UNIVERSAL
LIBRARY
OU_148869
UNIVERSAL
LIBRARY

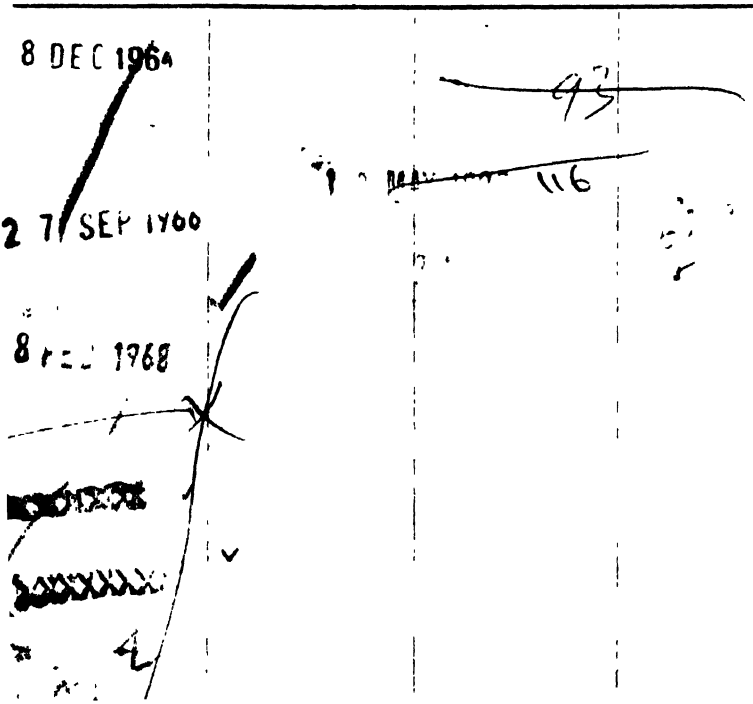
OSMANIA UNIVERSITY LIBRARY

Call No. Accession No.

Author

Title

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



DESCRIPTIVE AND

SAMPLING STATISTICS

Under the editorship

of

GARDNER MURPHY

DESCRIPTIVE AND SAMPLING STATISTICS

BY

JOHN GRAY PEATMAN

Associate Dean and
Professor of Psychology
The City College of New York

HARPER & BROTHERS PUBLISHERS

NEW YORK AND LONDON

DESCRIPTIVE AND SAMPLING STATISTICS

Copyright, 1947, by Harper & Brothers

Printed in the United States of America

A-B

All rights in this book are reserved
No part of the book may be reproduced in any manner whatsoever without written permission except in the case of brief quotations embodied in critical articles and reviews. For information address Harper & Brothers

To Lee, Alice, John, and Bill

Contents

PREFACE xv

PART I. DESCRIPTIVE STATISTICS

1. INTRODUCTION TO STATISTICS 3

A. Historical Background 3
Gamblers and Kings. The Mathematicians. The Census—Vital Statistics. Adolphe Quetelet—Social Scientist. Sir Francis Galton—Geneticist. Correlation. Statistical Prediction Actuarial, not Individual.

B. Descriptive vs. Sampling Statistics 9
The Concept "Statistics"—Its Various Meanings. Description vs. Sampling. The Reduction of Data.

C. The Nature of Statistical Data 13
Non-Variable Data. Variable Data. The Treatment of Statistical Data. The Mathematical and Logical Implications of a Variable—A Series. Exact and Approximate Measures.

2. THE REDUCTION AND ORGANIZATION OF CATEGORICAL DATA 19

A. Introduction 19

B. The Classification and Enumeration of Attributes 19
Dichotomous and Polytomous Classifications of Attributes. Classification vs. Division. Rules for Logical Division and Classification. Classification of Judgments, Attitudes, and Opinions. Classification of Don't Know's (*DK*'s) in Market Research Investigations. The Statistical Frequency. Enumeration vs. Measurement. Stratification—An Opinion Poll.

C. Methods for Treatment of Original Data 37
The Hand-Sorting of Statistical Data. Machine Tabulation. The Findex System of Coding and Analysis.

3. THE COMPARISON OF CATEGORICAL DATA: PROPORTIONS, PERCENTAGES, RATIOS, INDEX NUMBERS 43

A. Ratios and Percentages 43
Proportions. Rounding Off Numbers.

B. Use of Percentages for Comparing the Parts of Two or More Wholes 49

C. Ratios and Index Numbers 52
Per Capita Indices. Ratios as Index Numbers.

D. Confusion in the Use of Percentages 55
Confusion in Interpreting a Percentage Increase. A Percentage Decrease Can Never Be More Than 100%. Confusion Between Percentages and Proportions. Confusion from Large Percentages. Percentages from too Small a Base. Errors in Averaging Percentages.

E. Graphic Methods for the Presentation and Comparison of Categorical Data 58
Bar Graphs. Belt Graphs. Pie Diagrams. Maps. Pictorial Charts.

4. THE CORRELATION OF CATEGORICAL DATA 80

A. The Cross-Tabulation of Categorical Data 80
Cross-Tabulation Essential to Correlation. The Correlation of Non-Variable Attributes. The Correlation of Polytomous Attributes—Market Research Data.

B. Methods for the Correlation of Categorical Data 90
Yule's Coefficient of Association (A) for Dichotomized Non-Variable Attributes. The Correlation of Dichotomized Variables: The Phi Coefficient. The Correlation of Polytomous Attributes: The Contingency Coefficient.

5. THE REDUCTION AND ORGANIZATION OF VARIATE DATA 99

A. Introduction 99

B. The Range and Array 99
The Range as a Comparative Measure. The Array.

C. The Frequency Distribution 103
The Class Interval. The Tally. The Frequency Distribution.

D. The Histogram and the Frequency Polygon 113
The Histogram. The Frequency Polygon, or Line Graph. Comparative Usefulness of the Histogram and Frequency Polygon.

E. The Percentage Frequency Distribution 120

F. The Cumulative and Percentage Cumulative Frequency Distribution 121
The Cumulative Frequency Distribution. The Percentage Cumulative Frequency Distribution. Usefulness of Percentage Cumulative Graph for Comparing Distributions.

6. THE CENTILE POINT METHOD FOR VARIATE DATA 127

A. Centiles and the Description of Variate Data 127
Centile Point Values vs. Centile Intervals. Quartiles, Terciles,

Quintiles, Deciles, and Vigintiles. Comparative Implications of Centile Measures. The Determination of Centiles.

B. Centiles by the Graphic Method 131
The Centile Graph. Determining the Score Values of Centiles from a Centile Graph.

C. The Computation of Centile Values 134
The Location of a Centile Point. Interpolating the Score Value of a Centile Point. Checking the Computed Centile Value. Comparison of Estimated and Computed Centile Values.

D. Centile Measures 139
The Median (A Measure of Central Tendency?). The *D* Range —A Measure of Dispersion. The Quartile Deviation—A Measure of Deviation or Variability. The Tercile Deviation—A Measure of Deviation or Variability.

E. The Use of Centiles for Comparing the Results of Two or More Distributions of a Variable 142

F. The Use of the Centile Method for Comparing the Results of Two or More Variables 146

7. THE MEAN AND STANDARD DEVIATION 150

A. The Method of Moments for Variate Data 150
Basic Symbols.

B. The Mean 151
Definition. Method I: The Mean from Unordered Data. Method II: The Mean—Long Method with Data Grouped into a Frequency Distribution. Method III: The Mean—Short Method with Grouped Data.

C. The Standard Deviation 160
Definition. Method I: Standard Deviation from Ungrouped Data. Method II: Standard Deviation—Long Method with Grouped Data. Method III: Standard Deviation—Short Method with Grouped Data. Method IIIa: Standard Deviation—Short Method with Ungrouped Data. Sheppard's Correction for σ.

D. The Average Deviation 168
Definition. Method I: Average Deviation—Ungrouped Data. Method II: Average Deviation with Grouped Data.

E. The Coefficient of Relative Variation 171

8. COMPARATIVE IMPLICATIONS OF THE NORMAL, BELL-SHAPED CURVE 174

A. Implications of M and σ for Normal, Bell-Shaped Distributions 174
The Mean as Point of Reference. The Mean as a Fulcrum. The Median and Mean. Uni-Modality and the Mode. Bilateral

Symmetry. Points of Inflection and σ. Asymptotic Character of the Normal Curve. The Practical Limits Equal $M \pm 3.0\sigma$. σ as the Standard Measure of Variability. Measures as z Scores. z Scores Signify Relative Position in a Series. Centile Implications of Standard Measures. Summary of Commonly Used Measures of Dispersion About the Mean. The Normal Probability Curve. The Formula for the Normal Curve. Relationship Between Various Measures of Variability in a Normal Distribution.

B. The Use of z Scores and Standard Scores for Comparative Purposes 184

Standard Scores. Standard Score Norms. The Standard Score Profile Chart or Psychograph.

9. THE PRODUCT-MOMENT METHOD FOR THE CORRELATION OF VARIATES 195

A. The Linear Correlation of Bi-Variates 195

Pearson's Product-Moment r. The Cross-Tabulation of Bi-Variate Data. The Scattergram of Bi-Variate Data. The Assumption of Linear Correlation. Plotting the Bi-Variate Data of a Scattergram. The Correlational Frequency: Paired Associates. The Correlation Chart.

B. Estimation of Product-Moment r 208

Fitting Linear Regression Lines to Bi-Variate Distributions. The z Score Correlation Chart. The Regression Line for $\bar{z}_y$ on z_x. Estimating r. The Regression Equation of $\bar{z}_y$ on z_x. The Regression Equation in Descriptive Statistics. The Regression of z_x on z_y. The Regression Equation for $\bar{z}_x$ on z_y. The Regression Coefficients. Regression Equations Expressed in Terms of x and y. Standard Formula for r.

C. Computation of Product-Moment r 225

Summary of Mathematical Implications of r. Various Methods for the Computation of r.

D. Method I: Product-Moment r from Ungrouped Data (Long Method) 226

Order of Operations for Method I. Shortcomings of Method I.

E. Method II: Product-Moment r from Grouped Data (Short Method) 229

The Frequency Distributions of Each Variable from the Correlation Chart. The Standard Deviations of Each Variable from the Correlation Chart. The Product Deviations. Ratio for r. Checking $\Sigma(x'y')$. Means and Standard Deviations from the Correlation Chart.

F. Method III: Product-Moment *r* from Ungrouped Data (Machine Method) 236
Machine Computation. The Guessed Means Taken as Equal to Zero. The Formula for *r* (Method III). Inter-Correlation Coefficients. Work Sheet for Original Data and Computation of Means, Squares, and Cross-Products (Table 9:3). Computation of Standard Deviations of All Variables (Table 9:4). Computation of the Mean of the Product Deviations of Each Bi-Variate Distribution (Table 9:5). Computation of the Correlation Coefficients (Table 9:6).

G. Other Methods for the Computation of *r* 247
The Method of Sums for *r*. The Method of Differences for *r*.

10. SPECIAL METHODS FOR THE LINEAR CORRELATION OF VARIABLES 253

A. Correlation of Ranks 253
Purpose of the Method. Spearman's Rank-Difference Method. The Relation of *r* to Rho.

B. Serial Correlation 258
Biserial Correlation. Point-Biserial Correlation. Triserial, Quadriserial, and Quintiserial *r*.

C. Tetrachoric Correlation 275
Purpose of the Method. The Computation of Tetrachoric *r* (r_t). Estimating Tetrachoric Correlation with Thurstone's Diagrams.

PART II. SAMPLING AND ANALYTICAL STATISTICS

11. SAMPLES AND SAMPLING TECHNIQUES 283

A. Introduction 283
Census vs. Sample. Sampling Is a Research Technique.

B. Statistical Populations or Universes 288
The Statistical Universe. Finite and Infinite Populations. Actual vs. Hypothetical Universes.

C. Samples and the Techniques of Sampling 290
Representative Samples. Biased Samples.

D. Random Samples—The Principle of Randomization 294
Definition. The Technique of Random Sampling. The Sampling Unit.

E. Stratified-Random Sampling 299
Definition. Stratifying Factors. The Technique of Stratification. The Inter-Relation of Stratifying Factors. Sub-Universes in Stratified-Random Sampling. Internal Controls in Sampling.

Areal Sampling. The Technique of the Master Sample. The Random-Point Method of Sampling. The Stratified-Quota Method of Sampling. Chief Source of Error in Stratified Sampling. The "Representativeness" of Stratified Samples.

F. Some Further Considerations About Sampling 313
Precision and Adequacy in Sampling. The Character of Samples vs. the Size of Samples. Accidental Samples. Restricted Universes and Partial Investigations. The Analysis of Intra-Group Differences in Sampling. Sampling in the Experimental Method of Equated Groups. Experimental Method with Random Samples.

G. Some Terminological Distinctions for Sampling and Analytical Statistics 322
Parameters and "True Measures." Statistics. Symbols for the Differentiation of Parameters and Statistics. Sampling Distributions. Small Sample Theory vs. Large Sample Theory. The Standard Error of a Statistic. Statistical Hypotheses. The Probable Error of a Statistic. Sampling Error and Error of Measurement.

12. PROBABILITY AND STATISTICAL INFERENCE 328

A. The Statistical Concept of Probability 328
Definition of Probability. A Single Event Has No P Value—The Concept of Likelihood. Strict Causality vs. Statistical Relations.

B. The Binomial Distribution and the Normal Probability Curve 331
Normal Sampling Distributions. Binomial for Samples of $N_s = 2$. The Product and Addition Theorems of Probability. Binomial for Samples of $N_s = 3$. Binomials for Larger Samples. The Expansion of the Binomial for the Normal Probability Curve. The Probability of a Result Derived from the Normal Probability Distribution. A Test of Significance (T). The Evaluation of the Test of Significance. The Distribution of Frequencies in the Normal Probability Distribution.

C. Small Sample Theory—Leptokurtic Sampling Distributions 347
Kurtosis (Ku). The t Statistic. When Is a Sample Small?

D. Skewed Sampling Distributions and Normal Probability 349
The Binomial When $p \neq q$.

E. The Precision (Reliability) of Sample Results and the Size of Samples 353
Precision Measured by the Standard Error. Precision Generally a Function of $\sqrt{N_s}$. Precision and Reliability.

13. HYPOTHESES AND TESTS OF SIGNIFICANCE 360

A. Likelihood and Confidence Criteria 360
Postulation of Parameters. Hypotheses Give Direction and Meaning to Research. The Probability Estimate. The Test of Significance and the Test Ratio (T). Likelihood and Confidence Criteria. Confidence Criteria in Terms of T Ratios.

B. Confidence Limits: Testing a Continuum of Hypotheses 368
Many Statistical Hypotheses Can Be Tested. Fiducial Limits and Confidence Limits. The Reliability of a Statistic.

C. Summary of Steps for the Testing of Hypotheses 371

D. Tests of Significance for Some Commonly Used Statistics 373
Percentages. Proportions. Frequencies. The Arithmetic Mean. Test Scores and Other Measures. Standard Deviations. Average Deviation. Centiles. Product-Moment Correlation Coefficients. Other Correlation Coefficients. Skewness and Kurtosis of Distributions.

E. The Probable Error and Tests of Significance 393

F. Tests of Significance for Small Samples 397
Fisher's t Statistic. Probability Values for t.

14. TESTS OF SIGNIFICANCE FOR DIFFERENCES BETWEEN STATISTICS 401

A. The Standard Error of a Difference Between Any Two Statistics 401
Standard Error of a Difference for Independent Samples.

B. Tests of Significance for a Difference Between Any Two Statistics 403
Confidence Criteria for the Significance of a Difference.

C. A Difference Between Percentages (or Proportions) Derived from Non-Correlated Samples 404

D. A Difference Between Percentages Derived from Correlated Samples 407

E. A Difference Between Arithmetic Means Derived from Non-Correlated Samples 409
Fisher's Null Hypotheses for Differences.

F. A Mean Difference Between Correlated Samples 412
Effect of Heterogeneity of "Matched Samples."

G. A Difference Between Standard Deviations 416
Combining the Results of Several Groups for a Test of Significance.

H. A Difference Between Coefficients of Relative Variation 418

I. A Difference Between Product-Moment Coefficients of Correlation 419

15. CHI-SQUARE AND TESTS OF SIGNIFICANCE 424

A. Chi-Square for the Distribution of Non-Variable and Variable Attributes 425

Calculation of Chi-Square. A Chi-Square Test of Significance of Consumers' Brand Preferences (a Dichotomy). The Probability of Chi-Square. Degrees of Freedom (*d.f.*). Chi-Square as a Test of Significance. A Chi-Square Test of Significance for a Trichotomy. A Chi-Square Test of Significance for the Distribution of a Variate.

B. Chi-Square Tests of Significance for the Independence of Two Attributes 437

Chi-Square Tests of Significance for Correlation Between Dichotomized Attributes. Pearson's Short-Cut Computation of χ^2 for 2 by 2 Cross-Tabulations. Chi-Square Test of Significance for Correlation Between Attributes with More Than Two Categories. Contingency Coefficient. Relation Between χ^2 and ϕ.

16. THE PREDICTIVE MEANING OF CORRELATION 445

A. Making the Prediction 447

Predictions on a Correlation Matrix.

B. The Accuracy or Efficiency of Predictions 451

The Standard Error of Estimate. The Interpretation of the Error of Estimate. Graphic Representation of the Accuracy of Predictive Estimates. The Index of Predictive Efficiency (E). Standard Error of Estimate for the Mean (σ_{est_M}). Tests of Significance for Predictive Estimates. Summary

17. CORRELATION METHODS FOR THE EVALUATION OF PSYCHOLOGICAL TESTS 464

A. The Reliability and Validity of a Barometer and of a Psychological Test 465

The Barometer. The Psychological Test.

B. The Determination of Test Reliability 470

Test Reliability by the Method of Test-Retest (r_{xx}). Test Reliability by the Method of Alternate Forms ($r_{xx'}$). Test Reliability by the Split-Half Method ($r_{\frac{x}{2}\frac{x'}{2}}$). Test Reliability by the Method of Item-Intercorrelation. Effect of Range of Ability on Test Reliability.

C. The Determination of Test Validity 478

Operational Validity. Functional Validity. Validity Criteria—Abilities vs. Aptitudes. Effect of Range of Ability on Test Validity.

D. Test Item Analysis 481
Item Reliability and Validity. Biserial and Fourfold Correlation Techniques.

E. Multiple Correlation (R) 482
Predicting Academic Success from Two Variables. Predicting Clerical Efficiency from Two Variables. The Multiple Regression Equation and the Standard Error of Estimate of R.

F. Partial Correlation 485
Partial Correlation with Scholastic Aptitude Held Constant. Partial Correlation with Age Held Constant. Spurious Correlation.

18. CLUSTER AND FACTOR ANALYSIS 489

A. Theory of the Organization of Human Traits 489
The Coefficient of Determination (r^2). Spearman's Two-Factor Theory. Multiple-Factor Theories. Sampling Theory and Cluster Analysis.

B. Methods of Factor Analysis 492
Tryon's Method of Correlation Profile Analysis. Cluster Analysis of Body Measurements. Cluster Analysis of Psychological Variables. Some General Implications of Factor Analysis.

APPENDIX A. Bibliography of Statistical Tables and Nomographs, Periodical Literature, and Chief References in Mathematical and Advanced Statistics. 505

APPENDIX B. Tables of Statistical Functions. 507

I. Areas and Ordinates of the Normal Probability Curve 508

IA. Ordinate Values of the Normal Curve Expressed as Proportions of the Ordinate at the Mean 511

II. Probability Values for T of Normal Sampling Distributions of Large Sample Theory 512

III. Distribution of t for Small Samples 514

IV. Distribution of Chi-Square 515

V. Values of Functions of r 516

VI. Values of Fisher's z Function for Values of r 518

VII. Values of Proportions, p and q 519

APPENDIX C. Tables of Squares, Square Roots, Reciprocals, and Random Numbers 521

I. Squares, Square Roots, and Reciprocals of Integers from 1 to 1000 522

II. A Table of Random Numbers 543

GLOSSARY OF STATISTICAL SYMBOLS 547

GLOSSARY OF PRINCIPAL STATISTICAL FORMULAS 551

INDEX 565

Preface

Statistical method is a fundamental and necessary tool for research workers in the social and biological sciences. It needs no more justification for its existence in these fields than does applied mathematics in the fields of engineering and the physical sciences. The methods of statistics are methods of applied mathematics; they are essential working tools for social and biological scientists because they provide the necessary scientific methodology for obtaining, organizing, summarizing, and analyzing research data.

Statistical method is not presented in this book as a discipline to be studied for its own sake; such an approach would be essentially mathematical. Rather, the emphasis is on its presentation as a useful and necessary tool for research problems in psychology and the closely related fields of education, cultural anthropology, and sociology; and considerable attention has been given to the use of statistics in public opinion and market research.

The presentation of statistical method as a research tool can be treated in various ways. Thus interest can be focused solely on the methods of computation, with the reasons for the methods, the logic of their application, and their value for particular problems left to the student's imagination (or the instructor's); on the other hand, computational methods may be given practically no emphasis. We have attempted a balanced presentation that will teach the student not only how to compute a statistical measure but when to use a particular technique and how to interpret a result. Some mathematicians may feel that no student can attain a satisfactory grasp of statistics without a knowledge of the mathematical bases and their implications. Certainly there is no question but that this knowledge is both important and helpful. However, the student who is interested primarily in a social science and only secondarily in statistics—as a means to an end, a tool—can obtain a sound working knowledge of the subject without, for example, being able to differentiate the normal probability distribution by means of the calculus. Such a student has fundamentally a fourfold need: (1) an appreciation of the usefulness of statistical method in his field; (2) an understanding of the logic underlying its application; (3) the ability to select the most relevant statistical technique and to make the necessary computations with a minimum of error; and (4) the ability to interpret a statistical result in a way justified by the character of the data.

Descriptive and Sampling Statistics is designed as a text for an introductory one-year course for either undergraduate or graduate students. Each of the two parts into which the book is organized—Descriptive Statistics, and

Sampling and Analytical Statistics—contains sufficient material for a semester course of 45 to 60 hours. Most of the various statistical methods are developed by presenting both the type of problem for which each method is required, and the logical basis for the statistical solution. Part II contains a chapter on probability, presented as a preliminary to the development of Tests of Significance, and also a chapter on sampling methods, because methods of sampling are as integral to sampling and analytical statistics as is the manner of treating the data derived from the samples. Contrary to the belief held by some lay persons that conclusions based on statistics are dubious or useless, adequate methods of sampling and measurement make it possible to draw conclusions that are as reliable and useful as those based on other scientific methods. Only in the hands of the inept or the fraud is there any justification for the popular saying that there are three kinds of lies—defensive lies, base lies, and statistics.

Acknowledgments to various authors and publishers have been made through the book. I am especially indebted, however, to Professor R. A. Fisher and to Messrs. Oliver and Boyd, Ltd., of Edinburgh for permission to reprint Table Nos. III and IV of Appendix B from their book, *Statistical Tables for Biological, Agricultural and Medical Research*, and for the adaptation of Table VI of Appendix B from this same work. For permission to reproduce various charts, I am also indebted to the Editors of *Broadcasting Magazine* and to Radio Station WOR, New York City (Fig. 6:4); to the Editors of *Fortune Magazine* (Fig. 3:19); to the Institute of Public Administration, New York City (Figs. 3:7, 3:8, 3:11, 3:12, and 3:13); to the Public Affairs Committee, Inc., New York City (Figs. 3:14, 3:16, and 3:18); and to the New York *Times Magazine* and the Pictograph Corporation of New York (Fig. 3:17).

I wish also to take this opportunity to acknowledge my indebtedness to Frederick E. Croxton, my first teacher in statistics, who inspired a lasting interest in the subject; to Gardner Murphy, Editor of this Series, for his many helpful and constructive suggestions; to Harriet Clemenson, Clare Luhman, Mary McDonald, Georgette Schneer, Madeline M. Sherwood, and Jean Brown Trapnell for their able and scrupulous assistance in the preparation of my original manuscript; and to Dorothy Thompson, Production Editor of the College Department of my publishers, for her competent and careful work in the final preparation of my manuscript for the press.

To my students I wish to acknowledge a great indebtedness for their curiosity and stimulation which have long been a great satisfaction to me in the teaching of statistics.

John Gray Peatman

PART ONE

Descriptive Statistics

CHAPTER 1

Introduction to Statistics

A. HISTORICAL BACKGROUND *

Statistics is a form of applied mathematics. It is a logical tool used in all the sciences and employed by all modern cultures. It is especially a tool of the biological and social sciences, a tool whose development has paralleled the practical demands of man's needs in a diverse and complex world.

Gamblers and Kings

Statistics had its beginnings many generations ago as a result of the interests and needs of gamblers and kings. The gamblers wished to develop systems that would improve their skill at cards and dice. Kings wished to know more about their subjects so as to work out more efficient taxing systems. Out of the interests and needs of gamblers came the foundation of our modern theory of probability, a theory basic to sampling statistics. Out of the interests and needs of kings emerged vital and social statistics, statistics as a descriptive method for enumerating and classifying hundreds and thousands of classes of useful data.

The Mathematicians

After the gamblers and kings came the mathematicians. In 1657 there appeared a brief treatment by Christian Huygens, the great Dutch mathematician and physicist, of the chances of winning at certain card and dice games. Three years earlier Pascal and Fermat had had their famous correspondence, in which they established the fundamental principles of probability. A little later Jacques Bernoulli, the Swiss mathematician, wrote the first book on the subject of probability. It was published in 1713, after his death, by his nephew, Nicolas Bernoulli. The work is an historical landmark, especially because of its emphasis on the practical value of the theory of probability for social problems. But Jacques Bernoulli's untimely death cut short the immediate development of many practical possibilities of statistics in social affairs. Such development waited another century, until the work of the Belgian, Adolphe Quetelet.

* Cf. H. M. Walker, *Studies in the History of Statistical Method*, Williams & Wilkins, Baltimore, 1929.

In the meantime, the theoretical development of statistics centered about the concept of probability initiated, as we have indicated, by Pascal and Fermat, and Jacques Bernoulli. In 1733, de Moivre gave the first mathematical formulation of the *normal* probability curve (the curve of error), but little attention was paid to it at the time. De Moivre attempted to remove the stigma of gambling from the problem of probability and to give the theory a divine flavor by maintaining: "And thus in all cases it will be found, that although chance produces irregularities, still the Odds will be infinitely great, that in process of Time, those Irregularities will bear no proportion to the recurrency of that Order which naturally results from Original Design." *

It was not, however, until toward the end of the eighteenth century and the beginning of the nineteenth that the theoretical development of statistics got under way as a broad and continuous enterprise. It was with the work of the great European mathematicians, Laplace and Gauss, and of the physicists and astronomers that the scientific foundations were laid for the theory of probability and the measurement of errors of observation. Gauss, "the Prince of Mathematicians," † was especially concerned with the practical as well as the theoretical problems of astronomical measurement, and the *normal curve of error* was developed for the variable results of observation with the *mean* of a series of observed values taken as the most probable value of the measure sought.

In this work of the mathematical astronomers it is evident that theoretical statistics was developing in conjunction with some empirical problems of measurement. However, the broad foundations of *descriptive statistics* (which Quetelet later integrated with the theoretical) for the study of social phenomena were established by government officials and political economists.

The Census—Vital Statistics

We have seen that kings had long been interested in enumerating those of their subjects who could pay taxes. They had also long been interested in the number of subjects who could render military service. The registering of baptisms, marriages, and deaths was begun in a few places in Europe during the fourteenth and fifteenth centuries. Such data formed the basis for the beginnings of *descriptive statistics*, and by the seventeenth century census taking had its *systematic* start. According to Godfrey,‡ the first census of modern times to be conducted under that name was taken in Canada in 1666. The data reported filled 154 pages and included facts about the population such as sex, family and conjugal status, age, profession, and trade. More

* Cf. H. M. Walker, *Studies in the History of Statistical Method*, Williams & Wilkins, Baltimore, 1929, p. 17.

† Cf. E. P. Bell, *Men of Mathematics*, Simon & Schuster, New York, 1937, chap. 14.

‡ E. H. Godfrey, Section on Canada, in John Koren (ed.), *The History of Statistics; Their Development and Progress in Many Countries*, Macmillan, New York, 1918, pp. 179–198.

recently, however, Dr. Carlos Casteñada, Latin-American authority at the University of Texas, has reported that the first census on the North American continent was conducted by the *alcaldias mayores* of New Spain between 1570 and 1580 at the command of King Philip II of Spain.* Philip wanted to know how many people there were, the family income, members per family, the amount of taxes they paid, and on what and with what they paid their taxes. Altogether there were 150 questions for each family to answer.

The end of the seventeenth century saw the publication of mortality tables by the English astronomer, Halley, in 1693. Annuity tables for insurance societies made a marked empirical development in the eighteenth century because of the vital statistics which had been collected by that time. The revolutions in America and France further stimulated the interest in data about the masses of population. Our Articles of Confederation provided for a triennial census, but this was changed to a decennial basis when the Constitution was adopted; and 1790 saw the first official census of the newly formed United States of America.

Adolphe Quetelet (1796–1874)—Social Scientist

It was Quetelet who developed statistical method as a scientific research tool in the study of man and the social sciences. Quetelet was a university teacher, mathematician, astronomer, and anthropometrist, as well as his country's supervisor of official statistics and hence responsible for the first nation-wide census. It was Quetelet who brought together the theoretical and empirical foundations of statistics, integrating and developing them for the investigation of social phenomena. He combined a mathematical interest in the theory of probability with a passion for the collection of data about people. Time and again, during the nineteenth century, he emphasized that the basic techniques of statistical method are the same whether we are studying the stars or man, the weather or morals. It was Quetelet who developed the concept of *the average man—l'homme moyen*—insisting that in the sphere of human activities all is not individual and unmeasurable. In 1831 he reported a study on tendencies to crime at different ages, in which he analyzed the role of such factors as sex, education, and climate on criminal tendency. Just as we are often startled by predictions about the number of deaths from accidents to be expected on the Fourth of July or for a given period from automobile traffic, so Quetelet was impressed by the relative constancy of the number of crimes from year to year: "Thus we pass from one year to another with the sad perspective of seeing the same crimes reproduced in the same order and calling down the same punishments in the same proportions. Sad condition of Humanity! . . . We might enumerate in advance how many individuals will stain their hands in the blood of their fellows, how many will

* C. D. Casteñada, in the New York *Herald Tribune*, July 7, 1940; also direct correspondence.

be poisoners; almost we can enumerate in advance the births and deaths that should occur. There is a budget which we pay with frightful regularity; it is that of prisons, chains and the scaffold." *

Quetelet was criticized as a materialist by many of his contemporaries because he dared to suggest that the moral worth of a man might be inferred from measurements of his actions, that the intellectual vitality of a man might be deduced from what he produced. He was confident that the mental and moral traits of man could be measured and that, when measured, the distributions of such traits would be shown to conform to the so-called *normal law*. The normal probability curve, which is illustrated in Fig. 1:1, came practically to be deified—and no wonder. As large samples of data of various characters of men, of biological and social phenomena, came to be measured, the distributions were often found to approach the form of this curve.

Sir Francis Galton (1822–1911)—Geneticist

After Quetelet, but contemporary with him as a statistician for a generation, came Sir Francis Galton. Like Quetelet, Galton also made extensive use of the normal probability curve in the description of biological and social phenomena. Like Quetelet, Galton saw in statistical method the means of discovering regularity and lawfulness in phenomena which otherwise, by their diversity and complexity, seemed individual and unique. Galton suggested the use of the normal curve in the assigning of grades, or class marks, in the schoolroom. Like Quetelet, Galton had a great passion for observation, for recording data and analyzing them by the methods of statistics, many of which he himself developed as the need arose. It was Galton who discovered the method of statistical correlation, a discovery made in connection with the need for analysis in his studies of the inheritance of traits. It is no exaggeration to describe this discovery of Galton's as one of the greatest contributions ever made to the empirical development of the biological and social sciences.

Correlation

The need for the technique of correlation is aptly illustrated by some of Bowditch's problems which he was unable to answer adequately, as he himself recognized. With the object of improving school application in growing children, the Massachusetts Board of Health sponsored the study by Bowditch, reported in 1877.† Descriptive statistics of nearly 25,000 children were obtained, including not only bodily measurements and age, but also nationality, place of birth, and occupation of parents. Bowditch wished to analyze

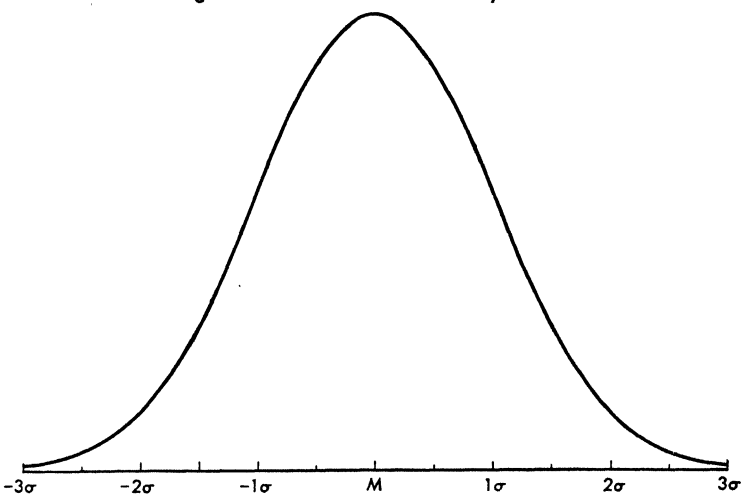
* F. H. Hankins, *Adolphe Quetelet as Statistician*, Columbia University Studies in History, Economics and Public Law, No. 84, New York, 1908.

† A. P. Bowditch, "The Growth of Children," Report of the Board of Health of Massachusetts, 1877; reprinted in Bowditch's *Papers on Anthropometry*, Boston, 1894.

this tremendous mass of data for such relations as might be relevant to the original problems of the inquiry. He wanted to know, for example, what relation there was between the height and weight of the children. He saw that there was a relationship, but the technique of correlation was not yet available with which to formulate a determinate answer regarding the degree or character of the relationship.

That there is an empirical basis for a possible relationship is obvious since height and weight are both attributes or traits of individuals. The real question, however, does not relate to the individual case. It is an *actuarial* or

Fig. 1:1. The Normal Probability Curve



The abscissa (horizontal) axis represents the *scale* of measures or scores of a variable attribute or trait. The ordinate (vertical) axis represents the *frequencies* of the distribution. The higher the curve at any point, the greater the number of frequencies or instances for the measures at that point. The point of greatest concentration of frequencies is in the center of the distribution, at *M*, the mean.

group question. We know that some people are likely to be tall, some short, some heavy, some light. Persons, even of the same age, thus *vary* in height and weight. Height and weight are therefore called *variables* or *variates*. Quetelet and others established the fact that there is a very real tendency for a large random sample of persons of a given age to have weights or heights which, when systematically organized into a series according to size, form a distribution which is similar to that of the normal probability curve (see Fig. 1:1).

The question of *relationship* between two variable attributes like height and weight is whether individuals of average height are also of average weight; whether very tall individuals are also very heavy; whether very short in-

dividuals are also very light. In other words, the question is whether weights and heights, when paired according to the persons from whom they are obtained, vary together in any systematic way. This is the problem of *co-variation* or *correlation*, and it is complicated by the fact that rarely, if ever, do the measured attributes of biological or social phenomena exhibit perfect or complete correlation. Persons who weigh, say, 160 pounds do not all have the same height; rather, they vary in height. Similarly, persons of a given height, say 6 feet, vary in weight. The statistical problem here is one of determining the *form and degree* of any tendency for weight and height to vary together. The details of the statistical technique of correlation demanded by this kind of problem, the problem of possible co-variation, will be considered in Chapter 9. Here we wish only to emphasize that the technique discovered by Galton has been indispensable to the modern development of the biological and social sciences.

It is by the statistical technique of correlation that we are today able by comparatively simple methods to investigate relations between the attributes of individuals, or of organisms generally, as well as relations between the attributes of other kinds of natural and social phenomena. What is the nature of the relation, if any, between the I.Q.'s and school grades of children, between the tested achievements of parents and their offspring, between the manual abilities of siblings? Is there any relation between temperature and plant growth, between neighborhood status and delinquency, between the protein content and proportion of vitreous kernels in wheat grains? Although methods of investigating such questions as these are sometimes complicated, the method of correlation itself remains a most powerful tool for the study of possible relations among the variable attributes of natural and social phenomena. It is again to be emphasized, however, that this method, as well as statistics generally, is for the study of group phenomena—of masses of instances. Inferences which can be made legitimately from statistical results are about the group, not about the individual instance. *Descriptively*, such results give us information about the group as a whole. *Analytically*, such results may often be used for predicting what may happen *in the long run* or *on the average*, but not in the individual case.

Statistical Prediction Actuarial, Not Individual

We say that the chances are even that a tossed coin will land heads or tails. We mean that *in the long run* a series of such tosses should give half heads and half tails. What happens in the given, individual toss is strictly determined, although we are unable to ascertain the determining conditions so as to predict which side of the coin will lie uppermost. Our ignorance of the many factors operating in the determination of the result is such, and our knowledge of what happens in the long run for a *fair* coin is such, that we say the chances are even, or fifty-fifty, that the coin will land heads or tails. This is thus a verbal, somewhat metaphorical expression of our ignorance about what will happen in the individual instance. Similarly, we say that the

chances are about even that a child to be born will be a boy or a girl. Again, the metaphor is based (1) on our ignorance of the determining factors in the given, individual instance, and (2) on the empirical facts of vital statistics which have revealed for thousands of births that the ratio of boys to girls is about 51 to 49.

These two examples should serve to illustrate the actuarial or group character of statistics. What is true for the proportion of heads and tails in coin tossing, and of the sex ratio of births in vital statistics, is also true for all statistical inference, in that predictions are actuarial and not individual. It is well established in psychological and educational measurement, for example, that there exists a real correlative relationship between the academic attainments and intelligence test achievement of the school population in our culture. Given a particular I.Q. score, say 70, obtained under optimum conditions of measurement, we can predict that school children with such an I.Q. will, *on the average*, be below average in their academic attainments. That this is an actuarial or group inference should be obvious; nevertheless, such a prediction is sometimes made for the individual child who is, after all, either below average or not, in his academic attainments. And what he will continue to do in his school work can be effectively and logically predicted with confidence only as a result of studying him as the psychological individual that he is. In dealing with the individual child, the psychologist finds it useful and valid to draw upon his fund of statistical or actuarial experience and information so long as he continues to focus his analytical attention on the unique totality of the particular child.*

That a child has an I.Q. of 70 is useful information so far as the psychologist determines as precisely as possible what the actual intelligence test performance means for that particular child. In fact, the competent psychological investigator uses an intelligence test chiefly for such a purpose, for the light which the child's performance may throw on his total personality. In individual diagnosis and prognosis, the calculation of the I.Q. score itself is incidental to this fundamental purpose.

We see, then, that the data and methods of statistics are for the study of group or mass phenomena. And statistical inferences are actuarial in character, i.e., they are inferences about what happens or may happen in the long run, or on the average.

B. DESCRIPTIVE VS. SAMPLING STATISTICS

The Concept "Statistics"—Its Various Meanings

We have been using the term statistics mainly to refer to a method. This is because we are primarily concerned in this book with statistics as a scientific method of description and analysis. However, it is well to note that the

* Cf. in this regard the comments on prediction by G. W. Allport in his presidential address to the American Psychological Association, 1939: "The Psychologist's Frame of Reference," *Psychological Bulletin*, 37:1–28, 1940, especially pp. 16–18.

word *statistics* is also used to denote the data or information about populations, about biological and social phenomena, that can be measured or enumerated. Although this latter use of the term has been suggested in the preceding pages, we wish specifically to differentiate *statistics as information* from *statistics as method.*

Statistics as information represents perhaps the most general use of the concept. Today there are literally thousands of publications presenting statistical information of various kinds: vital statistics, statistics of health and medical care, statistics of education, of social security and of labor, statistics of crime, statistics of governmental finance, statistics of agriculture, manufactures, minerals, of housing and building construction, of wholesale and retail trade, of public utilities, of money and banking, of security markets and corporations, statistics of international trade, of business activity, of commodity prices, of consumption, and of national income and wealth. Although we are not directly concerned with statistics as information, the student of psychology, anthropology, sociology, or education should be familiar with sources of statistical information relevant to his field of research. A short bibliography of source material is appended to serve this purpose (see Appendix A).

Statistic vs. Parameter Values

Another distinction in the use of the concept *statistics* arises in the study and analysis of populations by sampling methods. Any summary numerical values obtained from *samples* of data, such as measures of an average, of deviational tendency, of correlation, etc., are characterized as *statistics.* Such statistics are contrasted with *parameter* values, which are these same types of measures but are for a statistical population as a whole, rather than for only a sample of the population.

Statistical Method vs. Statistical Inference

A distinction is also sometimes made between statistical method and statistical inference. This is, however, a somewhat ambiguous and unnecessary distinction, since statistical inference is integral to statistics as a method.

Description vs. Sampling

A more useful distinction can be made with respect to the nature of statistical method itself, viz., the methods of *descriptive statistics,* on the one hand, and those of *analytical* or *sampling statistics,* on the other. Since this distinction has fundamental and important implications for statistical method in scientific research, and since this book is organized on the basis of the contrast, we shall describe the difference between descriptive and sampling statistics at this point and see more fully the implications of the distinction as we proceed.

The fundamental distinction between descriptive and sampling statistics is essentially as follows: In sampling statistics we study populations in terms of the data of *samples*. In other words, the data of a *part* are used as the basis for investigating or studying the *whole*. In descriptive statistics, on the other hand, no distinction between part and whole is made; the data obtained in a study are treated as if they constitute a whole. A *census* characteristically presents data for the methods of descriptive statistics, since, by definition, a census is a set of observations or measurements made for all members of a group or population. A *sample*, by definition, characteristically presents data for the problems and methods of analytical statistics. In both cases, however, the initial task in the statistical treatment of results is the *reduction of data*.

The Reduction of Data

One of the fundamental purposes served by statistical method is the reduction of data. What, then, do we mean by the reduction of data? It consists in the organization and summarization of data into forms that can be readily perceived and understood. Consider the schedule for information or data shown in Fig. 1:2, part of the school record for a child.

Fig. 1:2. A Schedule

Case No. _____ Date of Record _____
Name _____ Sex _____
(Surname) (First)
Address _____ School _____
Birth date _____ Place _____ Grade _____
Father's name _____ Occupation _____
Mother's name _____ Occupation _____
Brothers _____ Sisters _____
Mental Age _____ I.Q. _____ Test _____
Examiner _____ Date _____

Educational Achievement Record:

Area	Rating	Test	Date
---	---	---	---
---	---	---	---

Whether or not data are obtained in conjunction with the plan of an experiment or in conjunction with the plan of an educational system (or other agency) for maintaining relevant records, it should be obvious that a systematic schedule for recording the data is a labor-saving device. When possible, it is most convenient to arrange the data of a case on a card, the size of which will of course depend upon the amount of information to be recorded. A card 3 by 5 inches is about as small as can be easily manipulated; only rarely is a card or sheet larger than 9 by 12 inches needed. The procedure of recording is facilitated if the schedule is printed with appropriate descriptive terms for the various categories of data or information to be entered.

In recent years the problem of recording and handling great masses of data has been met by the development of the punch card, on which can be recorded by machine any type of information that can be coded by a number system, as well as original data which are numerical. To handle the coded data of such cards, sorting machines and various kinds of tabulators have been developed (cf. Chapter 2, Section C).

A schedule for a child's school record is generally useful in two ways. (1) It is valuable for the teacher, psychologist, etc., who works with the individual pupil and attempts to iron out problems of that child's adjustment to the school or other situations. This is a non-statistical, individual use of such information. (2) It is valuable for statistical purposes, which means that it is useful as one case in hundreds or thousands of such records which, when considered en masse or according to relevant groupings, may provide valuable information in the planning, financing, and management of an educational system. When the purpose is statistical, a given child's schedule is more appropriately signified by a convenient number rather than by his name, since the investigator is no longer dealing with the individual child but with one case in a group or mass of statistical information. Such results as are obtained in the statistical treatment of the data will apply to the group as a whole rather than to an individual case.

It should be apparent that it is impossible adequately to interpret the records of hundreds or thousands of such schedules of information unless the data are somehow classified and summarized. An investigator or research worker is thus faced with the very practical problem of reducing a great bulk of data to a form that will be more readily perceived and understood. The procedures to be adopted for such classification and summarization depend specifically upon the purposes of the investigation or inquiry. In general, however, the statistical procedures that may be used include only a few alternatives; they will be described in detail in the following chapters on methods of *descriptive statistics.* Here it is emphasized that descriptive statistical procedures serve a need which arises as soon as the observations or data of any survey or inquiry become at all sizable or bulky. For, as R. A. Fisher says, "No human mind is capable of grasping in the entirety the meaning of any considerable quantity of numerical data." * The statistical methods used for the reduction of data are of three kinds:

1. Graphic methods
2. Computational methods yielding numerical measures
3. Tabular methods

The aim, then, of descriptive statistics is the reduction of data so that the results of observation and measurement may be (1) made more immediately meaningful, and (2) presented in a form that will make interpretation

* R. A. Fisher, *Statistical Methods for Research Workers,* Oliver & Boyd, London, 7th ed., 1938, p. 6.

and comparison of results easy and unambiguous. In the light of the preceding discussion, descriptive statistics can now be defined as the organization and summarization of collections of numerical data, including data arrived at by the simple method of enumerating instances. Descriptive statistics consists in the reduction of groups or masses of data by means of tables, graphs, and numerical measures such as percentages or proportions, averages, measures of deviation or dispersion, coefficients of correlation, etc.

That the methods of descriptive statistics are essential to the methods of analytical or sampling statistics is apparent. Whether the data are of a census or of a sample, the first step in their treatment consists in their appropriate reduction or simplification.

C. THE NATURE OF STATISTICAL DATA

In general, statistical data are of two kinds. They are derived either from variables or from non-variables. Consider, for example, the kinds of statistical information collected about human beings. Census data provide us with information concerning the incidence or number of people by geographical areas, their ages, distribution with regard to sex, etc. Psychologists and sociologists bring together many kinds of information concerning human behavior and intelligence. The statistical data of the latter investigations consist of various psychological measurements, the frequency of different behavior patterns, scores from questionnaires, interest inventories, etc. Some of these data are variable and others are non-variable. Let us see what the distinction between them is.

Non-Variable Data

The incidence of the two sexes in a population provides a common example of non-variable data. People are either male or female. Sex is a non-variable attribute. A person can be categorized as either one or the other. Furthermore, no order is inherent in the arrangement of these two categories; that is, there is no basis in measurement for putting the male class first and the female class second, or vice versa. A non-variable attribute is thus one that exists with respect to distinct categories rather than with respect to a particular degree.

Non-variable data are often referred to as the *data of categories*. Categorical data are generally obtained simply by the enumeration of instances that occur, or that are observed to exist, with respect to the classes or categories under consideration.

Variable Data

In contrast to categorical data, variable data represent quantitative differences (variation) in the manifestation of a property or trait or attribute. Thus, the age and height of persons are examples of attributes that are variables.

The essential characteristics of a statistical variable are as follows: (1) The attribute being studied is capable of *quantitative differentiation* (at least, theoretically); (2) the data differentiated have *order* inherent in their nature, an order ranging from least to most. Thus, the age of individuals is infinitely variable (within the age range of human beings), inasmuch as *age* is susceptible to quantitative differentiation in terms of years, months, and days. Furthermore, a collection of age data can readily be brought together with respect to the order inherent in them, namely, an order that ranges from *least age* to *most age*.

Although sex was seen to be an attribute which yields non-variable rather than variable data, it should be observed that the ratio of males to females provides a measure—the *sex ratio*—which is a variable attribute. The sex ratio is an index that may vary in size for different calendar periods or places.

The Treatment of Statistical Data

Most of the methods of statistics have been developed for the treatment and analysis of variable data. This is because *variation* has, historically, been practically synonymous with the concept of *statistical phenomena*. We have already seen that Quetelet and Galton pioneered in the nineteenth century in the development of statistical methods. By and large, the methods they were responsible for were concerned with the variations characteristic of human beings and other natural phenomena. We saw that Bowditch, studying the growth of children, was faced with the problem of somehow relating two variable attributes—height and weight—but that it remained for Galton to develop a method of determining the correlation between the measurements of two such variables. Furthermore, the statisticians of the nineteenth century were inclined to consider the data of variables as forming a distribution of measures similar to the normal probability curve. For as large samples of data of various attributes or characteristics of man and other biological and social phenomena were observed and measured, the distributions of the collections of data obtained for a given attribute were often found to approach the form of this curve. The normal probability curve thus came to epitomize a fundamental property of a variable. Nevertheless, not all variable attributes yield distributions of this form. On the other hand, categorical data do not yield distributions of any kind. This is the case because the essence of any distribution is an *ordered series of measures* ranging from the *least degree* of the attribute observed or measured, to the *most degree*. Variable data are often referred to as the data of variates, and categorical data as the data of non-variates.

In consequence of the kinds of statistical problems arising during the nineteenth century and the early part of the twentieth, the bulk of the methods of statistics developed have been for the treatment of variable data. However, non-variable data are also important and accordingly some special methods

have been devised for their treatment and analysis.* These methods are especially relevant to many market research investigations, as well as to studies in social psychology and sociology. In Chapters 2–4, we shall present the basic statistical methods for non-variables as developed for problems of descriptive statistics, and in Chapters 5–9, the fundamental methods that have been developed for the descriptive treatment of variables. However, the distinction between these two sets of methods is not always sharp. The data of variables are sometimes treated by methods developed for categorical data. For example, in order to determine whether a particular aptitude test is satisfactory, the criterion therefor may be taken simply in terms of *successful and non-successful* performance. Obviously, *performance* is itself a variable attribute. However, we often lack satisfactory methods for quantitatively differentiating degrees of success or non-success and we obtain, at best, broad non-quantitative distinctions or differentiations of such attributes.

The Mathematical and Logical Implications of a Variable—A Series

The essence of a statistical variable resides in the two properties already mentioned: (1) the capacity of a characteristic or attribute to be *quantitatively differentiated* (by some process of measurement or observation), and (2) the presence of an inherent *order* in the data. When the statistical data of an investigation satisfy these two conditions, they yield a *series*, or *scale*, of measures. Such a series of measures ranges in numerical size from least to highest values. The concept of a series is thus implied by the order inherent in the quantitatively differentiated data of a variable. Some variables, however, can be studied and ordered into a series, but not quantitatively differentiated. Thus, the *social interests* of a group of people can be rated as "above average," "average," and "below average" (yielding a series with three broad classes), although there may not be available a satisfactory process of measurement that will yield quantitative differentiations of varying degrees of the attribute *social interest.*

Continuous vs. Discontinuous Series

Even though the data of a variable may satisfy the two properties of quantitative differentiation and order, there is a third property characteristic of such data that gives rise to a distinction among variables themselves. The data of variables may form either a *continuous* series of measures or a *discontinuous* series.

A continuous series of measures is one that, by definition, is theoretically susceptible to numerical subdivisions of any degree of fineness. A series of

---

* Cf. G. V. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, Griffin, London, 12th ed., 1940, chaps. 1–5.

age data is theoretically capable of such subdivisions. In practice, we may have no need to differentiate ages to finer degrees than years or months, but theoretically finer subdivisions in days or hours, etc., could be made. Such data thus form a continuous series, or *continuum*, that ranges from the least observed value to the highest observed value. This is the case, even though each subdivision in a continuum may not actually have an empirical datum.

On the other hand, the data of some variables do not, either in fact or theoretically, form a continuum, or continuous series of measurements. A collection of statistical information indicating the number of listeners per radio program, or the number of children per family, will yield data that satisfy the two basic properties of a variable, namely, quantitative differentiation and order. Such variable data may be arranged in a series ranging from the least number of listeners per radio program to the greatest number of listeners. However, it is obvious that only integral values can occur; there are no fractions of radio listeners. A distribution of such data thus yields a series that is non-continuous. There are real gaps between the integral values lying within the limits of the series. Such *discontinuous* series are often referred to as *discrete*, and the data of such a series are sometimes called *discrete data*. However, the latter terms are likely to be misleading, because discrete data are often confused with categorical data. It should be clear, however, that discrete data, as just defined, are the data of a variable rather than of a non-variable. Categorical data of a non-variable do not have an inherent order such that they can be arranged in a series of from least to most.

In statistical practice the data of a discontinuous series are usually treated *as if* they formed a continuous series. Thus the average number of children per family is usually calculated to a fractionate value, as for example, 3.5, despite the fact that such a value is an obvious abstraction. An average is useful because, for a collection of such data, it indicates that the typical number of children per family is midway between three and four children.

Exact and Approximate Measures

The data of statistical investigations are obtained by various methods. In general, however, the methods may be divided into two classes. (1) A great deal of statistical information is obtained by the simple method of enumerating or counting instances. (2) Statistical data are also obtained by a process of observation and measurement that is more complex than the method of simple enumeration.

The *method of simple enumeration* always yields an integral value. Such a value is an *exact* measure except for the possibility of errors in making a count. Categorical data are usually obtained by counting "noses," but the data of some variables, such as the number of children per family, are also obtained in the same way.

On the other hand, the data of variables are often *approximations*. They are

usually obtained by methods that yield estimates of *location* or *position* in a continuous series of values. Most of the measurements in the physical sciences are approximations obtained by well-defined methods of observation and measurement. Although they are approximations, they have, from a practical point of view at least, very small margins of error; in fact, oftentimes the errors are so small that they can be neglected. In psychology and the social sciences, a test score is an example of an approximate measure. It is usually obtained by a method of observation and measurement that provides an estimate of a person's position in a series or scale of test scores.

By definition, an approximate measure is one theoretically capable of greater exactness if the methods of measurement are continually refined. It is apparent that a continuous series of numbers is implied by the concept of an approximate measure. The fact that a statistical datum may be an approximation rather than an exact number is not, however, to be interpreted as thereby belittling its significance or usefulness. On the contrary, the difference between exact and approximate measures is a difference that results from the methods used in obtaining them. As just indicated, the method of the enumeration of instances (basic to all statistical data of censuses, many market research investigations, etc.) yields numbers or measurements which are exact in the sense that they are the result of a count. Nevertheless, it is to be observed that so far as a research investigation may consist of a sample drawn from a population of instances (as is characteristic of most market research studies), the count made of a sample is necessarily treated as an approximation of the population. Even though the count of the sample may be an exact number per se, from the point of view of its use as an estimate of a population value it is an approximation.

Similarly, the initial measurements obtained from many psychology tests are based upon a count. Thus, a vocabulary test score may be simply an enumeration of the number of correctly defined words in a list. Originally the vocabulary test score is simply an enumeration of correct responses, and from this point of view it is an exact number. However, as an estimate of a person's vocabulary ability, it is an approximation. This is true because the particular list of words used for the vocabulary test is only a sample of the test material that could be used for such a purpose. Since all psychological tests necessarily employ but a *sample* of test material, the measurements of ability yielded by a test are always approximations and never exact measurements. All such measures are estimates of people's positions in a series or scale of test performance. All such estimates are approximations.

EXERCISES

1. In what sense is statistics a form of applied mathematics?
2. What are the implications of Quetelet's work for the development of descriptive and sampling statistics?
3. State the different ways in which the concept "statistics" is employed.

4. What is the essential difference between descriptive and sampling statistics?
5. What is meant by the reduction of data?
6. What different kinds of methods are utilized for the reduction of data?
7. Distinguish between a non-variable and a variable attribute.
8. Distinguish between a continuous and a discontinuous variable.
9. What is the difference between exact and approximate measures?

CHAPTER 2

The Reduction and Organization of Categorical Data

A. INTRODUCTION

In this chapter we shall present some of the elementary but at the same time indispensable statistical methods for the treatment of the non-variate type of data often obtained in psychology, anthropology, sociology, and related fields. The data of non-variable attributes, of *categories*, their collection and statistical treatment are of basic importance to the research worker, even though a majority of research problems yield variate data.

From the point of view of the practical problems of research the initial task to be dealt with is classification, or division, of large masses of non-variate data. The logic of classification and division is essential to a sound use of methods for their reduction and comparison. Just as the psychologist and related scientists need to know the logic of measurement underlying the treatment of the data of variables, so they also need to know how to handle masses of non-variate data which first need to be classified and the results then described through the use of appropriate statistical techniques.

We shall consider first the problem of classification. Then methods for the *reduction* of such data to a useful form will be presented. Basically, these methods are simply *tabulation* and *enumeration.* Methods for the *comparison* of such data will be developed in Chapter 3. These methods consist chiefly in the calculation of ratios or rates, such as percentages. Finally, in Chapter 4 we shall present methods for the correlation of categorical data.

B. THE CLASSIFICATION AND ENUMERATION OF ATTRIBUTES

Categorical data are enumerated instances of attributes or qualities of objects or individuals that are taken as existing or not existing, rather than as existing to some degree. Hence, categorical data are derived from non-variable attributes, rather than from attributes or qualities that are variable.

Dichotomous and Polytomous Classifications of Attributes

Dichotomous Classification

We saw in the preceding chapter that the sex of human beings constitutes a non-variable attribute. People can be identified as either MALE or NOT-MALE.

This division is a *dichotomous*, mutually exclusive differentiation of a qualitative attribute. That is to say, it is a twofold classification of human beings with respect to an attribute (quality or trait) that can be differentiated qualitatively (MALE-KIND and NOT-MALE-KIND), but not quantitatively. It is a division such that a person can be identified as being MALE (the presence of the attribute in question) or NOT-MALE (the absence of the attribute) but not both (the two categories are mutually exclusive).

In the case of dichotomous divisions there thus should be two distinct, mutually exclusive classes or categories, as in the case of the attribute of SEX-KIND. The negative class, NOT-MALE, is of course usually identified by the positive descriptive term FEMALE. Although a positive term for the negative class is not always available for dichotomous classifications, the positive description of a class is empirically more satisfactory than the negative, provided no ambiguity results.*

Persons can also be divided into one or the other of the two following mutually exclusive categories: (1) the BLIND (total absence of vision), and (2) the NOT-BLIND, despite the fact that the latter class is variable, in that acuity of vision varies from little to much. Similarly, people can be divided into the LIGHT-HAIRED and the NOT-LIGHT-HAIRED. Here, however, the differentiation of the characteristics for each category is not so easy because of (1) the many variations in hair color, and (2) the problem of establishing satisfactory objective criteria for the appropriate identification of borderline cases. The extremes in hair color would, of course, be easy to identify and enumerate, but persons with in-between shades would be more difficult to classify. In any event, the line of division for an attribute of this kind would be arbitrary, whereas the distinction between the BLIND and the NOT-BLIND is not arbitrary.

Polytomous Classification

Eye color is again an attribute that can be divided into two categories, the BLUE-EYED and the NOT-BLUE-EYED. This time, however, the dichotomy itself is arbitrary. Eye color is an attribute which may, for research purposes, be more usefully differentiated into more than two categories. In fact, so far as it can be correlated with variations in degree of pigmentation, human eye color may be considered as a variable attribute. But at the present time there are no entirely satisfactory empirical methods for dealing quantitatively with this attribute. The usual method for field and laboratory purposes in psychology and anthropology consists in using a set of artificial eyes differing in pigmentation. By a matching technique (a person's eye color being compared with the colors and shades of the artificial eyes until the best match is obtained), the color and lightness of an individual's eyes are identified

* Technically, dichotomous division is restricted by logicians to a positive statement of the differentia and its negative: A and not-A.

with one of several categories or classes that differ in hue as well as in intensity. EYE COLOR is thus treated as a *polytomous* attribute, i.e., as consisting of several exclusive classes of hues which also have varying degrees of lightness or darkness.

In order to avoid ambiguities in the differentiation and enumeration of people with respect to the attribute of HAIR COLOR, similar matching techniques are used. In one such method, sample strands of hair are used (braided like strands of rope); the different colors range from the lightest to the darkest shades. The attribute of HAIR COLOR is thus treated categorically, despite the fact that it might be possible to arrange differences in HAIR COLOR on a quantitative scale ranging from LEAST-DARK to MOST-DARK.

Division by Exact Criteria

The line of division between the categories of an attribute may be arbitrary but at the same time exact. This is especially true of any attributes or traits that can be differentiated by enumeration or by a standardized method of measurement. Thus, men can be divided into two *arbitrary* but nevertheless exact classes with respect to the variable attribute of HEIGHT. This can be done by taking six feet of stature, for example, as the dividing line between the two categories, TALL-MEN and NOT-TALL-MEN. Similarly, animals can be classified as BIPEDS or NOT-BIPEDS; trucks, as SIX-WHEELED or NOT-SIX-WHEELED; children, as LIVING-WITH-BOTH-PARENTS or NOT-LIVING-WITH-BOTH-PARENTS. All such dichotomies as these have an element of arbitrariness, but it should be evident that they can be exactly established, inasmuch as the objects or individuals can be identified with one or the other of the dichotomized categories by counting legs, wheels, parents, etc.

In the final analysis, the problem of collecting and enumerating categorical data is one of the empirical identification of instances with respect to such distinctions among attributes as are relevant to the research investigation. If the dividing line of a dichotomized attribute is so vague as to produce ambiguities, then the researcher needs to develop more satisfactory criteria of division and identification. Dichotomous classifications imply that we can determine in which category of a twofold division an object is to be identified. Such a classification further implies that a given individual or object can be in only one of the two categories of the attribute; in other words, it implies that the two categories are mutually exclusive.

Classification vs. Division

From a logical point of view, a distinction is made between classification and division. If, in arranging the attributes of things, we proceed from the whole to its characteristics or traits—from the general to the less general—we are engaged in division. Thus, human beings may be divided into two groups—males and females—so far as the attribute of sex is concerned.

In classification, on the other hand, we begin with individual instances and seek common qualities or attributes—principles of organization—for grouping the individual instances into two or more relevant classes. Thus, humans $a, b, c, d \cdots n$ may be *classified* into two sex groups, males and females.

It should be apparent that a procedure for division which will be useful will depend upon some knowledge of the character of the different kinds of individual instances that comprise the whole. Hence, in practice, division is often not empirically distinguishable from classification. In any event, the results of division and classification are similar in that each gives a logical arrangement of an attribute or quality into two or more exclusive categories.

Science invokes both processes of division and classification in its attempt to make coherent and intelligible the apparent chaos of natural and social phenomena; however, uncharted fields of inquiry usually begin with individual instances and hence with classification. When the subject matter of a field is not already systematically arranged and interrelated by relevant concepts, the initial task of an inquiry is to establish criteria for the classification of individual observations. With the further development of a field of research, it becomes increasingly important to utilize systems of classification that have been empirically tested for their general usefulness. It is out of such verifiable schemes about natural phenomena that the systematic foundations of a science are made.

Stratification: Classes and Subclasses

Early students of zoology may have found the classification of all animals into the following three categories a useful scheme:

1. Water animals
2. Land animals
3. Air animals

But with increased knowledge of animals, this trichotomy came to be increasingly unsatisfactory. Not only because of ambiguities in the classification of individual animals but also because of the lesser relevance of *habitat* as the principle of classification, this scheme was finally abandoned in favor of others that took various attributes of animal *structure* and *function* as criteria for classification. Thus, zoologists today usually classify all animals into two general categories: *

1. PROTOZOA (without gastric cavity, germ layers, or tissues)
2. METAZOA (with gastric cavity, germ layers, and tissues; animals develop from eggs through cleavage, blastula, and gastrula stages)

This is but a beginning toward the classification of animals. Protozoa are considered as constituting the first phylum, and Metazoa are divided into

* Cf. E. G. Conklin, *General Morphology of Animals*, Princeton Univ. Press, Princeton, 1927.

thirteen phyla, or general subclasses. Several phyla of Metazoa are further subclassified into subphyla, subphyla into classes, classes into orders, orders into families, families into genera, genera into species, there being about a half million of the last subclass. Ordinarily, however, an animal is identified by only its genus and species (the binomial nomenclature introduced by Linnaeus in the eighteenth century). Thus, human organisms are identified as of the genus HOMO and of the species *sapiens*. *Homo sapiens* is of the *Primate Family*, which is of the *Order Eutheria*, which is of the *Class Mammalia*, which is of the *Subphylum Vertebrata* of the *Phylum Chordata*, which is of the Metazoa kingdom.

This zoological classificatory system is cited here because it exemplifies some of the fundamental problems that arise in the classification of data into subclasses. The *stratification* of a radio audience or of a group of voters for public opinion polls is based upon analogous schemes. It is the principles of classification that differ.

The Classification of Children's Apperceptive Responses

Although in many respects psychology, anthropology, sociology, etc., are beyond the classificatory stage of development, relevant schemes for classifying the data of an investigation still constitute an initial problem in much original research. Thus, in the analysis of personality differences in children, various kinds of projective techniques are being employed today. In a study by Elizabeth W. Amen,* the responses of 77 pre-school children to each of a series of 15 pictures were analyzed. After citing numerous instances of responses by the children, Amen writes:

> In the foregoing examples, three major types of responses can be observed:
> (a) A simple naming or other identification of objects. This may be regarded as response in terms of static form or enumeration ("A boy, a lady").
> (b) The description of the picture situation in terms of overt activity ("This little girl is eating her breakfast").
> (c) Inference as to psychological states or inner activity ("A little boy doesn't want to eat and his mama's going to get him to").
> When we study these categories of response, with reference to age level, it is apparent, from the examples given, that the two-year-olds respond predominantly in terms of static form. Description in terms of overt activity is more common after three years, and the suggestion of inner activity, rarely shown at two years, is fairly common at four years.

In the case of uncharted fields of inquiry, relevant principles of classification require a good deal of insight on the part of the investigator. But once the categories for classification are well defined, the statistical techniques necessary for reducing the data and presenting them for comparative purposes are relatively simple.

* Elizabeth W. Amen, "Individual Differences in Apperceptive Reaction: A Study of the Response of Pre-school Children to Pictures," *Genetic Psychology Monographs*, **23**:319–385 1941.

Amen's trichotomous classification of the children's responses evidently not only proved to be usable from the point of view of identification and enumeration of instances (responses), but revealed, as indicated, distinctions correlating somewhat with age difference. She brings together the verbal summary of her results, just quoted, into a single table which, however, omits the enumeration of instances (N) and gives only the *mean percentage* of responses for each category by age group. Thus:

Table 2:1. Mean Percentages of Types of Pre-school Children's Responses to Pictures

[Category of Response] *	[Age Groups]		
	2 Yrs.	3 Yrs.	4 Yrs.
Static form	73	38	23
Outer activity	26	50	51
Inner activity	1	12	26
[Total]	[100%]	[100%]	[100%]

* Material in brackets not in Amen's original table.

This cross-tabulation of Amen's results clearly indicates a correlation between category or type of response and age. *Static form* response is predominantly associated with the two-year group. *Outer activity* response is associated more with the three- and four-year groups than with the two-year-olds. *Inner activity* response is practically absent among the two-year-olds but is present in about one-eighth of the three-year group and in about one-quarter of the four-year-olds.

Rules for Logical Division and Classification

The need for the classification of phenomena is apparent from the foregoing examples. Natural phenomena, whether they be rocks or people, or the habits, attitudes, and preferences of people, need to be classified on the basis of similarities and differences in order that we may perceive and understand what goes on in an otherwise chaotic world of multitudinous and apparently unrelated events. The classification of things into categories is itself such a "natural" process of the human mind that we often overlook the bases for the sound classification of phenomena, or we become aware of them long after having engaged in the process of classification itself. Many children, for example, grow up with the stereotypes of their social environment and unthinkingly classify all people of a given skin color as good or bad, or all people of a given religion as more or less virtuous than other people.

Scientific method in the classification of phenomena aims to assure two things: (1) that a class or division be so defined and established that there will be no ambiguity or error in the identification of a thing or event in the

class; and (2) that the exploration and establishment of relations for phenomena within classes or between two or more classes be done on the basis of empirical information or evidence rather than on the basis of personal whim or prejudice. To classify a person as a Negro because of his skin color is a problem of identification; to conclude by virtue of the classification that he also has certain habits or attitudes is a separate problem of the relationship between two or more kinds of attributes or traits. In other words, there is the problem in scientific method not only of the identification of a phenomenon with a class, but also of the correct association of the relations between attributes or characteristics of things or events.

There are three rules of logical division which are also applicable to the problem of classification. They are usually described as follows: *

1. *A division needs to be exhaustive.*

In division, a category needs to be provided for every instance or member of the whole that is being divided. From the point of view of classification, i.e., working from individual instances to the whole, it is of course sometimes impossible at the outset of an investigation to have all categorical divisions perfectly provided for. As data are accumulated, new categories sometimes have to be added in order to avoid ambiguities in classification.

2. *The divisions into which the whole is differentiated need to exclude one another.*

The import of this rule is no doubt obvious. There should be no overlapping of instances from one division or category to another, so far as a given attribute is concerned. This exclusion is essential for the elimination of ambiguities in identification.

3. *The division should be based upon a single principle of differentiation.*

This is the principle of the *fundamentum divisionis*. Theoretically at least, division should be made with respect to a single quality or attribute. If a series of divisions and subdivisions is made, as in the stratification of data, this principle should be followed in each succeeding level of division. The division of children into the dull, the bright, and the fair-skinned obviously violates this rule, since two attributes, MENTAL STATUS and SKIN COLOR, are involved at the same level of division. The second rule of exclusion is also violated.

Classification of Judgments, Attitudes, and Opinions

Let us consider the application of these rules to the classification and division of people's judgments, attitudes, and opinions. In the comparison of pairs of weights, A and B, a subject is often doubtful as to which is the

* Cf. M. R. Cohen and Ernest Nagel, *An Introduction to Logic and Scientific Method*, Harcourt, Brace, New York, 1934, especially pp. 241, 242; also R. M. Eaton, *General Logic*, Scribner's, New York, 1931, especially pp. 282 ff.

heavier when the difference in physical weight is slight. His judgments are then frequently identified and enumerated with respect to one or the other of three categories, viz.,

1. *A* heavier than *B*
2. *A* lighter than *B*
3. Doubtful

This appears to be a trichotomous (or threefold division), but second thought suggests that what we really have here is, first, a dichotomous division of the character of all judgments into CERTAIN and DOUBTFUL, and second, a further dichotomous subdivision of the CERTAIN judgments into *A*-HEAVIER-THAN-*B* and *A*-LIGHTER-THAN-*B*. Instead of a true trichotomy, this situation illustrates the beginning of *stratification*, i.e., the division of classes into subclasses, subclasses into sub-subclasses, etc. Thus:

A. CERTAIN JUDGMENTS — B. DOUBTFUL JUDGMENTS

1. *A*-heavier-than-*B*
2. *A*-lighter-than-*B*

In recent years a multiple-choice answer method has been found useful in the development of attitude and interest questionnaires. E. K. Strong's Interest Inventory,* for example, asks whether the examinee likes "driving an automobile." The answer can be signified as

L (Like) or I (Indifferent) or D (Dislike)

The examinee is also asked whether he can write a concise, well-organized report. His answer can again be indicated in one of three ways, viz.,

YES or ? (not sure) or *NO*

Are these examples of truly trichotomous divisions? The latter is similar to the example of judging weights in that the examinee's judgments on this type of three-choice answer can be divided into SURE and NOT-SURE, with a further subdivision of the SURE judgments into YES and NO. However, the three-choice answer of the type L, I, and D (Like, Indifferent, and Dislike) appears to be in a somewhat different class because each type of answer expresses an attitude toward the verbalized situation (viz., driving an automobile). Does this threefold differentiation of attitudes satisfy the rules for satisfactory logical division and therefore yield a true trichotomy?

In order to answer this question, we need first to determine whether persons' attitudes are really exhausted by the threefold classification of *Like, Indifferent,* or *Dislike.* Pragmatically, i.e., from the point of view of the practical problems of research, it appears that many individuals have little difficulty in shaping their attitudes to fit into one or the other of these three categories, so long as it is clear that extreme dislike (hate) and extreme like are not

* E. K. Strong, *Vocational Interest Blank for Men (Revised)*, Stanford Univ. Press, Stanford University, 1938.

excluded. Sometimes fivefold or sevenfold divisions (cf. five-point and seven-point attitude scales) are used in the attempt to obtain finer distinctions in attitude.

As for the second rule: In practice there is only the respondent's initial difficulty at times in deciding between Like and Indifferent, or between Indifferent and Dislike. But once his judgment is made, there is no further difficulty in classifying it. The fundamental ambiguity that arises in psychological research with such choices hinges on the differences in meaning that a given type of answer may have for different people. Two persons may say that they dislike driving an automobile, but the experience of one may have been limited to a ten-year-old wreck of a car on corduroy roads, whereas the experience of the other may have been a $25.00 fine for speeding on a modern highway in a super-streamlined model. The point of course remains that both answer that they dislike driving automobiles. For some research purposes this "identity" of answer may, however, yield the desired and sufficient information.

It is the third rule that needs most carefully to be considered in examining the satisfactoriness of the threefold division (or classification) of attitudes into Like, Indifferent, and Dislike. Is this division based upon a single principle? The attribute or quality in question might be characterized as a FEELING ATTITUDE FOR A SITUATION. But how about feelings of anxiety, of joy, hope, sorrow, and pain? And is an attitude of indifference to be classed as a feeling attitude for a situation? Perhaps indifference usually means doubtful or uncertain. If the latter is the case, then, as already indicated, we have a division and subdivision of attitude rather than a trichotomous arrangement. The *certainty* of the attitude would be the first principle of division, and then the certain attitudes would be subdivided into attitudes of Like and Dislike.

In practice, the descriptive terms Like, Indifferent, and Dislike are often taken *as if* they yield a trichotomous division of feeling attitude for a situation and, consequently, *as if* a single principle of division of classification is involved. In practice, then, the third rule for division or classification is not always strictly adhered to, with consequences sometimes more useful than absurd. We have seen that no single principle but many principles are used in the biologist's classification of organisms. If, in the preceding example, we admit that a feeling attitude of Indifference introduces a principle of division which is not the same as that used in distinguishing attitudes of Like and Dislike, the practical gain of such a "trichotomous" division may nevertheless well offset the failure to satisfy a formal principle of logic. It is when a trichotomous division gives misleading or inaccurate information that it should be discarded.

Attitude Scales and Variable Data

On the other hand, answer choices that express attitudes are often treated *as if* they form a continuous scale of variable data ranging from *least* to *most*

in a given kind of attitude. If such an approach is attempted with the attitude choices of Like, Indifferent, and Dislike, the investigator is faced with the problem of locating attitudes of Indifference on the scale. The scale could be defined, for the quality of Like, as ranging from *Least Like* to *Most Like.* Dislike attitudes would be in the lower range of such a series, but Indifference would not fit into the scale at all, inasmuch as it does not represent any degree of Like. As a matter of fact, it is difficult to see how Dislike could be in such a scale in a genuine psychological sense. Attitudes of Dislike are qualitatively different from attitudes of Like. The verbal device of *Least Like* for *Dislike* is only a dodge for casting the results into the form of a variable. Essentially, as we have already pointed out, attitudes of Like, Indifference, and Dislike may be treated *as if* forming a trichotomy or *as if* forming a dichotomy of Certain and Not-Certain attitudes, with the Certain ones further subdivided into a dichotomy of Like and Dislike. Attitudes of Like may among themselves range from least to most, or from little to much, and hence be treated as if they form a scale of variable data. Similarly, attitudes of Dislike may vary in degree of Dislike, attitudes of Indifference may vary in degree of Indifference. But to attempt to force all these three *kinds* of attitudes into the same scale is to belie the basic facts.

An Ambiguous Trichotomy

In a poll of voters' opinions, the following question is asked:

WHOM DO YOU EXPECT TO WIN THE NEXT MAYORALTY ELECTION IN NEW YORK CITY?

The wording of this question suggests that each person polled will have an expectation as to who will win. However, three types of answers may be received:

1. Mr. X
2. Mr. Y
3. Don't know (DK)

Do these three classes of replies constitute a trichotomous division of three exclusive classes? No, because *Don't Know* here means NO EXPECTATION, and this is a class or category to be contrasted with that of EXPECTATION.

Instead of the trichotomy suggested by these results, we therefore really have a dichotomy with respect to expectation, and a further dichotomized subclass with respect to two candidates. In other words, two different strata of replies are involved. Unlike the categorical division of an attribute into two or more classes which should be exclusive and exhaustive, the categories between different strata (classes and subclasses) are necessarily overlapping. Thus, the attribute of the first stratum is EXPECTATION, with its negative, NO EXPECTATION. These categories are dichotomous because they are mutually exclusive and exhaustive of all such opinions. The second attribute, viz., EXPECTED CANDIDATE TO WIN may also be stated dichotomously as "Mr. X"

and "NOT-Mr. X." If there are only two candidates, the negative of "Mr. X" could be stated positively by the name of his opponent, "Mr. Y." If, however, there are more than two candidates, we may arrange at least a trichotomous division for the second stratum attribute. Thus:

EXPECTED CANDIDATE TO WIN

1. Mr. X
2. Mr. Y
3. Other

A tabulation of a poll of voters' opinions can thus be unambiguously arranged as follows:

A. THOSE WITH AN EXPECTATION (or Opinion)

1. Mr. X
2. Mr. Y
3. Other

B. THOSE WITH NO EXPECTATION (or no Opinion)—(*DK*'s)

Such a stratification of voters' opinions is not only unambiguous and exhaustive, but essential to an adequate interpretation of the results. For example, the percentages of voters' opinions need to be considered in relation to the total sample of replies, including those with NO EXPECTATION, for otherwise, if the latter answers are not included, Candidate X might receive a majority of the expectancies but in reality be named by less than a majority of all the people polled. Possible ambiguities and misleading interpretations can always be avoided if the researcher makes clear the base used in computing a percentage and at the same time indicates the character of the total sample result.

Classification of Don't Know's (DK's) in Market Research Investigations

The frequent occurrence of DON'T KNOW responses in market research investigations gives rise to a class of categorical data which needs careful analysis, else ambiguities will creep into the interpretation of results. Generally, we need first to distinguish between *DK*'s obtained from questions of *information* and *DK*'s obtained from questions of *opinion.* The *DK*'s for the question, Whom Do You Expect to Win the Next Mayoralty Election in New York City? cited in the preceding section, are of the latter kind; the respondents answering DON'T KNOW had no opinion. Now let us consider *DK*'s to questions of information. Generally, a respondent's *DK* will be symptomatic of either (1) ignorance or (2) failure to recall a once-known fact.

Lazarsfeld,* who has analyzed the problem of *DK*'s in market research,

* P. Lazarsfeld, *The Statistical Handling of the Don't Know's,* Office of Radio Research, Columbia University, New York, 1941.

points out that *DK*'s to questions of information may be generally classified into one or the other of two groups:

I. *DK*'s to questions of information, the answers to which the researcher knows.

II. *DK*'s to questions of information, the answers to which the researcher himself does not know.

Type I DK

When an investigator is chiefly interested in ascertaining whether a respondent knows or can recall a fact which the investigator himself knows, the *DK*'s obtained will be of Type I. These are the *DK*'s that occur on most radio quiz programs. Thus, *DK*'s to questions of the following kind are Type I:

WHO IS THE CHIEF JUSTICE OF THE U. S. SUPREME COURT?
WHO SPONSORS THE WEEKLY RADIO HIT PARADE OF POPULAR SONGS?
WHAT MAN HAD TWO NON-SUCCESSIVE TERMS IN THE WHITE HOUSE?

In all these questions, the aim is to find out whether the respondent is acquainted with certain facts which the investigator himself knows. Questions on most college examinations are obviously of this type.

Type II DK

Type II *DK* results when the investigator is chiefly interested in obtaining information which he himself does not have and which can usually be obtained most readily by asking an appropriate sample of respondents. Whereas, in questions leading to Type I *DK*, the investigator wants carefully to avoid giving the respondent any suggestions or clues to the answers, in questions leading to Type II *DK* he usually attempts to aid the respondent in recalling as accurately and completely as possible the information sought. Psychological questions in personality diagnosis are of this second class. Thus, the investigator asks:

DO YOU HAVE HEADACHES FREQUENTLY?

Here the fact is not known to the investigator; and if the respondent replies DON'T KNOW, the investigator may attempt to clear up the uncertainty of such an answer by asking the respondent further questions and by indicating more precisely what is meant by the modifier "frequently."

Questions leading to *DK*'s of Type II occur extensively in market research investigations. Thus,

WHAT KIND OF SOAP DO YOU NOW USE?
HOW MANY SUITS HAVE YOU PURCHASED DURING THE PAST YEAR?
WHEN DID YOU PURCHASE THE CAR YOU NOW OWN?
DO YOU TRADE AT X STORE?

DK's of Type II are much more difficult to classify than *DK*'s of Type I. The latter are usually unambiguous in their research implications. If a re-

spondent says he does not know who sponsors the weekly Hit Parade of popular songs, it is clear that his *DK* means ignorance of or failure to recall a fact. From the investigator's point of view, this *DK* is final in that it gives him the type of information he wants. If he finds that the great majority of listeners to a sponsored radio program give *DK*'s to this type of question, it is clear that the advertising on the program is not very effective.

On the other hand, *DK*'s of Type II are likely to be ambiguous in their research implications, even though they also may imply ignorance of or failure to recall a once-known fact. Type II *DK*'s give difficulty because the investigator himself usually does not know the answer; hence, the inclusion of such *DK*'s with other classes of answers may give misleading information.

Type II DK's in a Market Research Investigation (Example adapted from Lazarsfeld)

A psychologist conducting an investigation of consumers' habits and motives asks a sample of 3000 adults of City X

IN WHAT KIND OF STORE DID YOU BUY THE SHOES WHICH YOU ARE WEARING?

The hypothetical results are tabulated and summarized in Table 2:2.

Table 2:2. Replies of 3000 Adults to the Question: In What Kind of Store Did You Buy the Shoes Which You Are Wearing?

Type of Store	Number of Respondents
1. Department store	1620
2. Chain shoe store	360
3. Independent shoe store	510
4. Don't know	510
Total	3000

As it stands, this tabulation gives four categories for the attribute TYPE OF STORE. However, if we can assume that the three types of stores named are exhaustive of all possible types, the shoes of the 510 *DK*'s were necessarily bought in one or the other of the stores in the first three categories. The *DK*'s thus do not form a fourth, separate (exclusive) category for the attribute TYPE OF STORE. How, then, shall the classification be arranged?

The answer depends upon the nature of the problem under investigation and the consequent direction of research interest. If the basic problem is one of ascertaining whether the 3000 people interviewed could REMEMBER where they bought their shoes, then the data in Table 2:2 need to be reclassified into a major stratum for the attribute RECALL, and into a substratum for the attribute TYPE OF STORE, as in Table 2:3.

If, however, the chief research interest is in the fact of where the 3000 pairs of shoes were bought, rather than the respondents' ability to recall, the classi-

Table 2:3. Memory of 3000 Adults for Type of Store in Which Shoes Were Purchased

Type of Store	Recall	
	Remember	Don't Know
1. Department	1620	?
2. Chain	360	?
3. Independent	510	?
Total	2490	510

fication of the data in Table 2:2 can be modified as in Table 2:4 so as to yield this information and at the same time clarify the hidden character of the 510 *DK*'s.

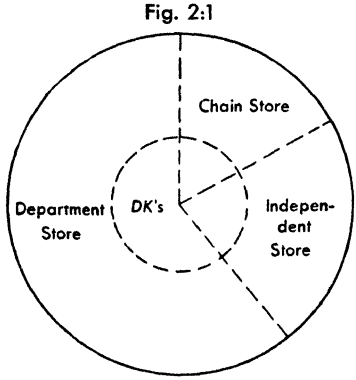
The logical possibility of the arrangement in Table 2:4 of the data in Table 2:2 is dependent upon the assumption already mentioned, viz., that the shoes of the 510 *DK*'s were purchased in one or the other of the three types of stores named. If these three types are exhaustive of all classes of stores in which shoes are purchased, the logic of the classification is sound.

Table 2:4. Type of Store in Which 3000 Pairs of Shoes Were Purchased

Type of Store	Frequency
1. Department store	1620
2. Chain store	360
3. Independent store	510
4. Department or chain or independent (*DK*'s)	510
Total	3000

The fact still remains that this fourth category (Department or Chain or Independent Store) for the 510 *DK*'s is not exclusive of the other three. Obviously, it is a class which, in an unknown manner, overlaps with at least one, possibly all, of the three categories of store types. Thus, logically, we have a situation such as is illustrated by Fig. 2:1. The radii and inner circle are drawn with broken lines because the actual breakdown of the 510 *DK*'s is not available and consequently the exact proportions of the three types of stores to the whole are uncertain. The inner circle is drawn to run through all three categories because

Fig. 2:1



there is the logical possibility that at least some of the *DK* purchases were made in each type of store.

There remains the research problem of reducing or completely eliminating the *DK*'s by securing more information from the respondents, perhaps by looking for labels in their shoes. In any event, the investigator would not be logically justified in distributing the *DK*'s among the three types according to the proportions of each type to the total number of purchases unless he had empirical evidence (as from a sample of the 510 *DK*'s) to warrant such a breakdown. Although it is possible that the *DK*'s came from respondents who generally trade in all three types of stores, it does not necessarily follow that their shoe purchases would be distributed among the three types in the same proportions as were the purchases of those who remember where they bought their shoes.

The Statistical Frequency

The statistical data of categories are obtained by (1) the *identification* of the presence (or absence) of the attribute or trait with its appropriate class or category, and (2) the *enumeration* or count of the number of individual instances thus identified with each category. The enumeration of all instances in a given category yields the *statistical frequency* of that category.

Statistical frequencies of categories or of classes constitute the *raw data* of an inquiry. Thus the data in Tables 2:2, 2:3, and 2:4 consist of raw data, whereas those in Table 2:1 are not *raw data* but *refined data*, i.e., they have been treated statistically and are presented as mean percentages.

Once the principles of classification have been determined for an investigation and the appropriate categories established, the next steps in treating the information available are *identification* and *enumeration.*

Enumeration vs. Measurement

In a broad sense, i.e., in the sense that measurement may be regarded as "the delimitation and fixation of our ideas of things, so that the determination of what it is to be a man or to be a circle is a case of measurement," * enumeration or counting may be described as a form of measurement. But in a stricter meaning of the term, and as ordinarily employed in sciences using quantitative methods, measurement means the determination of the magnitude or size of an attribute, i.e., the determination of the degree to which an object or an individual instance manifests a property or quality that varies in amount or extent, or from least to most, or from more to less. In this sense, measurement may be defined as "the correlation with numbers of entities which are not numbers." † Such measurement means the quantitative differentiation of an attribute or quality. It is the kind of measurement

* Ernest Nagel, *On the Logic of Measurement*, New York, 1930, p. 17.
† *Ibid.*, p. 17.

usually implied as characteristic of the study of individual differences in psychology.*

Stratification—An Opinion Poll

In the development of a student opinion poll of upperclassmen at the College of the City of New York during the academic year 1940–1941,† an attempt was made to increase the representativeness of a 10% *sample* of 200 upper-class male students by classifying the 2035 Juniors and Seniors with respect to several attributes. The 10% sample was then drawn randomly from each category of the established strata, according to the proportion of all students identified with each, as will be explained below.

When each individual (or object) of a group is classified with respect to more than one attribute, each will be a member of more than one class; this is characteristic of stratification. Thus, each student of the population of 2035 was classified with respect to the following attributes: COLLEGE CLASS, DEGREE GROUP, ROTC MEMBERSHIP (as well as several other attributes not included here). The principles involved in stratification are the same for any number of attributes as for three. The classification completed, the sample of 200 upperclassmen was drawn randomly and in proper proportions from each combination of categories of the stratified population. It is the development of the latter in which we are interested at this point.

In the stratification of a population for sampling, the first problem is to decide upon the attributes to be used. This choice is necessarily limited by the available information about the characteristics of the population to be studied. Within this limit, the investigator attempts to choose attributes that are significantly related to the problems of the investigation. In practice, it is usually necessary to make preliminary investigations in order to ascertain whether a given attribute (quality or trait) is relevant. In The City College opinion poll, the upperclassmen were stratified with respect to COLLEGE CLASS, DEGREE GROUP, and ROTC MEMBERSHIP on the presumption that individual differences in these attributes *might* be associated with differences in the opinions to be studied.

The Schedule of Information

The investigation began, then, with the population of 2035 upperclassmen. The first step consisted in obtaining the necessary information for each student and recording it in a form convenient for further use. Ordinarily, the construction of a schedule on a 3 by 5 card serves satisfactorily (see Figure 2:2). An individual card for each person has the advantage of easy manipulation and arrangement in the course of the investigation. When machines are available, the data can be transferred from the original cards to punch cards and a good deal of the sorting and analysis facilitated.

---

* Cf. M. R. Cohen and Ernest Nagel, *op. cit.*, chap. 15.

† M. Dreyfuss, *The City College Opinion Poll*, Honors Research in Psychology, 1941.

Fig. 2:2. Schedule of Information for Stratification of Sample

Name --JOHN JONES----------				No. ----1908---	
COLLEGE CLASS:	L.Jr.	(U.Jr.)	L.Sr.	U.Sr.	
DEGREE GROUP: Arts	(Soc. Sci.)	Sci.	Bus. Adm.	Tech.	Educ.
ROTC MEMBERSHIP: Yes ----X----		No ---------			

As indicated in the schedule card in Fig. 2:2, there were the following four categories for the attribute COLLEGE CLASS:

FIRST STRATUM
1. L.Jr. = Lower Juniors
2. U.Jr. = Upper Juniors
3. L.Sr. = Lower Seniors
4. U.Sr. = Upper Seniors

Within each of the four categories of the first stratum there were six different categories for the attribute DEGREE GROUP, as follows:

SECOND STRATUM
1. Arts (B.A.)
2. Social Science (B.S. in Social Science)
3. Science (B.S.)
4. Business Administration (B.B.A.)
5. Technology (Engineering, with several different degrees not differentiated here)
6. Education (B.Ed.)

Finally, for each of the preceding groups of the second stratum there was the following dichotomous substratum for the attribute of ROTC MEMBERSHIP:

THIRD STRATUM
1. ROTC (those who were at the time or had been members of the voluntary ROTC Unit at the College)
2. NR (those who were not at the time and had not been members of the ROTC at the College)

The stratification of 2035 students with respect to the three attributes COLLEGE CLASS, DEGREE GROUP, and ROTC MEMBERSHIP thus yielded three strata with four exclusive and exhaustive categories in the first stratum, six in the second, and two in the third. Again it should be noted that *within* a given stratum the categories are exclusive, whereas *between* strata they are not exclusive. Each individual is identified with respect to all three attributes: *his* COLLEGE CLASS, *and* his DEGREE GROUP, *and* ROTC MEMBERSHIP.

Exclusive Combinations of Attributes

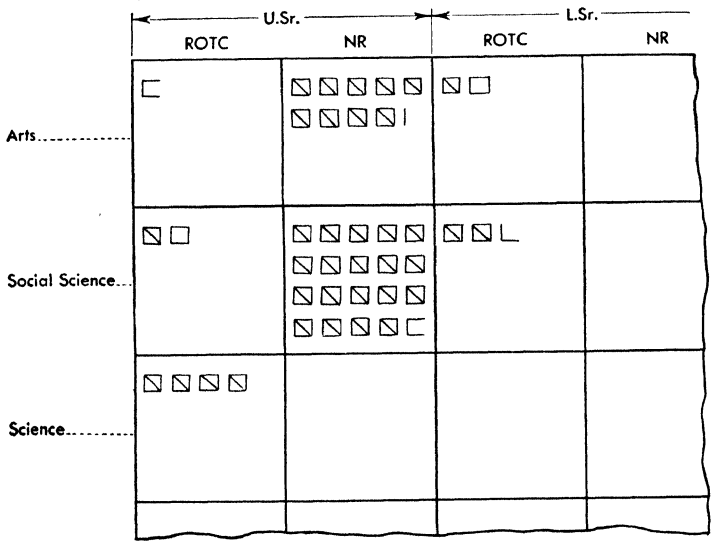
This particular stratification of a population yields the possibility of 48 *exclusive combinations* of attributes, since 4 (CLASSES) × 6 (DEGREES) × 2 (ROTC) = 48. Membership in any one of these 48 cells is exclusive; that is, a given student can be classified with respect to only one combination of the

three attributes. However, there are not necessarily members available for each one of the combinations or cells of a stratified matrix, i.e., the layout of a scheme for stratification.

Statistical Frequencies

With the necessary information available for each student, the next step consisted in arranging a matrix or table for the systematic tally of the frequencies of students for each combination of attributes. Each student was then identified and tallied in the appropriate cell of the stratified matrix and

Fig. 2:3. Part of Tally Sheet for Stratified Data



the frequencies per cell were summed. If the number of cases in any cell is likely to run into three figures, it is important that the original work sheet for tallying be laid off on a fairly large scale. Since the tallying procedure is the same throughout the work sheet, only part of it is reproduced in Fig. 2:3. The box method for tallying is used. Thus a box with a diagonal is made for each group of five cases in a cell, as follows: first case, |; second case, L; third case, ⊏; fourth case, □; fifth case, ⧅. This method eliminates errors that often arise in the 𝍸 || method of tallying. The boxes are easy to count since all cases are quickly identified in groups of five.

The final tabulation with the frequencies summed per cell is presented in Table 2:5. The Total column at the right of the table and the two Total rows at the bottom bring together succinctly the quantitative information for each category of each stratum. A stratified matrix like that in Table 2:5

has the advantage of presenting a summary picture of all the basic data, and in such a way that it can be readily condensed. It should be noted that the data of a stratified tabulation are *cross-tabulated*. That is, instead of separate and independent listings of the information for each category of each stratum, the students who are Upper Seniors *and* Members of the Arts Degree group *and* of the ROTC group are indicated ($N = 3$). In a similar fashion, each of the remaining 47 cells represents a cross-tabulation of the possible combinations of the three attributes. Such a cross-tabulation not only is essential for the determination of the proportion of individuals in each possible combination of the three attributes, but is also the basis for a correlational analysis. (See Chapter 4.)

Table 2:5. Stratified Classification of 2035 Students by College Class, Degree Group, and ROTC Membership

Strata		College Class and ROTC Membership								Totals by Degree Group
		U.Sr.		L.Sr.		U.Jr.		L.Jr.		
		ROTC	NR	ROTC	NR	ROTC	NR	ROTC	NR	
Degree Group	Arts	3	46	9	38	8	32	8	44	188
	Soc. Sci.	9	98	12	123	23	102	32	142	541
	Science	20	135	33	140	33	119	38	161	679
	Bus. Adm.	1	2	0	0	0	0	0	5	8
	Tech.	27	126	31	83	39	86	59	123	574
	Educ.	1	11	3	11	1	7	1	10	45
Totals		61	418	88	395	104	346	138	485	$N = 2035$
Totals by College Classes		479		483		450		623		$N = 2035$
Totals by ROTC		ROTC = 61 + 88 + 104 + 138 = 391; NR = 418 + 395 + 346 + 485 = 1644								

C. METHODS FOR TREATMENT OF ORIGINAL DATA

The Hand-Sorting of Statistical Data

We have already indicated that the tabulation and analysis of statistical data are often facilitated by the use of an individual schedule, say a 3 by 5 card, on which the information for each case is systematically recorded. Not only is it easier to work with the data from such cards, but the tallying procedure itself can often be simplified by hand-sorting the cards into appropriate classes or categories prior to the actual tabulation.

Thus, the data tabulated in Table 2:5 were obtained from the individual records of 2035 students. Accuracy in tallying as well as greater ease in the

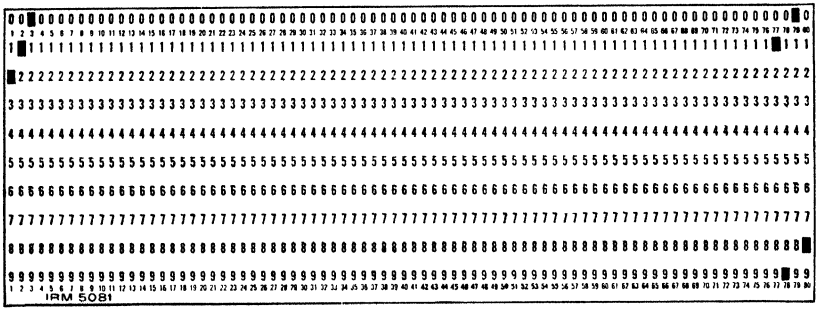
process was obtained by first sorting the members of the DEGREE GROUPS into their respective categories.

A further advantage that arises from recording the data of each case in a study on a separate card is that any *part* of the whole group of data can readily be studied by pulling out the cards of that part. By contrast, if all the data of, say, 50 or more cases are recorded on one sheet of paper, not only is hand-sorting of the cases impossible but it is very difficult to work with the data of only a part of the whole group. In the latter instance it is practically necessary to re-record the data of the smaller group on a separate sheet of paper in order to avoid errors.

Machine Tabulation

In recent years the treatment of statistical data has been greatly facilitated by the use of machines. A special card is used for coding the original data of each case. One type of card, which is reproduced in Fig. 2:4, is an

Fig. 2:4. An I. B. M. Card



I.B.M. (International Business Machine) card, which will take the data of many attributes or characters, whether they are variables or non-variables. Inspection of the card reveals 80 columns (numbered at the bottom) and 10 rows. Each column thus has a total of 10 positions numbered from 0 through 9.

The information in Table 2:5 can readily be coded and punched on such cards, and a machine called an Analyzer can quickly bring together the total number of students in each category, as well as the total number in any *combination* of categories that may be desired. The process of coding such a card will be briefly described for the three attributes in Table 2:5. The first column of the I.B.M. card will be used for coding COLLEGE CLASS. Inasmuch as there were four categories for this attribute, four of the ten positions of Column 1 will be needed. The 0 position can be used for coding those individuals who were Upper Senior students; the No. 1 position for those who were Lower Seniors; No. 2, for Upper Juniors; and No. 3 for Lower Juniors. The DEGREE GROUPS can then be coded in the second column of the card. Six

positions will be necessary, 0 to 5. Similarly, ROTC membership and non-membership can be coded in two positions in the third column. Technically only one position is needed to code a dichotomous attribute. Thus, those students who were members of the ROTC could be coded at 0, and those who were not members would not need to be coded at all (no entry). However, if a dichotomous attribute is coded by the use of only one position on a card, care is necessary to make sure that there are no cases for which information is not available. If this is not checked, all such instances would be counted with that group for which the code was "no entry."

Often the code number of each case is written at the top of the I.B.M. card; however, it may readily be punched. Since there were 2035 students in the population of upperclassmen, the students' case numbers will require four columns (Columns 77–80). Case No. 1 is punched as 0001, No. 10 as 0010, etc. Fig. 2:4 shows the data for Student No. 1908 coded as follows:

Column 1, Position 2: Upper Junior (College Class)
Column 2, Position 1: Social Science (Degree Group)
Column 3, Position 0: Yes (ROTC Membership)
Column 77, Position 1:
Column 78, Position 9:
Column 79, Position 0:
Column 80, Position 8:
} Case No. 1908

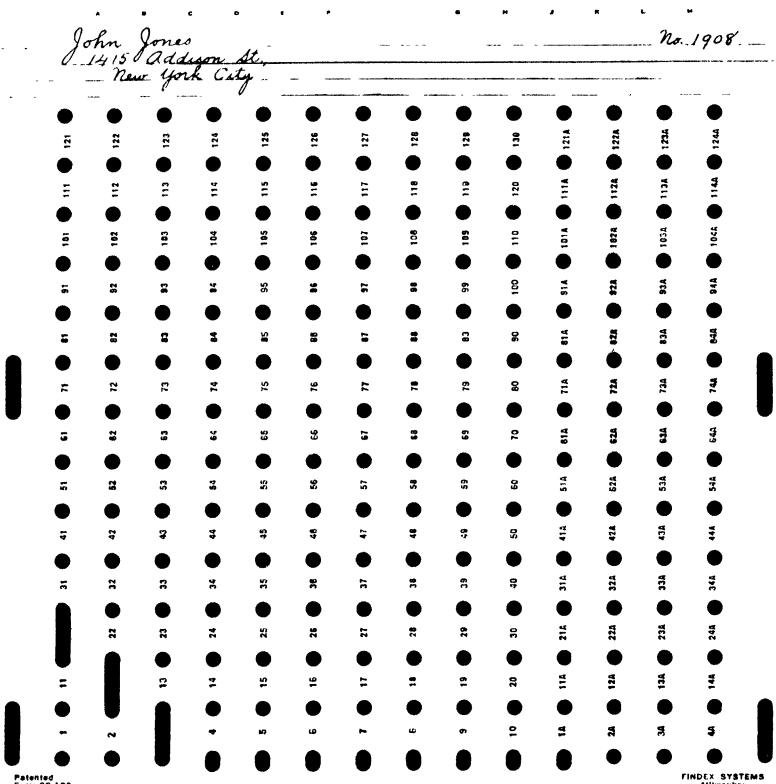
Machine tabulation and analysis are particularly useful when there are many cases in an investigation and many attributes or characters to be studied. The quantitative data of variables can readily be coded by the use of two, three, or four or more columns, depending upon the nature of the quantitative data obtained. If, for example, the scores of individuals on a psychology test were to be coded and the scores ranged from 1 to 99, only two columns on the I.B.M. card would be needed. A score of 96 would be represented by a slot at the No. 9 position in the first of a pair of columns, and at No. 6 position in the adjoining column.

The Findex System of Coding and Analysis

Machines for the analysis of card coded data are expensive and frequently not accessible to the research worker. Fortunately, there are available on the market several semi-machine methods for recording and analyzing data which are not so expensive. One of them is the Findex System. This consists of a special code card and a cabinet with special file drawers into which selecting rods can be inserted and the data of any attribute or character or combination thereof readily brought together. The method is semi-automatic, in that the individual code cards do not have to be hand-sorted and the selection of the appropriate data from a set of cards is done not by the eye but by a selecting rod.

A Findex card is illustrated in Fig. 2:5. The two slots on each side of the card are for guide rods which keep the cards in alignment in the file drawer. This particular card is the largest one manufactured by the Findex Company and has a total of 182 positions, there being 14 columns and 13 rows. As already

Fig. 2:5. The Findex Card



indicated, one position on a card code will serve to record the data of a dichotomous attribute or a variable which has been dichotomized. Furthermore, combinations of positions can be used to provide many more than the 182 simple positions on this card.

The method used with the Findex System is as follows: Each statistical datum is represented by a position on the card. An entry at a position is coded by a slot-cutting device. The slots in the card in Fig. 2:5 represent the same data for Student No. 1908 as does the I.B.M. card in Fig. 2:4. The bottom row, position Nos. 1 to 4, has been used for College Classes; in the

row above this, position Nos. 11 to 16 represent Degree Groups; and position No. 21 of the third row from the bottom has been used for ROTC membership. Thus, for Student No. 1908:

Position No. 3: Upper Junior
Position No. 12: Social Science
Position No. 21: ROTC Membership

"No entry" for No. 21 may be used to signify non-membership in the ROTC, provided there are no cases in the total group about whom the investigator does not have the relevant information. If there are such instances, then position No. 22 would be used to signify non-membership in the ROTC.

If the investigator wishes to determine how many members of the group are Arts students, he places the punched cards in the file drawer and inserts a rod at the No. 11 position. The cards are locked in the drawer by rods inserted through the four slots in the margins of the card. He then tilts the drawer to an upside-down position and ruffles the cards, i.e., separates them slightly from each other by running his fingers along the sides of the cards. All the cards on which an entry has been made at position 11 will drop about half an inch (the distance of the slot). A lock rod is then inserted through one of the holes at the bottom of the card and the drawer is returned to its original position. The lock rod prevents the cards which dropped from returning to their original position. Thus they can be quickly counted to give the total number of Arts students.

As with the I.B.M. cards, the data of a Findex card can also be cross-tabulated. This is done by the simultaneous insertion of selecting rods for the two or more characters which are to be studied. Thus, the total number of Arts students who are Upper Seniors and members of the ROTC can readily be determined by inserting rods in positions 1, 11, and 21 and repeating the operations just described. Only those cards with entries in each of the three positions will drop.

An advantage of the Findex System over a straight machine method is that the research worker can keep in closer touch with his original data and can examine the details of a case record at any time. For sizable groups of cases, the method is of course slower than machine analysis, but it is more rapid than hand-sorting, especially in the cross-tabulation of two or more attributes.

EXERCISES

1. Give five examples, not mentioned in the chapter, of dichotomous non-variable attributes.
2. Give five examples, not mentioned in the chapter, of polytomous non-variable attributes.
3. State the differences between Type I and Type II *DK*'s.
   a. Give five examples, not mentioned in the chapter, of Type I *DK*'s.
   b. Give five examples, not mentioned in the chapter, of Type II *DK*'s.

4. Set up a stratification for the student body of a college or university in terms of attributes or characteristics other than the three used in the example from the Dreyfuss opinion poll, and state the relevance of each attribute selected to a research problem on the attitudes or opinions of the student body.
5. Outline a research problem on the opinions or attitudes of the members of your community and select three attributes for stratification which might possibly be relevant to the research problem outlined.

CHAPTER 3

The Comparison of Categorical Data: Proportions, Percentages, Ratios, Index Numbers

A. RATIOS AND PERCENTAGES

The raw or original data of non-variable attributes are reduced to a form more manageable for interpretation by means of tabulations and enumerations into appropriate categories. Such reductions yield statistical frequencies, described in the preceding chapter. The *comparison* of two or more sets of categorical data is further facilitated by the reduction of enumerated values to appropriate *proportions*.

The most commonly used proportion is the *percentage* (from *per centum*), which is a proportion taken to a base of 100. That is, a percentage is a proportion multiplied by 100. Percentages are employed more generally than any other type of proportion for the comparison of categorical data. When percentage values for a given type of data repeatedly occur as very small values, the basic proportions involved are often taken to other bases, as to 1000 (per mille), or 100,000, or 1,000,000.

Since percentage values, per mille values, etc., are always derived from a proportion, the development of the latter will be considered first.

Proportions (*p*)

A proportion is a ratio of two numbers, such as the ratio of a part to the whole. This ratio is usually expressed in decimals. Thus, the ratio $\frac{1}{2}$ has a p value of .50. This may also be written as .5, but proportions are usually expressed to at least two decimal places.

The value obtained in computing a proportion always depends on the number taken for the *base*. Proportions are often misleading because they have not been taken to the base implied in their interpretation. Thus, in the comparison of the number of ROTC-Upper Seniors with the number of ROTC-Lower Juniors (Table 2:5), the statistical frequencies were found to be 61 and 138.*

Will a ratio of 61 to 138 give us the desired proportion for comparing these two groups?

$$p = 61/138 = .44$$

* For brevity throughout this discussion, the attribute ROTC will be employed to indicate, as previously, those students with past as well as present membership in the college ROTC unit.

Before answering this question, let us see just what the value .44 signifies. Aside from the fact that it means there are somewhat less than half as many persons in the first group as in the second, the value .44 expresses the mathematical fact that

$$61 : 138 :: .44 : 1.00$$

Sixty-one is to 138 as .44 is to 1.00. In other words, a proportion always signifies the ratio of a number taken to a *base of 1*. If the comparison is turned around and the proportion of ROTC-Lower Juniors to ROTC-Upper Seniors is obtained, then

$$p = 138/61 = 2.26$$

This means that there are more than twice as many persons in what is now taken as the first group than there are in the second group. Mathematically, the proportion 2.26 means that 138 is to 61 as 2.26 is to 1. Thus, in the Lower Junior Class there were 2.26 ROTC members for each one ROTC member in the Upper Senior Class.

Whichever way the comparison is made by means of a proportion, it gives the ratio of the first number to the second number taken to a base of 1.

Absolute vs. Relative Comparisons

Returning to the question of whether .44 gives us the proportion desired for comparing the two ROTC groups, we find that the answer depends upon the purpose of the comparison. The use of the proportion 2.26 rather than .44 does not alter the basic facts of the comparison; it merely shifts the direction of the comparison. That is, 2.26 indicates that there were more than twice as many ROTC-Lower Juniors as ROTC-Upper Seniors, whereas .44 indicates that there were less than half as many ROTC-Upper Seniors as ROTC-Lower Juniors.

From the point of view of the number of uniforms and other facilities to be provided, this *absolute increase* in ROTC membership in the Lower Junior Class has its significance, independently of what is happening to the size of *each Class as a whole*. The proportions .44 and 2.26 compare the absolute differences in the sizes of the two ROTC groups, independently of the Classes of which each is a part. For most purposes, the comparison of the absolute differences would be just as usefully made in terms of the numbers themselves as in terms of proportions derived from them.

If, on the other hand, it is relevant to compare these two groups *in relation to* the respective Classes of which they are parts, the procedure for calculation will be different. If the purpose is to determine whether there was a relative increase as well as an absolute gain in ROTC membership, it will be necessary to take into account the difference in the sizes of the Upper Senior and Lower Junior Classes. This is done by obtaining:

1. The proportion of ROTC-Upper Seniors to all Upper Seniors.
2. The proportion of ROTC-Lower Juniors to all Lower Juniors.

3. The comparison of the resulting proportions, either directly or by means of a proportion.

Thus:

$$(1)\ \frac{\text{ROTC-U.Sr.}}{\text{All U.Sr.}} = \frac{61}{479} = .13$$

$p = .13$ (a ratio of approx. 1 in 8)

$$(2)\ \frac{\text{ROTC-L.Jr.}}{\text{All L.Jr.}} = \frac{138}{623} = .22$$

$p = .22$ (a ratio of approx. 1 in $4\frac{1}{2}$)

Thus, of all Upper Seniors .13, or approximately 1 in 8, were in the ROTC. Of all Lower Juniors, .22, or approximately 1 in $4\frac{1}{2}$, were in the ROTC. However, let us make this comparison of the two groups by an additional proportion:

$$(3)\ p = \frac{.13}{.22} = .59; \quad \text{or} \quad p = \frac{.22}{.13} = 1.69$$

If ROTC-Upper Seniors are compared with ROTC-Lower Juniors, *relative to the size of their respective Classes*, there were about three-fifths (.59) as many ROTC members in the Upper Senior Class as in the Lower Junior Class. Or if ROTC-Lower Juniors are compared with ROTC-Upper Seniors, there were (again relative to the size of each Class) more than $1\frac{1}{2}$ as many ROTC members in the Lower Junior Class as in the Upper Senior Class.

It is evident, then, that the *relative* difference—relative to the size of their respective Classes as a whole—between the two ROTC groups was not as great as the absolute difference. Thus:

Ratio of Absolute Difference: .44 and 2.26
Ratio of Relative Difference: .59 and 1.69

An Absolute Difference, with the Absence of a Relative Difference

An absolute comparison will at times signify a difference—a gain or a decrease—whereas, relatively, there is no difference. Thus, Table 2:5 showed 33 ROTC men who were Science Degree-Lower Seniors, and 38 ROTC men who were Science Degree-Lower Juniors. Absolutely, there is a difference of 5—an increase of 5 ROTC men in the Science Degree-Lower Junior group.

However, *relative* to the size of their respective groups (Science-Lower Senior and Science-Lower Junior), there was no change. Thus:

$$\frac{\text{ROTC-Science-L.Sr.}}{\text{All Science-L.Sr.}} = \frac{33}{33 + 140} = \frac{33}{173} = .19$$

$$\frac{\text{ROTC-Science-L.Jr.}}{\text{All Science-L.Jr.}} = \frac{38}{38 + 161} = \frac{38}{199} = .19$$

The proportion of those in each group who were ROTC men therefore remains the same when considered in relation to each Science Degree group as a whole.

An Absolute Increase but, Relatively, a Decrease

It is also possible for there to be an absolute increase in the size of two groups, and at the same time a *relative* decrease. Thus, if the ROTC membership of Science Degree-Upper Juniors is compared with that of the Science Degree-Lower Juniors on the basis of the data in Table 2:5, the following results are obtained. There were 33 ROTC members in the Science Degree-Upper Junior Class and 38 ROTC members in the Science Degree-Lower Junior Class. In absolute terms, there was an increase of 5 members. But, relative to their respective groups as a whole:

$$\frac{\text{ROTC-Science-U.Jr.}}{\text{All Science-U.Jr.}} = \frac{33}{33 + 119} = \frac{33}{152} = .22$$

$$\frac{\text{ROTC-Science-L.Jr.}}{\text{All Science-L.Jr.}} = \frac{38}{38 + 161} = \frac{38}{199} = .19$$

There was therefore a relative decrease in ROTC membership among the Science Degree-Lower Juniors as compared with the Science Degree-Upper Juniors. The former group contained

$$.19/.22 = .86 \text{ of an ROTC member}$$

to each ROTC member in the Science Degree-Upper Junior group.

The ROTC Group Taken as a Whole

However, we have by no means exhausted the comparisons that can be made of the data in Table 2:5. For example, if we wish to consider *ROTC upperclassmen as a whole* and ascertain the proportion who are Seniors, as compared with Juniors, we use the following procedure:

Number of ROTC upperclassmen = 61 + 88 + 104 + 138 = 391
Number of ROTC Seniors = 61 + 88 = 149
Number of ROTC Juniors = 104 + 138 = 242

Proportion of ROTC Seniors:

$$p = 149/391 = .38$$

Proportion of ROTC Juniors:

$$p = 242/391 = .62$$

Check: The sum of the ratios of all parts of a whole should equal 1.00.

$$.38 + .62 = 1.00$$

Thus, nearly two-fifths (.38) of all ROTC upperclassmen were Seniors, whereas about three-fifths (.62) were Juniors. With respect to each other, then, the proportion of Junior Class-ROTC Members to Senior Class-ROTC Members was better than $1\frac{1}{2}$ to 1.

$$.62 : .38 :: 1.64 : 1.0$$
since $.62/.38 = 1.64$

Rounding Off Numbers

The proportions for the comparisons in the preceding section have been given to two decimal places. The actual arithmetical operations can, of course, be carried much further. However, when the smallest groups involved have proportions of .01 or greater, ratios to two (or three) decimal places are adequate for ordinary comparative purposes.

A standardized procedure to be used in determining the value of the last figure of the decimal when there is a remainder will now be described. The general rules are as follows:

Rule 1: If, in division, the remainder is less than one-half the number value of the divisor, the value of the last obtained digit of the quotient is unchanged. Thus,

```
        .86
.22 ) .1900
      176
      ---
       140
       132
       ---
         8 (remainder)
```

Since 8, the remainder, is less than one-half of 22 (the divisor) the quotient remains .86.

Another way to formulate this rule for rounding off numbers is as follows:

If the digit to be dropped from the quotient is less than 5, the preceding digit is unchanged. Thus:

$$.19/.22 = .863 = .86$$

Rule 2: If, in division, the remainder is more than one-half the number value of the divisor, the value of the last obtained digit of the quotient is increased by 1. Thus:

```
        .21
152. ) 33.0
       304
       ---
       260
       152
       ---
       108 (remainder)
```

Since 108 is more than one-half the value of 152, the quotient is written as .22.

This rule can also be reformulated as follows:

If the digit to be dropped from the quotient is more than 5, the preceding digit is increased by 1. Thus:

$$33./152. = .217 = .22$$

Rule 3: If, in division, the remainder is equal to exactly one-half the number value of the divisor,

(a) Leave the value of the last obtained digit of the quotient unchanged, *if the digit is even.*

(b) Increase the value of the last digit of the quotient by 1, *if the digit is odd.*

Thus, for (a):

$$
\begin{array}{r}
3.4 \\
.12\overline{)\,.41\,4} \\
\underline{36} \\
5\,4 \\
\underline{4\,8} \\
6 \text{ (remainder)}
\end{array}
$$

Since 6, the remainder, is exactly one-half of 12, and since the value of the last obtained digit of the quotient is even (4), it is unchanged.

And for (b):

$$
\begin{array}{r}
.55 = .56 \\
.70\overline{)\,.3885} \\
\underline{350} \\
385 \\
\underline{350} \\
35 \text{ (remainder)}
\end{array}
$$

Since 35, the remainder, is exactly one-half of 70, and since the value of the last obtained digit of the quotient is odd (5), the quotient becomes .56.

This rule for rounding off numbers can be reformulated more simply as follows:

If the digit to be dropped from a quotient is exactly 5 (followed only by zeros), make the preceding digit *even*. Thus:

.525000000 remains .52; .53500 becomes .54

This third rule is to be contrasted with the lay practice of always increasing the preceding digit by 1 when a 5 is dropped from the quotient. That the latter makes for a cumulative error is obvious from the following example:

Division: Even	Rounding Off to Two Decimals	
to 3 Decimals	Correct Method	Incorrect Method
.265	.26	.27
.105	.10	.11
.085	.08	.09
.205	.20	.21
.155	.16	.16
.145	.14	.15
.005	.00	.01
.035	.04	.04
$\Sigma = 1.000$	.98	1.04

The preceding example also illustrates the principle that the sum of the proportions of all the parts of a whole should be equal to unity. The method used in the last column is less accurate for rounding off the numbers to two decimal places than is the method used in the middle column. It should be

noted that with the recommended ("correct") method the sum of the eight parts is not exactly equal to unity. This is because there are more even than odd numbers in the hundredth column of digits. Except in such a circumstance, the sum of the proportions of all parts of a whole should, of course, exactly equal 1.0.

Dropping More Than One Digit

If in rounding off a number, several digits are to be dropped, the preceding rules still apply. The following numbers are rounded off to two decimal places, as indicated:

.47896 becomes .48
.47396 remains .47

.45550 becomes .46
.44550 becomes .45

.44500 remains .44
.44450 remains .44

.43500 becomes .44
.43499 remains .43

B. USE OF PERCENTAGES FOR COMPARING THE PARTS OF TWO OR MORE WHOLES

We have seen that the sum of the proportions of all parts (categories or classes) of a given whole should equal 1.0, or unity. The comparison of the parts of two or more wholes by proportions means that the data of the parts are reduced to a *common base of 1.0*. However, in practice, *percentage* values are more frequently used than proportions for comparing the composition of two or more groups. As we have seen, a percentage value is simply a proportion multiplied by 100. In other words, a percentage is a proportion taken to a *base of 100* instead of to a base of 1.0.

The question sometimes arises as to which "wholes" are to be compared and, therefore, what value is to be used in determining a base. Consider, for example, the data in Table 2:5 as presented in Table 3:1, with the attribute of ROTC membership omitted.*

The absolute number of students for each combination of College Class and Degree Group is given in Table 3:1, and the *percentage* of students of each College Class in each Degree Group is shown in Table 3:2. The latter therefore compares the proportion of different Degree students in each College Class. Whereas 10.2% of all Upper Seniors were Arts Degree students, only 8.3% of all Lower Juniors were Arts Degree students, etc.

* It is to be noted that, whereas Table 3:1 can readily be derived from Table 2:5 as a condensation of the latter, the reverse is obviously not possible. The advantage of a complete work table in classifying the original data of all attributes in an investigation thus lies in the time and labor saved in not having to go back to the individual record cards for the data of each comparison to be made.

Table 3:1. Classification of 2035 Students by College Classes and Degree Groups

Degree Groups	College Classes				Totals
	Upper Senior	Lower Senior	Upper Junior	Lower Junior	
Arts	49	47	40	52	188
Social Science	107	135	125	174	541
Science	155	173	152	199	679
Business Adm.	3	0	0	5	8
Technology	153	114	125	182	574
Education	12	14	8	11	45
Totals	479	483	450	623	2035

Table 3:2. Comparison of Four College Classes for Differences in Relative Size of Respective Degree Subdivisions
(Each College Class Taken to a Base of 100 Per Cent)

Degree Groups	College Classes			
	Upper Senior	Lower Senior	Upper Junior	Lower Junior
Arts	10.2%	9.7%	8.9%	8.3%
Social Science	22.3	28.0	27.8	27.9
Science	32.4	35.8	33.8	31.9
Business Adm.	0.6	0	0	0.8
Technology	31.9	23.6	27.8	29.2
Education	2.5	2.9	1.8	1.8
Totals	99.9%	100.0%	100.1%	99.9%
$N =$	[479]	[483]	[450]	[623]

If, on the other hand, it is relevant to compare the DEGREE differences of the four CLASS groups, the tabulation and percentages will be as indicated in Tables 3:3 and 3:4. Thus, according to the latter table, 26.1% of all Arts Degree upperclassmen were Upper Seniors, whereas only 19.8% of all Social Science Degree upperclassmen were Upper Seniors, etc.

Table 3:3. Classification of 2035 Students by Degree Groups and College Classes

Classes	Degree Groups					
	Arts	Soc. Sci.	Science	Bus. Adm.	Tech.	Educ.
Upper Senior	49	107	155	3	153	12
Lower Senior	47	135	173	0	114	14
Upper Junior	40	125	152	0	125	8
Lower Junior	52	174	199	5	182	11
Totals	188	541	679	8	574	45

Table 3:4. Comparison of Six Degree Groups for Differences in the Relative Size of Respective College-Class Subdivisions

(Each Degree Group Taken to a Base of 100 Per Cent)

Classes	Degree Groups					
	Arts	Soc. Sci.	Science	Bus. Adm.	Tech.	Educ.
Upper Senior	26.1%	19.8%	22.8%	37.5%	26.7%	26.7%
Lower Senior	25.0	25.0	25.5	0	19.9	31.1
Upper Junior	21.3	23.1	22.4	0	21.8	17.8
Lower Junior	27.7	32.2	29.3	62.5	31.7	24.4
Totals	100.1%	100.1%	100.0%	100.0%	100.1%	100.0%

Although the decision as to which attribute is to be laid off horizontally and which vertically is somewhat arbitrary, the arrangement in Tables 3:2 and 3:4 is the type ordinarily used. In both tables the wholes are subdivided into proportionate parts within the columns rather than in the rows. Therefore, the tables can be read horizontally (by rows) in comparing the proportions of members of each stratum in a given category of the substratum. Thus, it is readily observed from Table 3:2 that there is a constant decrease in the relative size of the Arts Degree group, beginning with the Upper Senior Class and going on across to the Lower Junior Class. Furthermore, the columns of this table give the proportionate composition by Degree groups of each College Class as a whole.

Sometimes the statistical frequency (symbolized by N, the number of cases) of each cell is included in tabulations like Tables 3:2 and 3:4. However, if the N's for each cell are not included, the total N's used for the *bases* of each whole should always be stated, so that the *base* for each percentage of a table will be clear. Moreover, if the base N is omitted and the reader is accordingly not informed of the sizes of groups being compared, he is likely to make absurd interpretations. In Table 3:4, for example, 62.5% of the Business Administration students were Lower Juniors, whereas the highest proportion of Lower Juniors in any other Degree Group was 32.2% (Social Science). But the total N for the Business Administration group was only 8; hence the percentage values for this Degree Group are more likely to be misleading than useful. In fact, percentage values derived from wholes that are composed of much less than 100 cases should always be interpreted with caution.

There is one further difference in the cross-tabulated layout of the two attributes, COLLEGE CLASSES and DEGREE GROUPS, in the preceding sets of tables which should be noted. The horizontal arrangement of College Classes in Tables 3:1 and 3:2 is in a sequence that has a *logical order*. That is, in going from Upper Senior to Lower Senior to Upper Junior to Lower Junior, the progression is from the college group with the most degree credits to the group with least. The order, then, is *from most to least.* "Concealed" within

these four categories of College Classes is a *variable* attribute, a variable for which the individual instances differ *quantitatively* in the number of degree credits per student at a given time. This variable, however, is ordinarily divided for convenience into a few categories and the data treated *as if* categorical.

The horizontal arrangement of Degree Groups in Tables 3:3 and 3:4, on the other hand, has no such logical order. The actual order employed is one corresponding in the main to the number notation for each degree used by the Registrar's Office; it is an order with a history but with no inherent logical structure. Therefore, for purposes of comparison, any other arrangement of the six categories of Degree Groups might have been used. For example, it might perhaps have been more useful to arrange these six categories in the order of the total number of frequencies in each. If this were done, the order would be:

	Sci.	Tech.	Soc. Sci.	Arts	Educ.	Bus. Adm.
N =	679	574	541	188	45	8

In any event, it should be clear that the arrangement of the four categories of College Classes corresponds to a scale or continuum characteristic of a variable, whereas in the arrangement of Degree categories there is no such inherent order ranging from *most to least*. The Degree subdivisions are categories of a non-variable attribute.

C. RATIOS AND INDEX NUMBERS

Per Capita Indices

Ratios are commonly used to bring together data in such a way that they can be readily compared by the use of a common "yardstick." For example, comparison of the costs of education in a school system from year to year, or among several school systems or among the states of the Union, can readily be made by a ratio that gives the *cost per capita*. In fact, per capita cost exemplifies one of the most common ratio techniques for indexing statistical information. Per capita cost is obtained by taking the ratio of the total cost to the number of individuals represented in the total cost. Thus, if a school system costs $115,000 a year to operate and the average daily attendance is 1000 pupils, the per capita cost is $115:

$$\frac{\text{Total educational cost}}{\text{Number of pupils}} = \frac{\$115{,}000}{1000} = \$115 \text{ per capita cost}$$

Table 3:5 summarizes the per capita cost of education for several suburban school systems near New York City, as reported for the school year of 1942–1943.

Table 3:5. Comparative Costs of Several School Systems
(Index: Per Capita Cost for 1942–1943*)

School System	Total Educational Costs	Number of Pupils (Average Daily Attendance)	Per Capita Cost
Mount Vernon	$2,397,336.81	9202	$260.52
Mamaroneck	1,013,516.36	3177	319.02
Great Neck	996,866.07	2980	334.52
Scarsdale	866,881.09	2072	418.38
Port Chester	756,407.93	3743	202.09
Garden City	763,985.23	1803	423.73
Pelham	627,975.04	1643	382.21
Manhasset	554,650.16	1706	325.12
Bronxville	483,624.96	1159	417.28
Peekskill	443,538.40	2205	201.15

It is apparent, from the data in Table 3:5, that the fairer way of comparing the educational costs of two school systems is in terms of per capita cost (last column), rather than absolute cost (Total Educational Costs, second column). The Mount Vernon School System cost the most, but it served over 9000 students and consequently ranked seventh of the ten systems in per capita cost. On the other hand, the Bronxville System ranked ninth in total cost, but served only 1159 students and was third in per capita cost.

Per capita ratios provide indices useful for the comparison not only of relative costs but also of the relative incidence of many social phenomena, such as crimes, diseases, various social services, etc.

Ratios as Index Numbers

Ratios presenting per capita costs are in reality *index numbers;* that is, they are values that not only index cost but index it with respect to a yardstick that is useful for the comparison of costs in two localities. Essentially, this is the purpose of index numbers, whether they are obtained in educational statistics or in economic statistics or other fields.

The main problem in the development of an index number is determining what the *base* shall be. Even in per capita costs, this problem is not always simple. It might appear to be a problem merely of counting the number of cases in the group being studied. However, the per capita costs of education cited in Table 3:5 were based on the *average daily attendance* rather than on the total enrollment. In order to get such a base—namely, average daily attendance—systematic records have to be kept by each classroom teacher during the entire school year. When two or more indexes are being directly

* Vernon G. Smith, *Cost Study and Salary Study: Cities and Villages of the Metropolitan Area*, Scarsdale, New York, 1943. Cf. also, *Fortieth Annual Report of the State Education Department of New York*, 1945, vol. 2, pp. 76–79, 148–151.

compared, it is obviously essential that they all be computed to the same base.

The I.Q. Index

A commonly used ratio, that yields an index of intellectual maturity, is the *intelligence quotient*, or *I.Q.* In the Stanford-Binet and other intelligence tests, this ratio is taken as equal to:

$$\frac{\text{Mental age}}{\text{Chronological age}} = \text{I.Q.}$$

An individual whose I.Q. is 1.00 is therefore one whose mental age (as derived from the Binet tests) is equal to his chronological age. If a child 8 years old has a Binet mental age of 10 years and 6 months, his I.Q. index is (years being converted to months):

$$\frac{10(12) + 6}{8(12)} = \frac{126}{96} = 1.31$$

On the other hand, if an 8-year-old child has a Binet mental age of $6\frac{1}{2}$ years, his I.Q. is:

$$\frac{6(12) + 6}{8(12)} = \frac{78}{96} = .81$$

The I.Q. index is often multiplied by 100 to eliminate the decimal in the ratio. The above index numbers would then be 100 (average intelligence), 131 (above average intelligence), and 81 (below average intelligence).

Standard Scores

A widely used index of relative ability is the *Standard score* (see Chapter 8). It is initially the ratio of a person's performance on a test (expressed as the difference between his test score and the arithmetic mean of the group) to a standard measure of the variability of the test taken in terms of a measure known as the standard deviation of the test scores of the group and symbolized by the Greek letter σ. This ratio is then converted to a scale whose mean equals 5.0 and whose standard deviation equals 1.0. Thus:

$$S \text{ (Standard score)} = \frac{X - M_x}{\sigma_x} + 5.0$$

If a person's score on an ability test is one standard deviation above the mean of the group, his Standard score index is therefore 6.0.

If $(X - M_x) = \sigma_x$

then $\frac{X - M_x}{\sigma_x} = 1.0$

and $1.0 + 5.0 = 6.0$

Index Numbers as Percentages

Many index numbers are expressed in terms of a percentage rather than as a proportion. The method of expression chosen is simply a matter of con-

venience in interpretation. For example, the index of the relative proportion of the sexes in a population, called the *sex ratio*, is usually expressed as a percentage, the percentage of males to females. This is illustrated by Table 3:6, containing population data for eight age groups in New York City.

Table 3:6. Sex Ratio for New York City by Age Groups *
(Based on Data of 1940 U.S. Census)

Age Groups	Males	Females	Proportion of Males to Females	Sex Ratio
Under 5 yrs.	221,415	212,479	1.042	104.2
5 to 9	238,798	231,758	1.030	103.0
10 to 14	283,453	277,655	1.021	102.1
15 to 19	300,717	306,225	.982	98.2
20 to 24	304,862	344,291	.885	88.5
25 to 44	1,302,761	1,383,554	.942	94.2
45 to 64	836,920	795,688	1.052	105.2
65 and over	187,367	227,052	.825	82.5
All ages	3,676,293	3,778,702	.973	97.3

The sex ratio in New York City for children under 5 years of age was 104.2 in 1940. This means that for every 100 girls there were approximately 104 boys. On the other hand, the sex ratio was 88.5 for young men and women in the 20–24 age group. This means that at this age there were but 88.5 men for each 100 women. The sex ratio for all age groups, given in the bottom row of the table, was 97.3, which means that for every 100 women there were approximately 97 men.

D. CONFUSION IN THE USE OF PERCENTAGES

Misinterpretations or confusion in the use of percentages is likely to occur unless the research worker is aware of such possibilities and is careful to avoid them. Some of the most common errors in using percentages are illustrated in the following examples.

Confusion in Interpreting a Percentage Increase

In a recently published study, "How Leading Questions Determine Answers," the following results were reported for two field survey questions. Half of the respondents were asked:

> As you know, this war is costing a lot of money. Do you think that advertising in wartime is a *necessary* or *unnecessary* expense?
>
> 42% said Necessary
> 38% said Unnecessary
> 20% said Don't Know

* Sixteenth Census of the U.S., *Population*, 1943, vol. 4, pt. 3, p. 663.

The alternative question asked of the other respondents was:

Do you think advertising in wartime is *necessary* or *unnecessary*?

64% said Necessary
26% said Unnecessary
10% said Don't Know

These results were misinterpreted as follows: "Here [for the second question] the per cent of answers in favor of advertising is 22% higher." As an inspection of these data shows, the *difference* in the percentage figures between the 64% saying "Necessary" to the second question and the 42% saying "Necessary" to the first question was 22 per cent points. However, the percentage increase is in reality the ratio of:

$$\tfrac{64}{42}(100) = 152.4\%; \quad \text{and} \quad 152.4\% - 100\% = 52.4\%$$

This is considerably different from "22% higher." The percentage of respondents in favor of advertising was 52.4% higher on the second question than on the first. The ratio of answers was thus about $1\frac{1}{2}$ to 1.

A Percentage Decrease Can Never Be More Than 100%

This is so obvious that little or no reflection should be necessary. However, this misinterpretation is sometimes made. Thus, if an article regularly costs \$1.00 and the price is increased to \$2.00 and then later reduced to 90 cents, one might in error interpret this as a 110% decrease in price. Actually, of course, the decrease from the price of \$2.00 is 55%. The base for the decrease has to be the higher cost value, not the original cost value of \$1.00.

$$\frac{.90}{2.00}(100) = 45\%; \quad \text{and} \quad 100\% - 45\% = 55\% \text{ (the decrease)}$$

Another example of this type of misinterpretation: A person may have increased his *accuracy* score in a card-sorting test, say from an average of 16 errors per trial to an average of only 8 errors per trial. This would be a *gain* in the accuracy score of

$$\tfrac{8}{16}(100) = 50\%$$

Let us now assume that his accuracy score reverts to 24 errors per trial with the experimental introduction of a "distraction." Even though there is a threefold increase in errors, the decrease in accuracy is *not* 300%

$$\tfrac{24}{8}(100) = 300\%$$

but rather 67%:

$$\tfrac{8}{24}(100) = 33\%, \quad \text{and} \quad 100\% - 33\% = 67\%$$

The base for the percentage decrease is 24 instead of 8.

Confusion Between Percentages and Proportions

Values of less than 1% are sometimes written to two decimal places without a zero being placed to the left of the decimal. Thus, one-half of one per cent

is sometimes written as .50. That this may readily be confused with the proportion of .50 is apparent. Such confusion will be avoided if a zero is always placed on the left of the decimal for any figure which is a fraction of one per cent. Thus, one-half of one per cent should be written as 0.5%.

Confusion from Large Percentages

An incapacity to grasp the implications of very large percentage values is another frequent cause of confusion. Thus, a population of 1,000,000 people is 5000% as great as a population of 20,000:

$$\frac{1{,}000{,}000}{20{,}000}(100) = 5000\%$$

In such a case, it is usually better to indicate that the larger group is 50 times as great as the smaller, rather than 5000% as great.

Percentages from Too Small a Base

Misleading conclusions are likely to result from percentages computed from too small a base. As previously pointed out, the expression of a ratio as a percentage usually implies at least 100 members in the total group for which the proportion is computed. Examination of Table 3:4 indicates that $62\frac{1}{2}\%$ of the Business Administration Degree students were Lower Juniors, that none were Upper Juniors or Lower Seniors, and that $37\frac{1}{2}\%$ were Upper Seniors. On the face of it, these appear to be astounding differences. However, as indicated in the table, these percentages are based on only eight cases, and consequently the differences are not as significant as they appear.

Errors in Averaging Percentages

A fallacious result follows the averaging of percentages unless the size of each group being averaged is taken into consideration. In other words, an average of percentages needs to be weighted by the respective size of each group from which the percentage figures are derived. Let us assume that the following four groups of respondents expressed a liking for a motion picture according to the percentages indicated for each:

Group	Percentage Liking Motion Picture
Men	40%
Women	60
Boys	10
Girls	50
	Total = 160%
	Average = $\frac{160}{4}$ = 40%

If each of these four groups of respondents is the same size, then the average of 40% is correct. But if the sizes of the groups vary, we cannot be sure that

40% is a satisfactory average. Each group needs to be weighted by its size and the average computed from the weighted percentages, as follows:

Group	Percentage Liking Motion Picture	N (Size of Group)	Percentages Weighted by N
Men	40%	1000	40,000
Women	60	1000	60,000
Boys	10	500	5,000
Girls	50	500	25,000
Total	160%	3000	130,000

With this information, a correct average can now be obtained to represent the percentage of all respondents liking the motion picture. It is as follows:

$$\frac{\text{Total percentages weighted by } N\text{'s}}{\text{Total } N\text{'s}} = \frac{130{,}000}{3000} = 43.3\%$$

E. GRAPHIC METHODS FOR THE PRESENTATION AND COMPARISON OF CATEGORICAL DATA

During the past ten years, the art of portraying statistical information has expanded and flourished. That one graph or picture of a statistical result is "worth a thousand words," as the saying goes, serves to emphasize the psychological value of a picture that drives home through the immediate comprehension of the eye a set of otherwise dull or uninteresting statistical tabulations. The art of graphic representation has penetrated our daily press, and magazines and pamphlets for the lay public. Today children in the secondary schools—even in some elementary schools—not only are familiar with statistical portraits but also make them. Particularly popular are graph and pictorial techniques for categorical data.

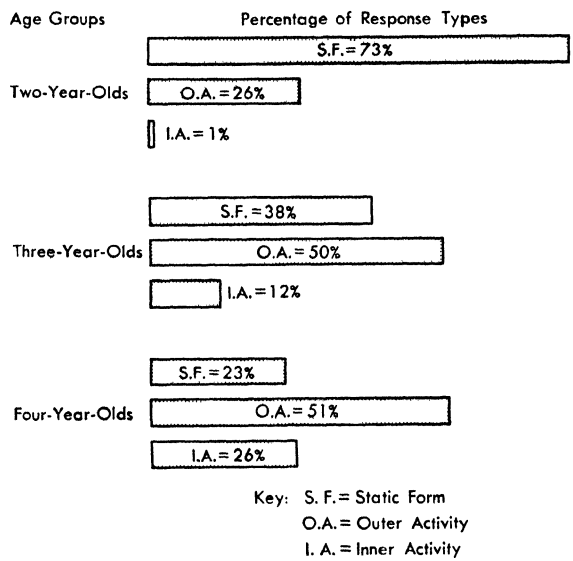
The purpose of a graph or chart is to tell a story in a simple and intelligible but striking manner. So far as possible, the story should be complete. That is, the reader should need to have little recourse to text material or to a table of statistical information in order to understand the meaning of a graph. A graph should therefore have a descriptive title and the details should be adequately labeled so that the statistical results to be conveyed the reader can be understood immediately.

The chief types of graphs or charts for categorical data are (1) bar charts, (2) belt graphs, (3) pie diagrams, (4) maps, and (5) pictorial charts. In addition, line graphs and belt graphs are used to portray trends in categorical data over a period of time. With respect to the passage of time, the data of a non-variable attribute per se may thus become a variable, as for example the changes from year to year in the number or proportions of the people in a geographical area engaged in various occupations, etc. Learning curves are examples of *time series*, for which line graphs are usually employed.

Bar Graphs

The simplest but not always the most interesting type of graph for categorical data is the bar graph. The bars may be drawn vertically or horizontally. Their length is scaled according to the number of frequencies or percentages of the categories to be shown. Their width is mainly a matter of aesthetics; that is, the determination is governed by what pleases the eye rather than by any logical rule other than uniformity of width for the several bars of a given chart.

Fig. 3:1. Comparison, by Age Groups, of Pre-School Children's Types of Verbal Responses to Pictures (Amen's Data)

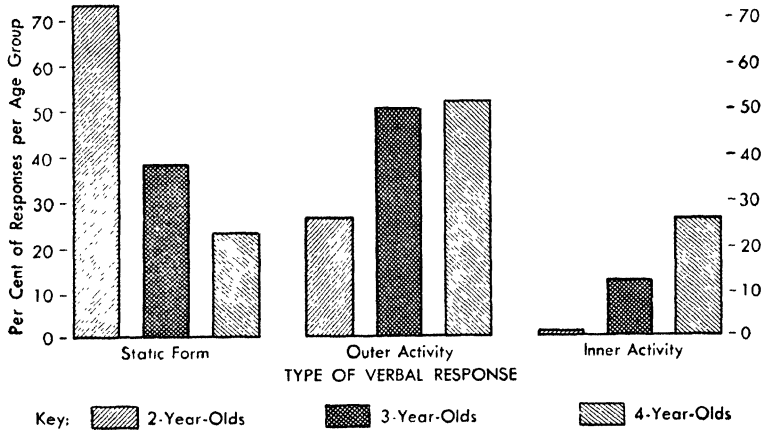


Figs. 3:1 and 3:2 are horizontal and vertical bar graphs, respectively, for Amen's data in Table 2:1. Both graphs are based on the same data and each tells a similar all-over story, but with a different emphasis. The horizontal bars in Fig. 3:1 are arranged by successive age groupings to emphasize the differences in the total composition by response types for each age group. The vertical bars in Fig. 3:2 are arranged by successive response types in order to emphasize the different proportions of the three types of response among the three age groups.

Sometimes the spaces within the rectangles in bar charts are utilized for descriptive phrases and notations, as in Fig. 3:1. Or the rectangles may be cross-hatched or shaded in order to differentiate subdivisions within categories or to contrast categories themselves; their length is scaled at the side, and

the descriptive phrases for each subdivision are *keyed* on the chart, as in Fig. 3:2.

Fig. 3:2. Comparison of the Relative Incidence of Types of Verbal Responses to Pictures Among Pre-School Children (Amen's Data)



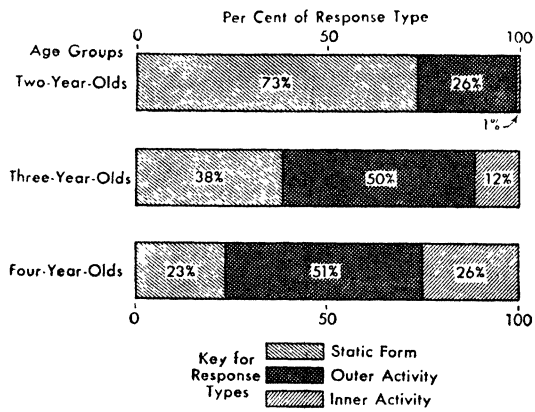
These types of charts are used for statistical frequencies as well as for proportions. When contrasting *colors* can be employed for the different categories or subdivisions to be compared, the result is considerably more effective than the black-and-white patterns used in these figures. When black-and-white patterns have to be employed in photographic or printed reproductions, the best effects can be obtained by using already prepared paper that comes in a great variety of patterns and can be cut to fit the surface area that is to be shaded or cross-hatched. Such prepared paper was used in these, as well as in several of the ensuing figures.

Bar Charts for the Proportions of Wholes

A further bar chart technique for the graphic comparison of Amen's data on pre-school children's responses to pictures is shown in Fig. 3:3. This type of graph is useful for comparing the proportionate composition of two or more groups (wholes) (a) when the number of instances (N) is the same for each, or (b) when the data in each categorical group have been reduced to percentages of the whole. Fig. 3:3 exemplifies the latter. The length of each horizontal rectangle is taken as equal to 100% and the proportions of each type of response are laid off to scale and differentiated by contrasting patterns within the area of the rectangle. Although pie charts, as we shall see, are also commonly employed to differentiate the proportionate parts of a whole, rectangles or bar charts are more suitable for the types of comparisons made in Fig. 3:3. It is easier to *compare* the composition of several cate-

gories shown in a series of rectangles in the same vertical plane than when the categories are shown in several pie charts.

Fig. 3:3. Comparison of the Relative Incidence of Types of Verbal Responses to Pictures Among Two-Year, Three-Year, and Four-Year-Old Pre-School Children



Of the three types of bar charts for Amen's data shown in these three figures, the last is the most effective in the way in which it tells the whole story and emphasizes both (a) the response-type composition of each age group (reading across the chart) and (b) a comparison of the proportionate incidence of each response type for the successive age groups (reading down the chart).

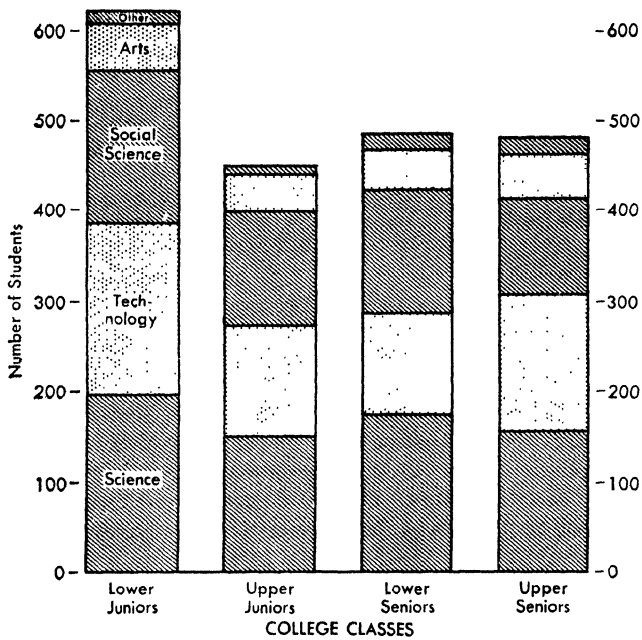
Bar Trend Graphs

Charts of the type shown in Figs. 3:4 and 3:5 emphasize the change in the composition of successive categories. These are trend graphs and their construction is based on the bar chart principle. They may be employed with either frequencies or proportions. The categories in these particular charts are the four groups of City College upperclassmen. The changes in the composition of each College Class by students' degree objectives are portrayed on the basis of the data in Tables 3:1 and 3:2. Because of the few students choosing Education or Business Administration as a degree objective, these two subdivisions have been combined into "Other" at the top of each rectangle.

Fig. 3:4 is based on the actual frequencies of each category and its subdivisions. Hence the changes described are *absolute*, rather than relative to the size of each category. The frequencies are scaled at the left and right of the chart. The choice as to which subdivision is to be represented at the base of the rectangles, which next, etc., is of course arbitrary. However, once the

choice is made for the first one (Lower Juniors), the same order must be maintained throughout. The order of the arrangement of the five degree-objective subdivisions in this figure is based on the size of each in the Lower Junior Class: the Science subgroup had the most students; Technology was

Fig. 3:4. Student Composition of Upperclass College Groups According to Their Different Degree Objectives



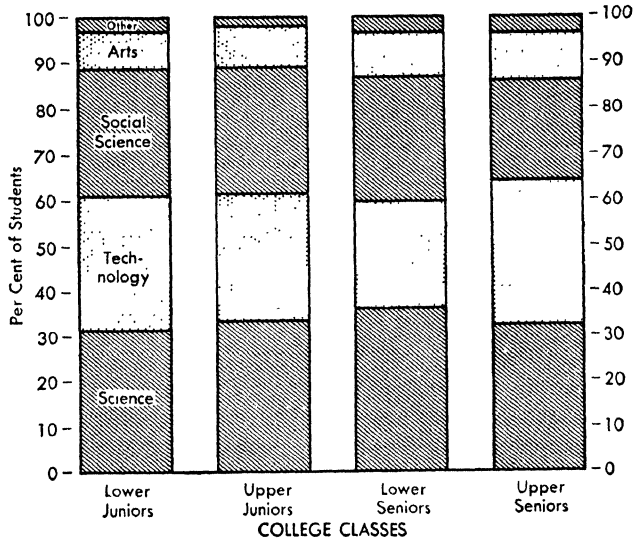
next; Social Science third, etc. To facilitate the actual drawing of the rectangles to the frequency scales at the left and right, the number of students per subdivision for each Class can be *cumulated* as follows (for the Lower Juniors): Science = 199; Science 199 + Technology 182 = 381; Science 199 + Technology 182 + Social Science 174 = 555; these 555 + Arts 52 = 607; and these 607 + the 16 "Other" = 623.

It will be observed that the emphasis in Fig. 3:4 is on *gross* changes in frequencies for the degree subgroups of the four categories of Classes. For a more accurate comparison of the change in frequencies, the original data (Table 3:1) must also be employed. Figure 3:4 gives an over-all view of the situation but shows at a glance whether there are any marked changes in the situation from one Class to another.

Fig. 3:5 is similar to Fig. 3:4 in that it also indicates the changes in the

composition of the degree subgroups for the four categories of upperclassmen. This time, however, the *relative* changes are emphasized by the use of proportions (percentages) instead of frequencies. Each of the four Classes is used as the base, 100 per cent, and the proportions of students in each subdivision (degree objective) are indicated. Despite the greater absolute size of the Lower Junior Class, as seen in Fig. 3:4, it is apparent from Fig. 3:5 that *relatively* there is little change in the degree-objective composition of the four upperclass groups of the 2035 City College students.

Fig. 3:5. A Comparison of the *Relative* Composition of Upperclass College Groups According to Their Degree Objectives



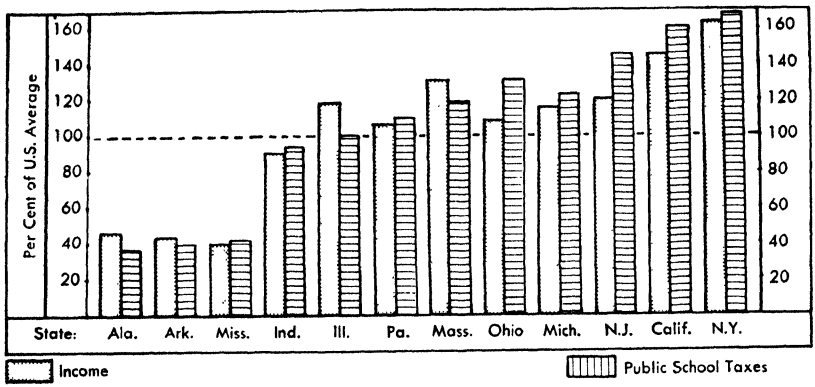
Bar Graphs for Relationships

Fig. 3:6 illustrates the use of vertical bar graphs in comparing two sets of percentage values for several categories—in this case, 12 states in the Union. The relationship between per capita income and per capita tax expenditures for public school education is shown. Each is taken as a percentage of the United States average as the base (projected at 100% on the chart by the horizontal broken line).

It is apparent that the relationship between income and tax expenditures for public school education in these 12 states is strikingly close. Thus, per capita income in Alabama, Arkansas, and Mississippi is about 40% of the U.S. average, as is also per capita tax expenditure for public school education. In New York State, at the other extreme, per capita income is about 160%, and

per capita tax expenditure for public school education about 165%, of the U.S. average.

Fig. 3:6. Showing the Relationship Between Income and Public School Tax Expenditures per Capita in New York and Other States, 1935 *



* From "Public School Costs in New York and Other American States," *Public Education Information Bulletin*, vol. 14, No. 2, 1939. New York State Teachers Association, Albany, New York. Reproduced by permission of A. J. Burke, Director of Studies.

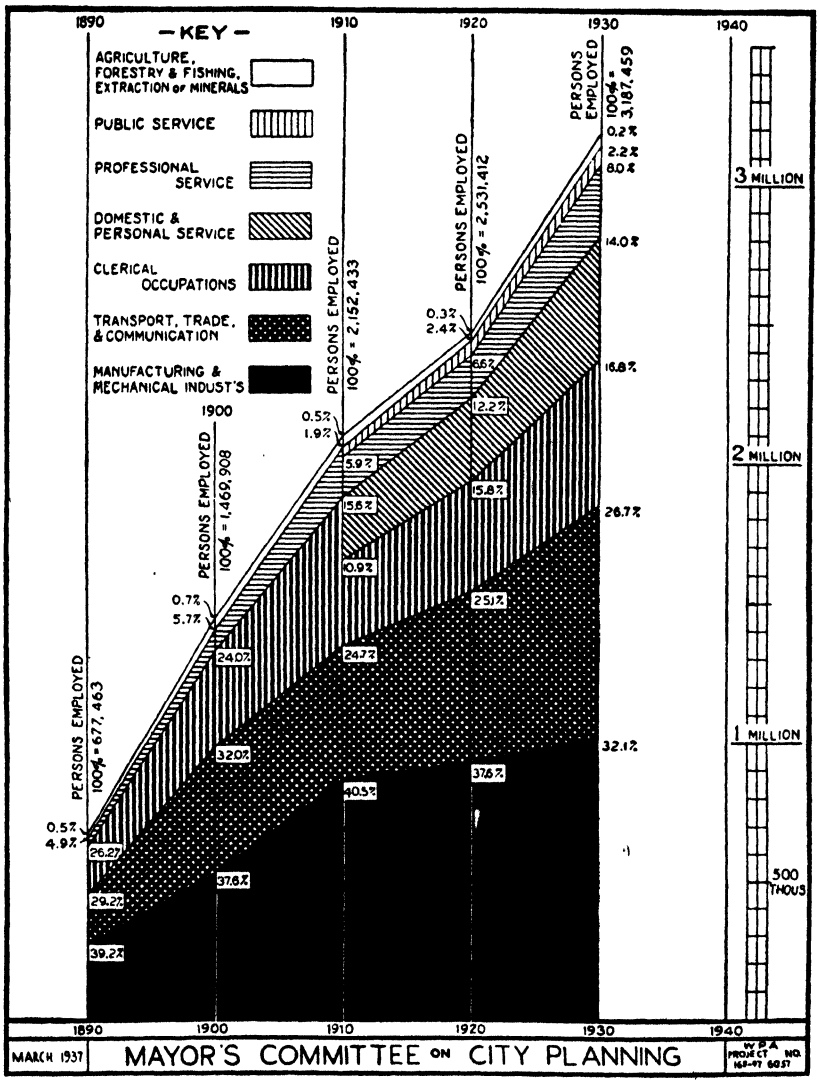
Belt Graphs

When emphasis in a chart is to be focused on changes in the composition of a category that occur over a period of time or in continuously successive stages, as is characteristic of *time series* (changes in an attribute per unit of time over a period of time), the bar trend graphs in Fig. 3:4 and 3:5 are converted into belt graphs of the type illustrated in Fig. 3:7 and 3:8. Instead of separated rectangles being used, the continuity and trend of the statistical information are emphasized by plotting line graphs and shading with contrasting patterns the areas of each subdivision.

Fig. 3:7 shows the trend in the occupational composition of New York City's population of employed persons over a period of five decades. Successive decades are plotted on the base line and at the top of the chart. The frequency scale of persons employed is drawn in units of 100,000 at the right. Each major occupational category is keyed in the upper left-hand corner of the chart. The total number of employed persons each decade is stated on the chart and is taken as the base for the percentages keyed on the body of the chart. Thus both the absolute and the relative changes are presented. The category, Domestic and Personal Service, was not differentiated from Clerical Occupations until 1910, and Public Service was also differentiated at that time. The greatest absolute increase in employed persons was in the Manufacturing and Mechanical Industries, and the greatest relative increase

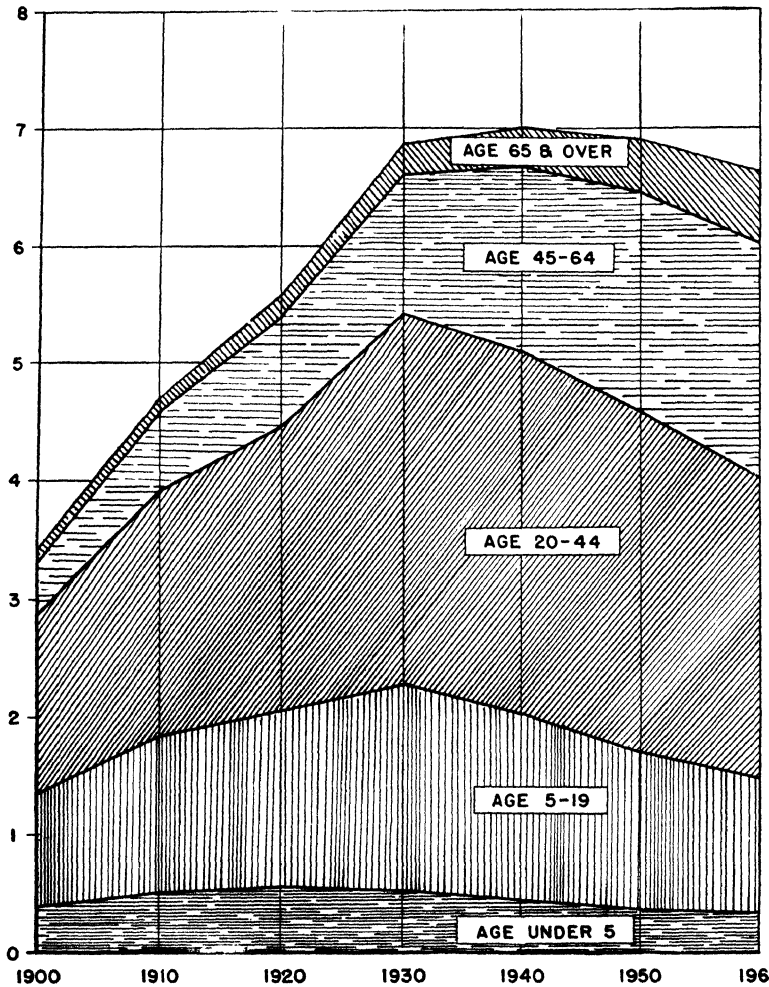
was for the original category that combined Public Service and Professional Service prior to 1910.

Fig. 3:7. Distribution of Persons by Occupational Groups, New York City, 1890–1930*



* Reproduced by permission of the Institute of Public Administration, New York City.

Fig. 3:8. Age Composition of Population, New York City*



* Reproduced by permission of the Institute of Public Administration, New York City.

Fig. 3:8 is a belt graph showing changes and trends in the age composition of New York City's population since 1900. Although AGE is a variable attribute, the original data have been organized into five broad categories (or class intervals) and the trend of each is indicated in frequencies, scaled in millions at the left of the chart.†

† This particular graph, made in the '30's, is interesting because of its projection of the anticipated composition of the population in 1940, 1950, and 1960. The 1940 U.S. Census

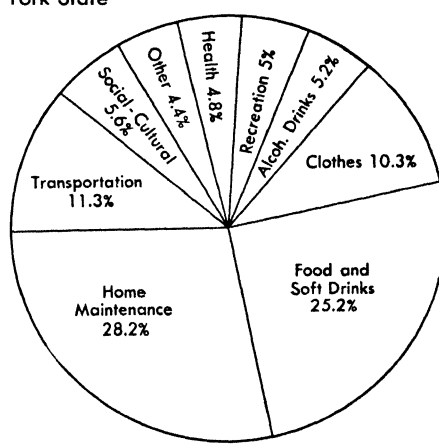
AGE is an attribute that figures importantly in practically all polls of public opinion. Therefore knowledge of a population's age composition by broad categories is essential to the stratification of this variable in sampling statistics.

Pie Diagrams

Circular charts have long been used to portray the relative size of the parts of a whole. Each category, or subdivision, is represented by a pie-shaped piece (a sector of a circle) drawn to give an area equal to each category's appropriate share of the whole area.

In Fig. 3:9 a pie diagram is used to show the relative importance of consumer expenditures in New York State prior to World War II. The name of each category and its relative size are indicated within each sector. A semicircular protractor scaled in percentages as well as in degrees facilitates the construction of such a chart. If percentage calibrations are not available on a protractor, the percentage values of each category or subdivision must be converted into degrees. For example, 25% would be represented by a sector with an angle of 90° [since .25(360°) = 90°], which is a quadrant of a circle.

Fig. 3:9. Showing the Relative Importance of Various Consumer Expenditures in New York State*



* From "Public School Costs in New York and Other American States," *Public Education Information Bulletin*, vol. 14, No. 2, 1939. New York State Teachers Association, Albany, New York.

If frequencies rather than percentages are to be represented on a pie chart, the size of the angle for each sector can readily be determined by multiplying 360° by the proportionate value of each sector's frequency to the total number of frequencies (N).

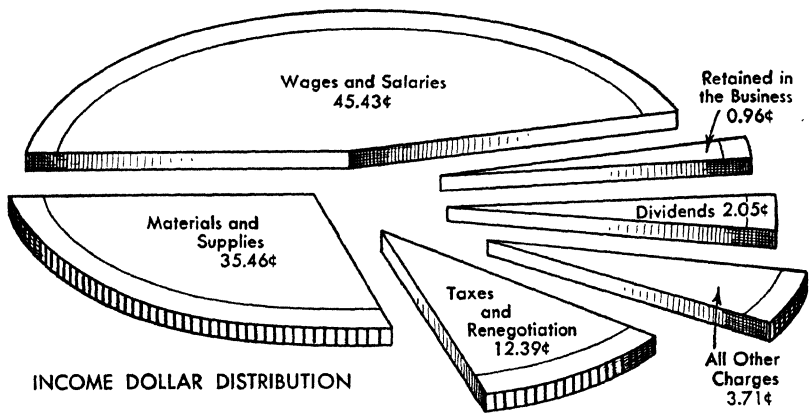
Pie diagrams are useful for comparative purposes when the sectors represent proportions of the whole. However, if subdivisions of *frequencies* of two or more groups of data are to be compared and N (the total number of frequencies for each group) varies considerably, the proper construction of the diagram becomes somewhat complicated because the relative size of the area of each circle must be proportionate to the N's of each group compared. If group A is twice the size of group B, the *area* (not the radius) of circle A needs

returns gave New York City a total population of nearly seven and a half million (in contrast to the predicted seven million).

to be twice as great as that of circle B. For such comparisons it is better to use the type of bar chart shown in Fig. 3:4, or the technique used for Fig. 3:8.

The pie diagram in Fig. 3:10 illustrates one of the most common graphic devices used to show how the money of an organization is spent. In this particular chart, emphasis is given to the relative size of each expenditure cate-

Fig. 3:10. Income Dollar Distribution



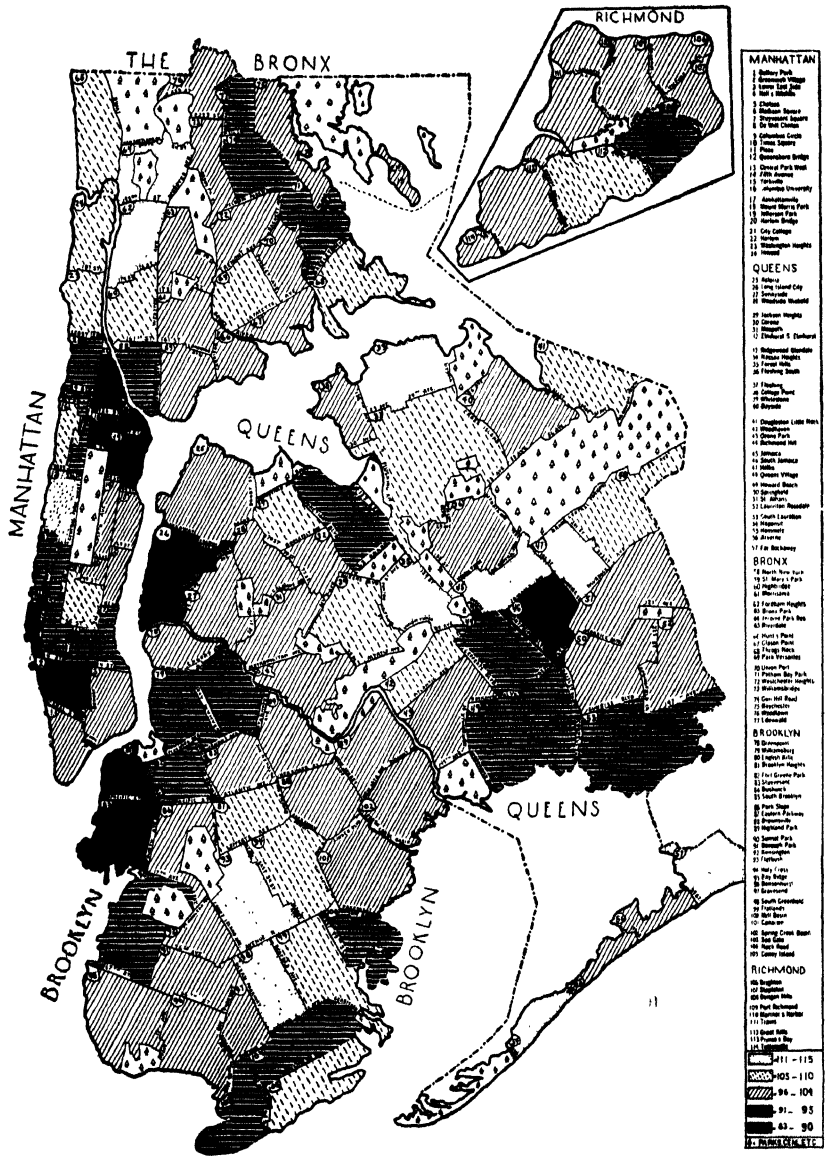
gory by the separation of each sector. The catchall category, common to the procedures of classification and division, is labeled "All Other Charges." The base taken for the divisions of the total income is *one dollar* and the circle is portrayed in three dimensions, the edge being milled to represent a coin. Thus for each dollar spent, 45.43 cents went for wages, and slightly less than one cent was retained in the business.

Maps

Maps have long been used in economic statistics for the protrayal of statistical information related to geographical areas but are used less frequently for data of a psychological character. Figs. 3:11 and 3:12, however, are illustrative of the potentiality of maps for the presentation of psychological and educational information. Both are maps of the five boroughs of the City of Greater New York, which are under the jurisdiction of a single Board of Education.

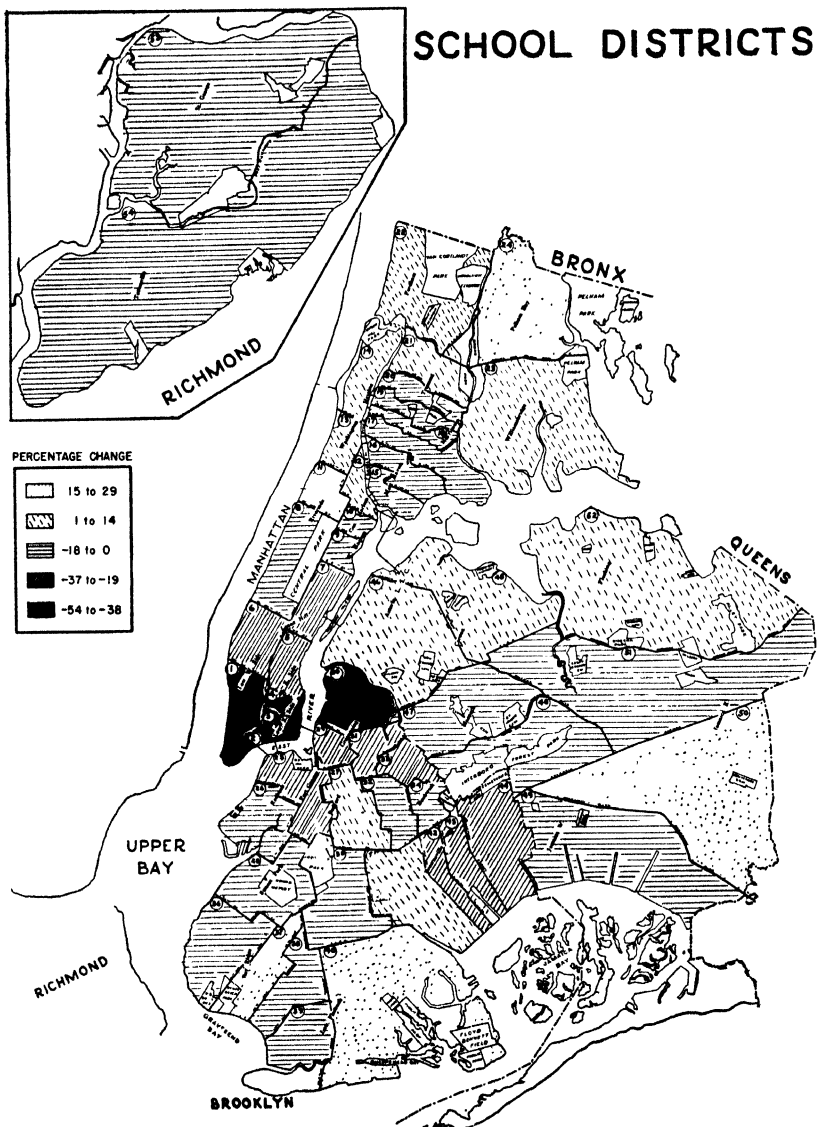
The *average* intelligence quotients of fifth-grade public-school children in 114 districts of Greater New York are differentiated into five categories, described in the key at the lower right-hand corner and portrayed on the map in Fig. 3:11. Familiarity with the industrial, business, and residential areas of Greater New York as they existed more than a decade ago would of course add considerably to the information to be derived from this map.

Fig. 3:11. Average Intelligence Quotients of Fifth-Grade Public School Pupils, New York City, 1932*



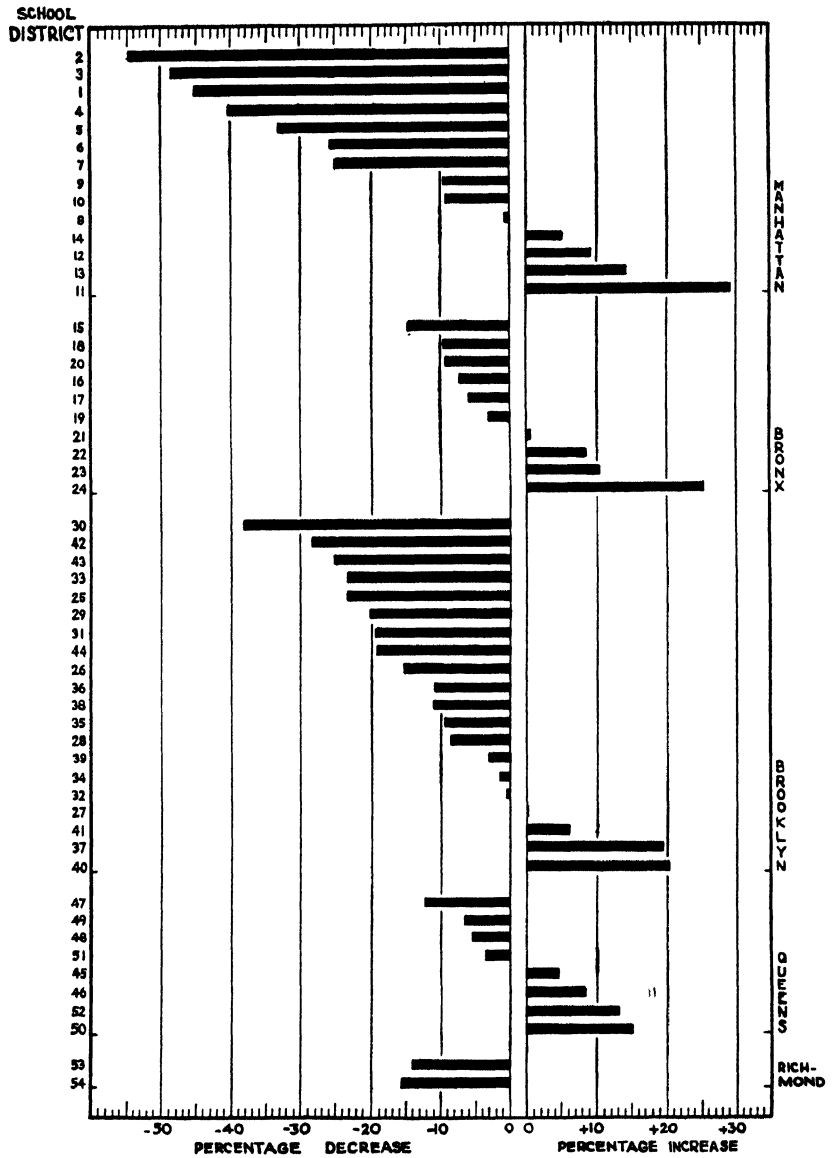
* Reproduced by permission of the Institute of Public Administration, New York City.

Fig. 3:12. Percentage Change of Enrollment in Elementary and Junior High Schools, 1929–1939, New York City*



* Reproduced by permission of the Institute of Public Administration, New York City.

Fig. 3:13. Percentage Change of Enrollment in Elementary and Junior High Schools, 1929–1939, New York City *



* Reproduced by permission of the Institute of Public Administration, New York City.

However, it is apparent that the five categories of I.Q. averages are represented in all the boroughs except Richmond (Staten Island) and that they are fairly well scattered in each borough. A high-average I.Q. district may be adjacent to a low one; thus District 47 in Queens has a high-average I.Q., whereas District 46 is low.

Aside from the proper layout of statistical information on the map, the main problem in its construction is (1) to establish appropriate geographical subdivisions and (2) to obtain the necessary descriptive (or sampling) statistics for each. The geographical subdivisions in the map in Fig. 3:12 are the 54 school districts of Greater New York, whereas the 114 subdivisions in Fig. 3:11 were set up in terms of smaller districts on the basis of the elementary schools serving them.

The map in Fig. 3:12 was constructed to portray the percentage *changes* of enrollment in elementary and junior high schools in Greater New York over the decade 1929 to 1939. Five categories of change are keyed at the left center of the chart. The trend of families away from the industrial and business sections of lower Manhattan (Districts 1, 2, 3, and 4) and from District 30 across the East River in the Borough of Queens is striking, as is the trend toward an increase in outlying districts in Brooklyn, Queens, and the Bronx.

In order to summarize the information on the map in Fig. 3:12, an additional graph is desirable. The bar chart shown in Fig. 3:13 is suitable for the purpose. The predominantly decreasing trend in enrollment in all five boroughs, and the relative extent of both decreases and increases are immediately apparent from an inspection of this chart.

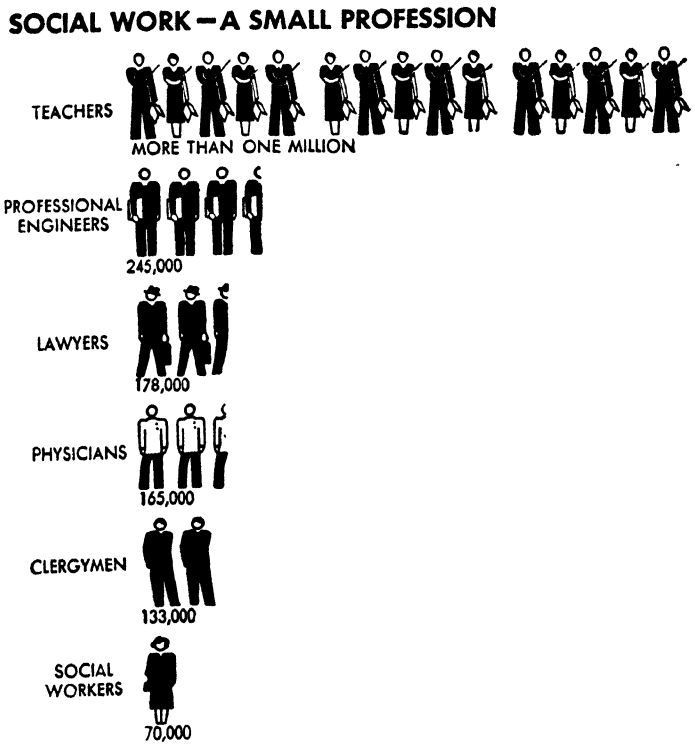
Maps that describe differences in the socio-economic character of neighborhoods in a city or in suburban and rural areas are used increasingly today in sampling statistics for the study of people's attitudes and opinions. When such maps are carefully constructed, a fraction of all the geographical subdivisions for each type of category can be sampled and the entire population satisfactorily studied from the information derived from only a small part.

Pictorial Charts

Pictorial charts are charts in which pictures or designs are employed that directly symbolize something essential about the character of the categories to be described or compared. Such charts are extensively used in the lay press to make interesting what would otherwise be uninteresting statistical information to the average reader. They are of special value in emphasizing the over-all result, bringing out contrasts, and showing the relation of the parts within a whole. Typical examples are illustrated in Figs. 3:14–3:19. The statistical information in each of these figures could readily be graphed in simple bar charts or pie diagrams. However, statistical charts obviously have greater interest value and attract and hold the eye, when symbols are em-

ployed for people in different professions (Fig. 3:14), for different types of motivation (Fig. 3:15), for the earnings of workers in different countries (Fig. 3:16), for different types of occupations (Fig. 3:17), for different types of adjustive behavior (Fig. 3:18), and for differences in public opinion (Fig. 3:19).

Fig. 3:14. Social Work and the Joneses*



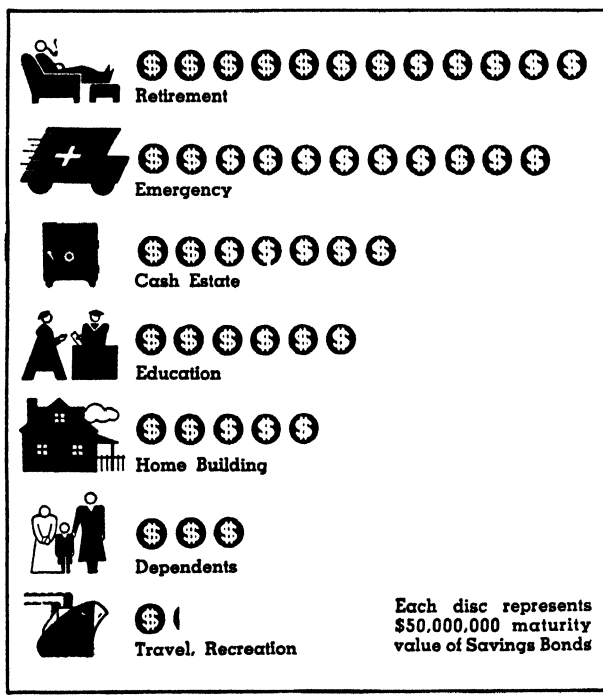
* From Public Affairs Pamphlet, *Social Work and the Joneses*, by Ruth Lerrigo and Bradley Buell. Published by the Public Affairs Committee, Inc., New York City.

The relative size of social work as a profession in relation to five other professions is strikingly brought out in Fig. 3:14. Each person symbolizes roughly 75,000 professional workers, but the rounded count is also given for each category. This pictorial chart makes vivid what otherwise might be a horizontal bar graph of frequencies.

Different classes of motives for the purchase of U.S. Savings Bonds are portrayed in Fig. 3:15. However, the size of each category is given in terms of

the maturity value of the bonds rather than in terms of the number of people or families buying them. Dollar discs are therefore employed to represent the amount of investment made by each group. This chart, like the preceding, is a pictorial substitute for what might otherwise be a simple horizontal bar graph.

Fig. 3:15. Reasons Given by Individual Owners for Systematic Saving Through Savings Bonds—and Amounts Invested*

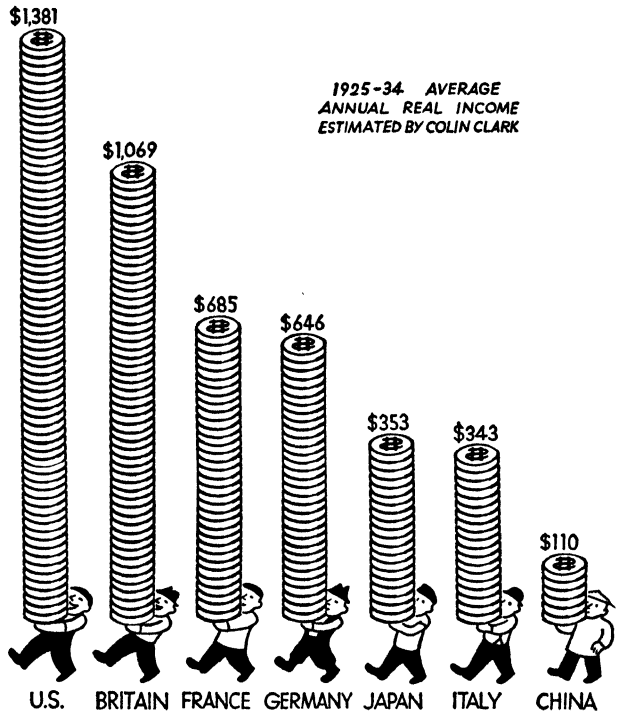


* From *The Graphic Story of United States Savings Bonds*, Pamphlet of the U.S. Department of Treasury, 1939.

A comparison of the average annual real income of workers in seven different countries is made possible by the pictorial chart in Fig. 3:16, a comparison for which a simple vertical bar graph might otherwise be used.

Fig. 3:17 shows a pictorial device for emphasizing the character and size of the proportionate parts of a whole, ordinarily shown by a pie diagram. The statistical base is taken as 100 and consequently the data of each category can readily be interpreted in percentages. Thus, at the beginning of 1942, 4 per cent of the "Labor Force" in the United States were in the Armed

Fig. 3:16. How Much Does a Worker Earn?*



* From Public Affairs Pamphlet, *What Foreign Trade Means to You*, by Maxwell S. Stewart. Published by the Public Affairs Committee, Inc., NewYork City.

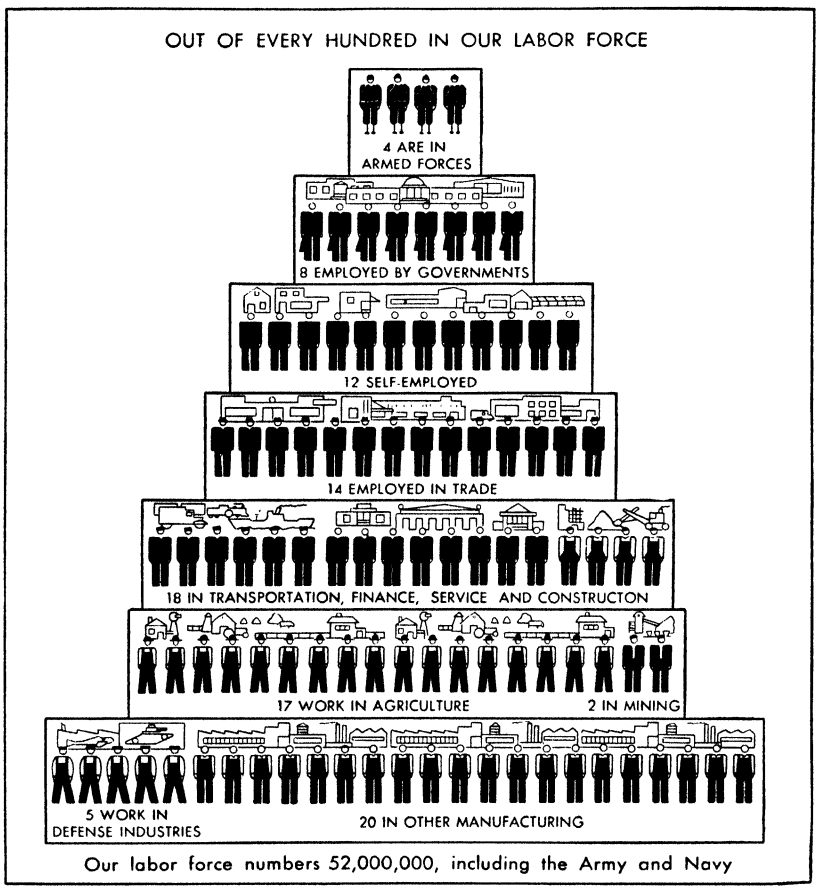
Forces; 8 per cent were in government service, etc. Our total labor force at the time was 52 million, according to the last line of the chart.

The extent to which epileptics can make an occupational adjustment is portrayed in Fig. 3:18. In this pictorial chart, 20 such persons are taken as the statistical base and the behavior of epileptics is differentiated into four categories. A pie diagram or a bar chart like that shown in Fig. 3:3 might otherwise have been used for a less striking presentation of this statistical information.

In Fig. 3:19 a pictorial device is combined with a statistical chart, viz., a series of grids that portrays the percentage *volume* of replies to a question used in a Fortune Survey of public opinion by Elmo Roper. The character of the replies is tabulated in percentages at the top of the chart, as is customary in reporting such survey data. These figures are then dramatized by both

the grids and the pictures. This particular survey question is an interesting historical "memento" of American public opinion in 1940.

Fig. 3:17. Out of Every Hundred in Our Labor Force . . .*

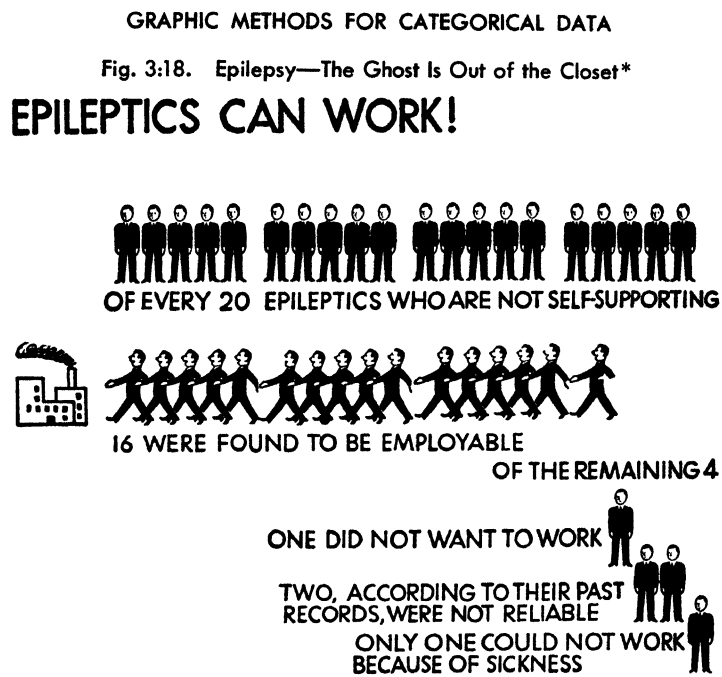


* From the New York *Times Magazine*, February 15, 1942. Reproduced by permission of the New York *Times* and the Pictograph Corporation.

EXERCISES

1. Given a total group of 385 people, 110 of whom are women, 125 men, 80 girls, and 70 boys, state the relation of each of these four sub-groups to the whole in terms of (a) proportions, (b) percentages.
2. In the preceding example, what proportion of the total group consists of adults? Of boys and girls?

Fig. 3:18. Epilepsy—The Ghost Is Out of the Closet*



* From Public Affairs Pamphlet, *Epilepsy—The Ghost Is Out of the Closet*, by Herbert Yahraes. Published by the Public Affairs Committee, Inc., New York City.

3. Round off the following numbers to two decimal places:
   a. 43.4083
   b. .11113
   c. 106.556
   d. 1.8555
   e. 3.1645

4. Set up a table for the following data, compare the results of each sub-group in terms of percentages, and give your reasons for choosing the base or bases for the percentage comparisons:

   A group of 2000 people is composed of 1000 men and 1000 women. Of the men, 446 voted for the Democratic candidate in the last election; 403 voted for the Republican candidate; 76 voted for other candidates; and the remainder did not vote. Of the women, 425 voted for the Democratic candidate; 437 for the Republican candidate; 42 for the other candidates; and the remainder did not vote. Of the men who voted for the Democratic candidate, 200 were over 35 years of age. Of the men voting for the Republican candidate, 250 were over 35 years of age. Of the men voting for the other candidates, 32 were over 35 years of age. And of the men who did not vote for any candidate, 60 were over 35 years of age. Of the women who voted for the Democratic candidate, 195 were over 35 years of age. Of the women who voted for the Republican candidate, 220 were over 35 years of age. Of the women who voted for other candidates, 10 were over 35 years of age. And of the women who did not vote for any candidate, 75 were over 35 years of age.

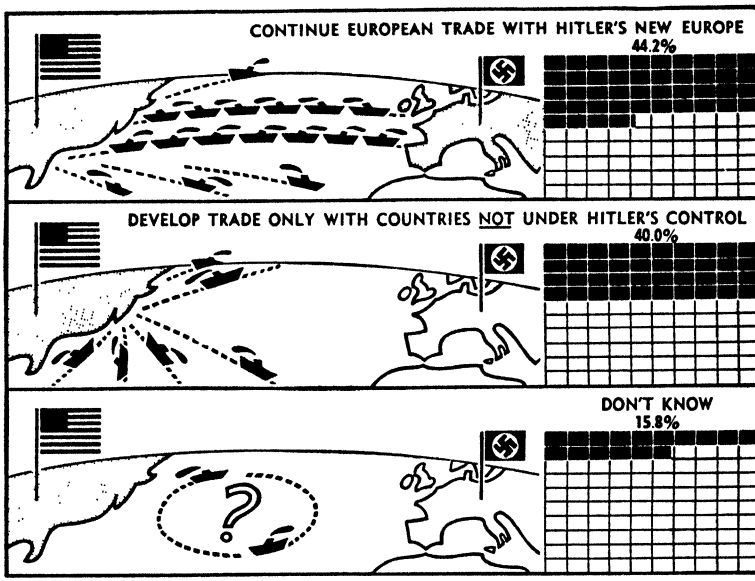
Fig. 3:19*

If Hitler wins, should we:

Find some way of continuing our European commercial business with Hitler's new Europe.....44.2%

Make every effort to develop business only with countries not under Hitler's control.....40.0

Don't know 15.8



EACH GRID = 100%; EACH BLOCK = 1%

* Reprinted from the August, 1940, *Fortune* survey by special permission of the editors.

5. Compute the ratio of college graduates and non-college graduates for the following groups:

Groups	College Graduates	Non-College Graduates
A	45	175
B	225	110
C	65	195
D	135	155
E	10	280

6. Gallup reported the following results in a public opinion poll on the question: "Which political party—the Republican or the Democratic—do you think is most interested in persons of ABOVE AVERAGE income? (New York *World Telegram*, February 11, 1946.)

Democratic	14%
Republican	57%
No Difference	17%
No Opinion	12%

Assume the total sample consisted of 2500 cases.

a. What is the percentage difference between those answering "Democratic" and "Republican"?
b. What is the percentage excess of those expressing "No Difference" over those with "No Opinion"?
c. How many more, in percentages, answered either "Democratic" or "Republican" as compared with those who did not answer either "Democratic" or "Republican"?

7. Compute the average percentage of the total group majoring in psychology for the following five college groups:

College	Total Number of Students	Per Cent Majoring in Psychology
A	3000	6%
B	500	10%
C	1200	15%
D	2000	11%
E	900	2%

8. What is the essential difference between a simple bar graph and a bar trend graph? For what purpose is each used?
9. What is the essential difference between a map chart and a pictorial chart? For what purpose is each used?
10. Choose an appropriate graphic device for portraying each group of data in Exercises 5, 6, and 7.

CHAPTER 4

The Correlation of Categorical Data

A. THE CROSS-TABULATION OF CATEGORICAL DATA

Whether there is any relationship between the data of two attributes or qualities is determined by methods of analysis known as *correlation*. As the origin of the word correlation implies, the procedure is a means of determining whether there is any *association* or "co-" relation, that is, *relation between* the differentiated data of two attributes.

Correlation methods are useful and relevant for problems in descriptive statistics, although in practice they are more often applied to problems in analytical and sampling statistics. As indicated in Chapter 1, these latter are problems in which the statistical data constitute *samples* of larger populations. When used for problems in descriptive statistics, a correlation coefficient simply summarizes the degree of co-relation found between two non-variable or variable attributes. When used for problems in sampling statistics, a correlation coefficient not only summarizes the relationship between the sample data of two attributes, but also provides the basis for an estimate of the correlation between the *populations* or *universes* from which the samples are derived.

In this chapter we shall first describe the basis for organizing categorical data in descriptive statistics so that we can determine by inspection whether the data of two attributes are correlated. We shall then develop the fundamental methods used for the actual computation of coefficients of correlation for the categorical data of non-variable attributes, as well as for the data of variables that are grouped into broad classes.

Cross-Tabulation Essential to Correlation

The essence of what is implied by correlation can be exemplified by the study of a 2 by 2 correlation chart, often referred to as a fourfold tabulation because of the four cells of cross-tabulated data. Table 4:1, which shows such a chart, is simply a cross-tabulated distribution of listener attitudes for two non-adjacent sequences, or parts, of a radio program.* The attitudes of each listener were studied in the Program Analyzer Laboratory of the Columbia Broadcasting System. Listeners whose attitudes were favorable to the two

* J. G. Peatman and Tore Hallonquist, *The Patterning of Listener Attitudes Toward Radio Broadcasts — Methods and Results*, Stanford Univ. Press, Stanford University, 1945.

sequences are indicated by a plus sign; those whose attitudes were unfavorable are indicated by a minus sign. In other words, despite the fact that attitudes constitute an attribute or trait which is a variable, the data in this study were dichotomized from a seven-point attitude scale to provide two classes: *favorable* and *unfavorable* attitudes. The dichotomization of the listeners' attitudes for each program sequence was made near the middle of each distribution. The actual number of cases in each class of the dichotomized attributes is given by the marginal totals at the bottom and at the right of the fourfold table. Thus, of the 59 subjects, 31 had favorable attitudes and 28 had unfavorable attitudes toward the earlier sequence. Similarly, for the later sequence, 30 had favorable and 29 had unfavorable attitudes.

Inspection of only the marginal totals does not provide any insight into a possible correlation between the attitudes of the listeners. The marginal totals simply summarize the results for each attribute separately. In order to study the possibility of correlation between the attributes, the cross-tabulations of the data into each of the four cells must be analyzed. The statistical frequencies of these cross-relationships are given in each of the four cells in Table 4:1. In the case of correlated data, a statistical frequency in any cell

Table 4:1. Cross-Tabulation of Listeners' Attitudes Toward Two Non-Adjacent Sequences of a Radio Program

Attitudes Toward Earlier Sequence	Attitudes Toward Later Sequence: −	Attitudes Toward Later Sequence: +	n_r
+	a: 2 (+, −)	b: 29 (+, +)	31
−	c: 27 (−, −)	d: 1 (−, +)	28
n_c	29	30	$N = 59$

represents a *pair of observations* or measurements, one for each attribute correlated, which are related by virtue of a common property. In this case, a paired observation represents the attitudes of *the same* individual for the two sequences of the broadcast.

The Correlation Chart as a Geometric Field

The cross-tabulations in Table 4:1 were thus made for attributes which are actually variables but were dichotomized into two broad classes. It is because of this fact that there is no question as to the order of the arrangement of each of the two classes. The more favorable attitudes in each case are symbolized by a plus sign, and the less favorable attitudes by a minus sign.

The fourfold table is analogous to a geometric field with the *origin* in the center at the right-angle intersection of the two lines dichotomizing each attribute. These are the usual four quadrants of such a field and they have been designated as *a*, *b*, *c*, and *d*. The order inherent in the attributes has been laid off for the vertical side of the square (the ordinate) and the horizontal side of the square (the abscissa) so that the quadrants themselves have the signs usually used in such a field; that is, the *b* and *c* cells are positive quadrants ($b = +, +$ and $c = -, -$), and the *a* and *d* cells are negative quadrants ($a = +, -$ and $d = -, +$). Having utilized the order inherent in the two attributes and laid off each one accordingly, we have conformed to the implications of a geometrical field whereby a relationship obtained for the data will be truly positive or negative. Thus, if the major proportion of the cases is distributed in the *b* and *c* quadrants, and a considerably smaller proportion therefore lies in the *a* and *d* quadrants, the correlation is positive; and, contrariwise, if a considerable majority of the cases is distributed in the negative quadrants, *a* and *d*, the correlation is negative.

Let us now examine the implications of the upper left-hand cell in Table 4:1, in which 2 of the 59 cases are entered. As indicated by the symbols at the left and top, these two cases represent listeners who had favorable attitudes toward the earlier sequence and unfavorable attitudes toward the later sequence. They constitute but a small proportion of the listeners in their respective column and row. That is, of the 31 individuals who had favorable attitudes toward the earlier sequence, only 2 had unfavorable attitudes toward the later sequence. And of the 29 individuals who had unfavorable attitudes toward the later sequence, only 2 had favorable attitudes toward the earlier sequence.

The 29 cases in the upper right-hand cell represent listeners whose attitudes toward both sequences were similar—all were favorable. The 27 cases in the lower left-hand cell also represent individuals with similar attitudes for both sequences—in this case, all unfavorable.

Thus it is apparent that inspection of a fourfold table like Table 4:1 usually reveals whether two attributes or traits are correlated. If a considerable majority of the cases are in either set of the diagonally related cells (*a* and *d*, or *b* and *c*) there is evidence of correlation. From the data in Table 4:1 we see that practically all the listeners had attitudes toward the later sequence which corresponded with their attitudes toward the earlier sequence (negatives with negatives and positives with positives). The correlation for these data is therefore not only high but positive. It is *positive* inasmuch as the individuals with favorable attitudes toward one sequence are, on the whole, the individuals with favorable attitudes toward the other sequence; and similarly, the individuals with unfavorable attitudes toward one sequence are on the whole the individuals with the unfavorable attitudes toward the other sequence. The correlation coefficient itself is high (for its computation, see page 94)

because all but three of the 59 cases are in the positively associated cells, *b* and *c*.

An example of a high degree of *negative* correlation is illustrated by the cross-tabulated data in Table 4:2. This represents a hypothetical redistribution of the data in Table 4:1. The correlation coefficient for these data would be identical in value with that for the data in Table 4:1, but it would be expressed as a *negative* value. Such a result is considered *negative* because a large majority of the listeners switched their attitudes toward the later sequence as compared with their attitudes toward the earlier sequence. The 27 individuals in the upper left-hand cell of Table 4:2 represent listeners

Table 4:2. Hypothetical Redistribution of the Cross-Tabulated Data in Table 4:1 to Illustrate Negative Correlation

Attitudes Toward Earlier Sequence	Attitudes Toward Later Sequence: −	Attitudes Toward Later Sequence: +	n_r
+	*a* 27	*b* 1	28
−	*c* 2	*d* 29	31
n_c	29	30	$N = 59$

whose attitudes toward the earlier sequence were favorable, but were unfavorable for the later sequence; similarly, the 29 cases in the lower right-hand cell represent listeners whose attitudes toward the earlier sequence were unfavorable, but were favorable for the later sequence. In other words, the majority of the cross-tabulated frequencies are now in the diagonal cells *a* and *d* which signify negative rather than positive associations.

An Example of No Correlation

No correlation between the cross-tabulated data in Table 4:1 would be indicated by a similar proportion of the cases being distributed in each of the four quadrants. Table 4:3 shows the data that yielded the marginal totals in Table 4:1, but redistributed so as to illustrate *no* correlation. The 31 listeners whose attitudes toward the earlier sequence were favorable are now about evenly divided in their attitudes toward the later sequence. Similarly, half of the 28 individuals whose attitudes toward the earlier sequence were unfavorable now have favorable attitudes toward the later sequence. There is thus no correlation between their attitudes, as distributed in the four cells

of Table 4:3. This is another way of saying that listeners who had favorable attitudes toward the first sequence are just as likely to have favorable as unfavorable attitudes toward the later sequence; similarly, listeners who had unfavorable attitudes toward the earlier sequence are just as likely as not to have favorable attitudes toward the later sequence.

Table 4:3. Hypothetical Redistribution of the Cross-Tabulated Data in Table 4:1 to Illustrate No Correlation

Attitudes Toward Earlier Sequence	Attitudes Toward Later Sequence: −	+	n_r
+	a 15	b 16	31
−	c 14	d 14	28
n_c	29	30	N = 59

The Correlation of Non-Variable Attributes

As already indicated, Tables 4:1, 4:2, and 4:3, illustrating the correlation of dichotomized data, were based upon the data of attributes which in reality are variables. By virtue of this fact, the dichotomized data were laid out on the horizontal and vertical sides of the 2 by 2 matrix so as to give a result that would correspond to the meaning of the usual quadrants of coordinate axes in a geometric field. However, for non-variable attributes, there is no order inherent among the categories of data themselves. In other words, for truly non-variable attributes the decision as to how the categories of each attribute shall be laid off on the vertical and horizontal sides of the cross-tabulation matrix is purely arbitrary. Similarly, the concepts of positive and negative correlation have no meaning for the cross-tabulated results of such data. It is rather a question of whether there is *any* correlation or association. If there is any correlation, it must be interpreted by means of verbalization.

This situation, characteristic of the correlation of non-variable attributes, is illustrated by Table 4:4. The data are of the kind often obtained in market research investigations. One hundred persons—50 men and 50 women—are asked: "Have you ever used K brand of soap?" The marginal totals at the right of the table indicate that 50 of the total group answer "Yes" and 50 answer "No." Thus, these data give an even division not only for the sexes but also for the responses. Inspection of the cross-tabulated data in the four cells indicates that there is considerable correlation between the sex

of the respondents and the character of their replies. Thus, 40 of the 50 women answer "Yes," whereas only 10 of the 50 men answer in the affirmative.

Table 4:4. Cross-Tabulation of the Data of Two Non-Variable Attributes (Commodity Use by Men and Women)

Respondents' Replies	Sex of Respondent: Women	Men	n_r
Yes	*a* 40	*b* 10	50
No	*c* 10	*d* 40	50
n_c	50	50	$N = 100$

If the results of this cross-tabulation could be interpreted as were the data in Tables 4:1, 4:2, and 4:3, the association would be described as negative correlation. The distribution of the cross-tabulated data in Table 4:4 is most like that in Table 4:2; 80% of the cases are in the "negative" quadrants (*a* and *d*). However, to describe the correlation in Table 4:4 as negative is not warranted. Whether the data on the sex of the respondents is arranged as in Table 4:4 or as in Table 4:5 is purely arbitrary. Table 4:5 presents exactly the same correlation as Table 4:4, but the replies of the male respondents are cross-tabulated in the left-hand cells (*a* and *c*) and those of the female respondents in the right-hand cells (*b* and *d*).

Table 4:5. Cross-Tabulation of the Data (from Table 4:4) of Two Non-Variable Attributes (Commodity Use by Men and Women)

Respondents' Replies	Sex of Respondent: Men	Women	n_r
Yes	*a* 10	*b* 40	50
No	*c* 40	*d* 10	50
n_c	50	50	$N = 100$

The interpretation of a correlation between two non-variable attributes thus involves verbalizing the relationship as it is observed to exist, rather than labeling the result as negative or positive. In this case, the women generally have used the brand of soap, and the men generally have not.

The foregoing remarks also apply to a correlation between any two attributes, one of which is a non-variable and the other a variable. Thus, a correlation between sex and height or between sex and ability cannot be described as either positive or negative. Men tend to be taller than women, and one sex might tend to do better than the other in a particular ability test.

The Correlation of Polytomous Attributes—Market Research Data

The data of non-variable attributes of research are often classified in more than the two categories characteristic of a dichotomy. We saw earlier that attributes divided into three categories yield a trichotomy, and, if into more than three categories, a polytomy. Methods for computing a correlation coefficient for the cross-tabulated data of polytomous non-variable attributes have not been fully developed. The available methods are based upon the assumption that trichotomous or polytomous divisions are derived from a variable (as they often are) rather than from a non-variable attribute. Pearson's method for the Coefficient of Mean Square Contingency, developed later in this chapter, is used to measure the degree of correlation for such situations. (See also pages 443 ff.)

Table 4:6. Attitudes of Economic Groups Toward Private vs. Government Management: The 3 by 4 Cross-Tabulation of Two Attributes

	Economic Groups				
	Low	Lower Middle	Upper Middle	High	n_r
Private Management	230	660	570	225	1685
Government Management	120	140	68	13	341
Don't Know	150	200	112	12	474
n_c	500	1000	750	250	$N = 2500$

Table 4:6 presents a 3 by 4 cross-tabulation based on data obtained from a market research investigation reported during the war by the Psychological

Corporation.* One of the questions asked of the nation-wide sample of respondents was:

> Do you think that business companies will do a better job if they are allowed to keep on under their own management, or if the Government takes them over and runs them completely?

Of the 2500 respondents, 67% thought that business companies would do a better job if kept under their own management. Only 14% thought that the companies would do better if under government management. Nineteen per cent did not know. These are the over-all results. If, however, the data are analyzed in relation to income groups, the results shown in Table 4:6 are obtained. These data reveal that there is a correlation between ECONOMIC STATUS and the nature of the answer to the question. A relatively greater proportion of the higher income groups felt that the companies would do a better job under their own management. Conversely, a relatively greater proportion of the lower income groups felt that a better job would be done under government management. Moreover, a greater proportion of the higher income groups had a *definite opinion*. In fact, only 12 of the 250 respondents in the highest income group gave a *DK* answer, whereas 150 of the 500 respondents in the lowest income group gave this answer.

The cross-tabulated results in Table 4:6 can be *interpreted* more readily if the frequencies of each cell are converted into percentages. Such a conversion immediately raises a question as to what total (N) or set of totals shall be used as the base. The answer depends upon the type of comparison to be made. Since the original differentiation of the respondents into income groups orients the comparison in this direction, the totals of each of the four income groups provide the most appropriate bases for the percentages of each cell.

A comparison of the respondents' answers to the question could be made by using the total 2500 cases as the base for the percentages in each cell of the 3 by 4 table. However, the most relevant picture of the relation between the respondents' opinions and their economic status is that shown in Table 4:7. According to this table, 90% of the high income, 76% of the upper middle income, and 66% of the lower middle income group felt that business companies would do a better job under their own management; but only 46% of the low income group were of this opinion. The proportion of the affirmative answers decreases as the economic status of the respondents decreases. On the other hand, the proportion of replies in favor of government management increases as the economic status decreases—from 5% for the high income group to 24% for the low income group; however, the rate of increase here is not the same as the rate of decrease in the former case. The difference is accounted for by the *DK*'s. As the economic status of the respondents decreases, the proportion

* The Psychological Corporation, "The Eighth Nation-Wide Social and Experimental Survey," New York, 1943.

of *DK*'s increases: only 5% of those in the high income group gave *DK* answers, whereas 30% in the low income group gave *DK* answers.

Table 4:7. The Cross-Tabulated Frequencies in Table 4:6 Converted into Percentages

	Economic Groups				
	Low	Lower Middle	Upper Middle	High	All
Private Management	46%	66%	76%	90%	67%
Government Management	24%	14%	9%	5%	14%
Don't Know	30%	20%	15%	5%	19%
	100%	100%	100%	100%	100%
(N)	(500)	(1000)	(750)	(250)	(2500)

From the point of view of a correlational analysis, the data in Tables 4:6 and 4:7 might perhaps be clearer if the respondents' answers were reclassified as in Tables 4:8 and 4:9. As indicated in Chapter 2, a complication arises in the analysis of market research data when there are a considerable number of *DK* answers. In the present case, the answers of the respondents can first be dichotomized into two categories: (1) those who have an opinion, and

Table 4:8. Reclassification of the Data in Table 4:7, Dichotomizing Respondents According to Those Who Had an Opinion and Those Who Did Not Have an Opinion

	Economic Groups				
	Low	Lower Middle	Upper Middle	High	Total Group
Those with an Opinion	70%	80%	85%	95%	81%
Those with No Opinion (*DK*'s)	30%	20%	15%	5%	19%
	100%	100%	100%	100%	100%
(N)	(500)	(1000)	(750)	(250)	(2500)

(2) those who do not have an opinion. This reclassification of the data yields the results shown in Table 4:8. It is now clear that having an opinion or having no opinion on this question is correlated with economic status, for those in the higher income groups were more likely to have an opinion, and those in the lower income groups were less likely to have an opinion.

The cross-tabulation in Table 4:9 presents the trend among the 81 per cent of the respondents who had an opinion one way or the other. It is now even

Table 4:9. Reclassification of the Data in Table 4:6, Dichotomizing the Replies of Respondents Who Had an Opinion

	Economic Groups				
	Low	Lower Middle	Upper Middle	High	Total Subgroup
Private Management	66%	83%	89%	95%	83%
Government Management	34%	17%	11%	5%	17%
	100%	100%	100%	100%	100%
(N)	(350)	(800)	(638)	(238)	(2026)

clearer than it was from Tables 4:6 and 4:7 that there is a correlation between economic status and answer to the question. The higher the income, the greater the proportion of answers in favor of private management; and conversely, the lower the income status, the greater the proportion of answers in favor of government ownership. However, even in the low income group, practically two-thirds (66%) of the members of this group who had an opinion felt that business companies would do a better job if allowed to keep on under their own management.

The data used for the preceding tables are based on attributes at least one of which is a variable, viz., income or economic status. In market research investigations, however, economic status is usually treated by a breakdown into three, four, or five classes, rather than by attempted measurements or ratings on a continuous scale. That this attribute is a variable rather than a non-variable should be apparent from the order inherent in the arrangement of the data; that is, the classes are arranged in order of economic groupings, from lowest standards of living or income to highest standards of living or income. Sometimes this attribute is quantitatively differentiated by using actual income in dollars as the index of economic status. However, research has indicated pretty clearly that a few economic groupings based upon several factors such as type of home, home conveniences, neighborhood, etc.,

as well as dollar income, give a better index of socio-economic status than dollar income alone.

The other attribute, respondents' replies to the question, can be interpreted as a variable less readily, if at all. This is because, as was brought out in the reclassification of the data in Tables 4:8 and 4:9, the replies actually yield two different attributes. The first attribute, as shown in Table 4:8, is "Having an Opinion" or "Not Having an Opinion." This is a dichotomy, and there is little question but that this attribute is non-variable—a person either has an opinion or he hasn't. The second attribute, as indicated in Table 4:9, represents a twofold division of the replies of the respondents who had an opinion. This situation is not so simple with respect to the logic of characterizing the results as a true dichotomy of a non-variable attribute. It might be argued, for example, that the form of the question itself has forced the dichotomy, that in reality the respondents may have had many different shades of opinion with respect to private management and government management, or a combination thereof. In practice, such an attribute is often treated *as if* it comprises a variable attribute, with the shades of opinion theoretically distributed according to a standard type of distribution. If the assumption can be made that the shades of opinion, as well as the differences in economic status, are distributed in a form similar to the normal probability curve, then the statistical method for the Contingency Coefficient can be used without further qualification to compute a correlation coefficient for the cross-tabulated data in Table 4:9. (See p. 94.)

B. METHODS FOR THE CORRELATION OF CATEGORICAL DATA

The extent to which the categorical data of two attributes or qualities are correlated can be expressed by a coefficient. Correlation is not an all-or-none affair; it is not a question of whether two attributes are perfectly correlated or not correlated at all. Correlation is always a question of the *degree* of such relationship as may be present.

Mathematical methods that have been developed to express the degree of correlation between two attributes, whether variable or non-variable, yield an index which may vary in value from no correlation (indicated by zero) to perfect correlation (indicated by a coefficient of 1.00). In the case of variable data for which positive and negative directions of correlation are relevant, the correlation coefficients may vary from a perfect positive correlation (1.00) through zero to a perfect negative correlation (−1.00). A coefficient of .90 expresses a high degree of positive association, whereas a coefficient of .10 expresses a very low degree of positive association. Similarly, a coefficient of −.90 expresses a high degree of negative association, and a coefficient of −.10 expresses a very low degree of negative association. We shall see that not all methods of correlation yield coefficients which are strictly comparable

in all respects. Coefficients of the values 1.00, 0.00, and −1.00 are always comparable for any method of correlation. They signify, respectively, perfect positive correlation, no correlation, and perfect negative correlation. However, coefficients with values between these limiting points vary in their implications according to the method of correlation used.

In the historical development of methods for computing correlation coefficients, most attention has been given to those that index the degree of correlation between *variable* attributes. This is because most research studies in psychology and related fields have been concerned with data of variables rather than non-variables. The most widely used method of indexing the degree of correlation between two variables is the one referred to in Chapter 1, the product-moment method originally developed by Galton and perfected by Karl Pearson. (See Chapter 9.) Many years ago, however, Yule presented methods that yield an index for the correlation between non-variable as well as variable attributes which are dichotomized or divided into only a few categories.

Yule's Coefficient of Association (*A*) for Dichotomized Non-Variable Attributes *

The method developed by Yule for indexing the degree of correlation between two dichotomized non-variable attributes yields a coefficient known as the Coefficient of Association. It is simple to compute, and we shall illustrate it with the data from Table 4:4, which are reproduced in Table 4:10.

Table 4:10. Correlation Between Sex and Use of Brand K of Soap

Respondents' Replies	Sex of Respondents: F	Sex of Respondents: M	n_r
Yes	*a* 40	*b* 10	50
No	*c* 10	*d* 40	50
n_c	50	50	N = 100

These data give in a fourfold table the cross-tabulation of the sex and answers of respondents to the question: "Have you ever used K brand of soap?"

* G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, Griffin, London, 12th ed., 1940, p. 44. Cf. also G. U. Yule, "On the Methods of Measuring the Association Between Two Attributes," *Journal of the Royal Statistical Society*, 75:576–642, 1912.

Cell *a* of Table 4:10 indicates that 40 of the 50 female respondents had used Brand K; cell *b* indicates that only 10 of the 50 males had used this brand; cell *c* indicates that 10 of the females had not used this brand; and cell *d* indicates that 40 of the males had not used this brand. We have already seen that this cross-tabulation implies a considerable correlation between sex and soap usage. Generally the women have used the soap and generally the men have not. The actual degree of correlation for these data may be expressed by the Coefficient of Association, which is .88. It is computed by the following formula:

$$A = \frac{ad - bc}{ad + bc} \qquad [4:1]$$

Yule's Coefficient of Association

in which *a*, *b*, *c*, *d* represent the statistical frequencies of these respective cells in the 2 by 2 table. *A* is therefore computed as follows:

$$A = \frac{(40)(40) - (10)(10)}{(40)(40) + (10)(10)} = \frac{1600 - 100}{1600 + 100} = \frac{1500}{1700} = .88$$

The Coefficient of Association is based upon the ratio of the difference of the products of the frequencies in the diagonal cells to the sum of the products of the frequencies in the diagonal cells. For dichotomized non-variable attributes, a negative value does not really signify negative correlation. Whether or not the coefficient *A*, as computed, is negative or positive is arbitrary, since the result depends upon the arrangement of the respective categories of each attribute in the fourfold table. Only in the case of cross-tabulated data of two variable attributes does a negative sign with a correlation coefficient signify correlation which is, in reality, negative rather than positive. It will be recalled that positive correlation in such cases signifies a tendency for the larger or greater values of each variable to be associated, and for the lower values of each variable to be associated, whereas negative correlation signifies that the larger values of one variable tend to be associated with the lower values of the other. The implications of a coefficient of correlation for non-variable attributes must be verbalized from the nature of the data correlated.

The Correlation of Dichotomized Variables: The Phi Coefficient

The Coefficient of Association can also be used to index the correlation between two variable attributes which have been dichotomized. However, a better index of their correlation is given by the phi (ϕ) coefficient. As in the case of the Coefficient of Association, ϕ is based upon the ratio of frequencies in the cells of a fourfold table. The ϕ coefficient is computed as follows:

$$\phi = \frac{bc - ad}{\sqrt{(a + b)(c + d)(a + c)(b + d)}} \qquad [4:2]$$

ϕ coefficient of correlation for dichotomized non-variable attributes

If the two attributes are both variables which have been dichotomized, a better estimate of their correlation is made by dividing ϕ by the constant .637, as follows:

$$\phi_r = \frac{\phi}{.637} \qquad [4:3]$$

ϕ coefficient of correlation for dichotomized variates

This correction factor yields a coefficient greater than 1.00 (the mathematical limit of perfect correlation) if ϕ is greater than .637. In such cases, the result is interpreted as approaching 1.00 as a limit.

If only one attribute is a dichotomized variable, the other attribute being a true dichotomy of a non-variable, a better estimate of their correlation is made by dividing ϕ by the constant .798, as follows:

$$\phi_r = \frac{\phi}{.798} \qquad [4:4]$$

ϕ coefficient of correlation for correlation of a true dichotomy with a dichotomized variate

That the ϕ coefficient for dichotomized non-variable attributes yields an index of correlation which differs from the Coefficient of Association may be shown with the data in Table 4:10. For these data ϕ is as follows:

$$\phi = \frac{(10)(10) - (40)(40)}{\sqrt{(40 + 10)(10 + 40)(40 + 10)(10 + 40)}}$$

$$= \frac{-1500}{\sqrt{6250000}} = \frac{-1500}{2500} = .60$$

This value of .60 (the negative sign is dropped as irrelevant) for the ϕ index of correlation between the two non-variable attributes in Table 4:10 is a measure of correlation which is more analogous in its implications to the measures of correlation for variates (cf. Pearson's product-moment coefficient, Chapter 9) than is the value of .88 obtained by the Coefficient of Association method.

Whenever it is logical to assume that one or both of the attributes being correlated is a variable rather than a non-variable, ϕ should be used instead of A. The ϕ correlation of the dichotomized data of the variable attributes in Table 4:1 is computed as indicated in Table 4:11. The ϕ coefficient for these data is found to be .90. However, both the attributes which are dichotomized in this table are in reality variables. Therefore it is relevant to estimate r (the product-moment correlation coefficient) for this ϕ value. By Formula 4:3, $\phi_r = \phi/.637$. Since ϕ is greater than .637 and the correction, if applied, would yield an estimated r value in excess of 1.00, we can conclude that ϕ_r approaches 1.00 as a limit and denote the estimate as equal to 1.00^-.

Table 4:11. ϕ Correlation of Listener Attitudes for Two Sequences of a Radio Program

Attitudes Toward Earlier Sequence	Attitudes Toward Later Sequence: −	Attitudes Toward Later Sequence: +	n_r
+	a: 2	b: 29	31
−	c: 27	d: 1	28
	n_c: 29	30	$N = 59$

$$\phi = \frac{(29)(27) - (2)(1)}{\sqrt{(2+29)(27+1)(2+27)(29+1)}} = \frac{783 - 2}{\sqrt{(31)(28)(29)(30)}}$$

$$= \frac{781}{\sqrt{755160}} = \frac{781}{869.0} = .899, \text{ or } .90$$

The Correlation of Polytomous Attributes: The Contingency Coefficient

Methods have also been developed for computing an index of correlation for the cross-tabulated data of attributes divided into more than two broad classes. The method most commonly used for this purpose is the one developed by Karl Pearson. It gives a statistical index of correlation called the Coefficient of Mean Square Contingency. In practice, this coefficient is symbolized by C and is referred to as the Contingency Coefficient.

The computation of C will be described for the data in Table 4:12, which were derived from the cross-tabulation in Table 4:6 and represent the relation between economic status and respondents' opinions concerning private vs. government management of business companies. It will be recalled that the attribute income status is definitely a variable differentiated not on a continuous, quantitative scale, but into four broad classes ranging from lowest to highest. It will also be recalled that the other attribute (having an opinion on the question) may also be construed as a variable representing a dichotomization of all shades of opinion ranging from strong convictions in favor of private management to very weak or no convictions in favor of private management (with government management presumably as the alternative). In any event, the data in Table 4:6 yield a 2 by 4 or eightfold cross-tabulation of two attributes whose data are arranged categorically. The percentage values of each cell in Table 4:9 have been reconverted to frequencies in

Table 4:12. Relation Between Economic Status and Respondents' Opinions About Private vs. Government Management of Business (from Data in Table 4:6)

	Economic Groups: Low	Lower Middle	Upper Middle	High	n_r
Private Management	a 230	b 660	c 570	d 225	1685
Government Management	e 120	f 140	g 68	h 13	341
n_c	350	800	638	238	N = 2026

Table 4:12, inasmuch as the contingency method of correlation is based upon frequencies rather than proportions. The Contingency Coefficient itself is equal to the following ratio:

$$C = \sqrt{\frac{S - N}{S}} \qquad [4:5]$$

Pearson's Coefficient of Mean Square Contingency

where S is equal to the sum of the ratios obtained in the last column of Table 4:13 and N is the total number of correlational frequencies or paired observations used in the cross-tabulation of the two attributes.

Steps in Computing C (*Table 4:13*)

Column 2: List the statistical frequencies obtained for each cell of the cross-tabulation shown in Column 1. These are the obtained frequencies, f_o.

Column 3: Square each of the cell frequencies to get f_o^2.

Column 4: Compute "independence values" for each cell. These values give the hypothetical frequency (f_h) for each cell to be expected on the basis of chance according to the total number of frequencies of the column (n_c) and row (n_r) in which the cell is located. The hypothetical "independence value" of any cell is equal to the following ratio:

$$f_h = \frac{n_r n_c}{N} \qquad [4:6]$$

Hypothetical frequency value for any cell of a correlation chart, on the assumption of *independence* between attributes

where n_r is equal to the number of frequencies for the row in which the cell is located, n_c is equal to the number of frequencies for the column in which

the cell is located, and N is equal to the total number of frequencies obtained and used in the cross-tabulation. The sum of these "independence values" (N_h) should equal (except for dropped decimals) the total number of correlation frequencies (N_o).

Column 5: Compute for each cell the ratio of its squared frequency value (f_o^2 of Column 3) to its theoretical "independence value" (f_h). These ratios are presented in Column 5. Sum these ratios to obtain the value of S for the computation of C.

The contingency coefficient for the data in Table 4:13 is computed at the bottom of the table and is found to be .23.

Table 4:13. Computation of C, the Contingency Coefficient, from Cross-Tabulation of Table 4:12

(1) Cells	(2) f_o (Obtained Frequencies)	(3) f_o^2	(4) $(n_c n_r)$	(5) $(n_c n_r)/N_o = f_h$	(6) $\frac{(f_o)^2}{f_h}$
a	230	52,900	350(1685)	589,750/2026 = 291.1	181.7
b	660	435,600	800(1685)	1,348,000/2026 = 665.4	654.6
c	570	324,900	638(1685)	1,075,030/2026 = 530.6	612.3
d	225	50,625	238(1685)	401,030/2026 = 197.9	255.8
e	120	14,400	350(341)	119,350/2026 = 58.9	244.5
f	140	19,600	800(341)	272,800/2026 = 134.6	145.6
g	68	4,624	638(341)	217,558/2026 = 107.4	43.1
h	13	169	238(341)	81,158/2026 = 40.1	4.2
	$N_o = 2026$			(check) $N_h = 2026.0$	$S = 2141.8$

$$C = \sqrt{\frac{2141.8 - 2026}{2141.8}} = \sqrt{\frac{115.8}{2141.8}} = \sqrt{.0541} = .2326 = .23$$

How this correlation is interpreted depends upon the way in which the data are distributed in the cells in Table 4:12. The coefficient C is written without a sign. The arrangement of the attributes of the table and of the data in each cell, however, indicates the direction of the correlation, viz., the higher the economic status, the greater the *tendency* for the respondents' opinions to favor private management of business, and conversely, the lower the economic status the greater the tendency to favor government management. That this is the case is emphasized by comparing the actual, obtained frequencies per cell (f_o) with the theoretical number of frequencies for each, as shown in Table 4:14. Cells *a* and *h*, for example, have fewer obtained than hypothetical frequencies, whereas cells *d* and *e* have more obtained than hypothetical frequencies. These are the extremes of the paired observations and such a trend of the frequencies in the diagonally located cells is, as we have seen, indicative of correlation. However, the distribution of the obtained

Table 4:14. Comparison of Obtained Frequencies (f_o) with Hypothetical Frequencies (f_h) for Data in Table 4:13

	Economic Groups			
	Low	Low Middle	Upper Middle	High
Private Management	a $f_o = 230$ $f_h = 291$	b $f_o = 660$ $f_h = 665$	c $f_o = 570$ $f_h = 531$	d $f_o = 225$ $f_h = 198$
Government Management	e $f_o = 120$ $f_h = 59$	f $f_o = 140$ $f_h = 135$	g $f_o = 68$ $f_h = 107$	h $f_o = 13$ $f_h = 40$

frequencies in relation to the hypothetical frequencies for all eight cells needs to be considered. This is done in computing C, which is equal to only .23. This degree of correlation is not very marked; at best it is indicative only of a *tendency* to a relationship between the attributes correlated.

Mathematical Limits of C *and Estimates of* C_{cor}

There is a limit to the computed value of a Contingency Coefficient in the correlation of broad or only a few classes. The value of C is not affected by variations in the total number of frequencies (provided they are considerable), but it is affected by the number of cells used in the cross-tabulation of two attributes. Yule and Kendall * have presented the maximum possible values of C for the cross-tabulations of attributes, each of which is divided into the same number of categories. These values, which are given in Table 4:15, are useful in correcting C values to obtain better estimates of the degree of correlation.

Table 4:15. The Maximum Values of C for Correlated Attributes Divided into the Same Number of Categories

2 by	2-fold,	C	cannot exceed	.707
3 by	3-fold,	"	"	.816
4 by	4-fold,	"	"	.866
5 by	5-fold,	"	"	.894
6 by	6-fold,	"	"	.913
7 by	7-fold,	"	"	.926
8 by	8-fold,	"	"	.935
9 by	9-fold,	"	"	.943
10 by	10-fold,	"	"	.949

The maximum possible computed value for C derived from a fourfold (2 by 2) cross-tabulation is .707. A better estimate of correlation by means

* Yule and Kendall, *op. cit.*, p. 69.

of the Contingency Coefficient can be obtained by dividing C by .707, when C is derived from a fourfold table. Similarly, an estimate from a 3 by 3 table can be obtained by dividing C by .816. Generally, C is most satisfactory for attributes polytomized into 5 by 5 or more divisions, provided, however, the subdivisions are not too fine (not greater than 9 or 10).

EXERCISES

1. Set up a fourfold table for the following data, determine the degree of correlation by means of (a) the Coefficient of Association and (b) the phi coefficient, and interpret your results:

   Of a total group of 400 adults, 150 are men; and 175 of the total group belong to labor unions. Of the 175 who belong to labor unions, 100 are men.

2. Set up a correlation table for the following data, determine the degree of correlation in terms of the Contingency Coefficient, and interpret your result:

   Of a total group of 3000 people, 500 graduated from high school and had at least some college education; 1500 graduated from high school but had no college education; and 800 had some high school education but did not graduate. Of the 500 with some college education, 100 were in favor of "large families" (as opposed to "small families"); of the 1500 high-school graduates with no college education, 450 favored "large families"; and of the 800 with some high-school education, 320 favored "large families." The total number favoring "small families" was 1980. (Note that part of the total group had no high-school education.)

CHAPTER 5

The Reduction and Organization of Variate Data

A. INTRODUCTION

This chapter will present the initial methods necessary for the reduction and organization of the data of variable attributes or qualities. They include procedures for rearranging the raw data of a variable into an ordered structure that will compactly portray the character of the distribution of data. Whether the form of a variable is similar to the normal, bell-shaped distribution, or some other type, can usually be ascertained from a graph of the *frequency distribution* of the variable. Additional methods of statistics to be used in the description of a result will in part depend upon the type of distribution that a variable yields. We shall be concerned here with the following:

1. The range and array.
2. The frequency distribution.
3. The histogram and the frequency polygon.
4. The percentage frequency distribution and polygon.
5. The cumulative and percentage cumulative frequency distributions.

B. THE RANGE AND ARRAY

The first step in the organization of variate data consists in determining the *range*. The numerical values of the lowest and highest (or smallest and largest) scores in a group of variate data constitute the range. It is readily determined *by inspection* for a group of only 100 or so cases. Looking through the intelligence test scores in Table 5:1, we see that the highest is 100 and the

Table 5:1. Intelligence Test Scores of 100 College Freshmen

49	66	66	86	75	34	21	12	58	17	34	30
52	56	67	58	80	40	21	17	73	56	13	40
79	73	65	61	43	30	85	21	40	66	14	75
91	65	50	38	85	94	26	56	76	24	71	73
100	53	30	11	40	64	38	56	10	11	3	59
62	52	61	76	11	39	99	52	19	73	24	77
58	44	36	26	38	15	64	63	19	45	42	64
31	48	62	89	60	8	76	21	89	47	98	29
47	63	91	32								

lowest is 3. The range is therefore 3 to 100. We employ the same procedure for the data in Table 5:2, putting the largest negative values at the lowest end of the scale. The largest rating is 126 and the smallest rating (largest negative value) is −165; the range is therefore −165 to 126.*

Table 5:2. Bernreuter "Sociability" Scores of 100 College Freshmen

53	−103	−65	−108	−31	−154	−49	−31	−37	−22
42	−56	126	−69	1	−33	−93	−25	−5	−77
−3	−49	−87	−91	−50	116	5	60	−30	−83
−66	−113	−42	−76	−132	63	−95	−70	104	−17
−137	24	−79	−29	0	−53	−45	1	30	−5
−78	−69	−21	−37	−106	−19	−17	−14	−58	−52
−94	−13	−27	−8	43	−67	−51	−120	−22	−60
−87	−124	−51	−97	39	−104	−86	−93	−30	−40
−165	−30	−131	63	6	−39	−65	−18	8	64
−45	−101	−86	−41	−49	−77	−57	−8	−29	−3

Table 5:3. Strong Interest Ratings for "Physician" of 100 College Freshmen

C+	C	C	C	C	B+	B	A	C+	C+	C+	B−	B
B−	B	B−	C+	C	C+	C	C	C	B−	C	C	B−
B	B−	C	C+	C	C	B	C+	B−	C	C	C+	C+
C	B	B	C	C	C	C	B−	C	B−	C	C+	C+
B	C	C+	B+	C	C+	C	C	B+	B	C	B−	C
A	B+	C	C	C+	C+	C	C	B−	B−	B	C	B+
C+	C	C	C+	C+	C	C+	B	C	C	C+	C+	C+
C	C	C	C	C+	B	C	C	B+				

The data in Table 5:3 consists of interest ratings that have been converted from numerical scores to broad classifications (hence, categorized) on a letter scale, with *order of interest pattern* represented by the order of the alphabet. *Most* interest is signified by the first letter of the alphabet, A, and *least* interest by C. The table is found on inspection to include both A and C ratings; the range is therefore C to A.

The range thus gives the outside limits of the numerical values or ratings present in a distribution of variate data. Not only is it a valuable aid in the initial steps of constructing the frequency distribution, but it also provides an index of the *spread*, or *dispersion*, of the scores, i.e., it provides an index of the extent to which the measures differ.

* What "most" and "least" mean on the Bernreuter scale, or any other scale of variable quantities, is ultimately a question of functional analysis; here we are concerned only with reducing such data to the form of a frequency distribution.

The Range as a Comparative Measure

The range as an index of spread or dispersion of the data of a variable is sometimes used to compare the variation of scores or ratings of one group of individuals with the variation of another group. If the data for each group are for the same variable—are obtained by means of the same test or rating device—this comparison procedure may be useful. However, it is obviously meaningless to compare the ranges of the three variables in Tables 5:1, 5:2, and 5:3. The fact that these ranges are

3 to 100 (Table 5:1)
−165 to 126 (Table 5:2)
C to A (Table 5:3)

is a summarizing item of information about each, but not a comparative one. But if a second group of college freshmen shows scores for each of these variables with the following ranges:

6 to 91
−100 to 53
C to B

the differences between this second group of freshmen and the first one suggest that the second group is not as variable as the first in the functions or traits differentiated.

It should be emphasized, however, that we cannot place too much confidence in differences in dispersion measured by the values of the ranges of two or more distributions of data, because the range tells nothing about the internal organization of the series of scores or ratings. For example, the bulk of scores in each distribution being compared may have about the same dispersion of measures over the scale, despite different ranges. Furthermore, range values are likely to be too erratic or scattered as limiting points of groups of data to serve usefully as comparative or even summary measures of dispersion for the groups as a whole. Even if the ranges of two groups of scores represent the maximum and minimum values that can be obtained on a test or rating scale (by virtue of the nature of the method employed in scoring the result), there still remains the possibility that the internal structure of each group of scores is markedly different.

The Array

A simple but usually not the most useful or easiest method of organizing the data of variables is to rearrange *all* the scores of each group in order of size or order of rating. This constitutes an array. The data in the preceding three tables have been rearranged into arrays in Tables 5:4, 5:5, and 5:6.

Two general characteristics of such arrays are perhaps evident from these three tables. (1) The array does not provide a very useful type of organization; with large groups of data the result hardly warrants the labor required. (2) The

Table 5:4. Array of Intelligence Test Scores
(Data in Table 5:1 Rearranged in Order of Size)

100	66	52	30
99	66	52	30
98	66	50	29
94	65	49	26
91	65	48	26
91	64	47	24
89	64	47	24
89	64	45	21
86	63	44	21
85	63	43	21
85	62	42	21
80	62	40	19
79	61	40	19
77	61	40	17
76	60	40	17
76	59	39	15
76	58	38	14
75	58	38	13
75	58	38	12
73	56	36	11
73	56	34	11
73	56	34	11
73	56	32	10
71	53	31	8
67	52	30	3

Table 5:5. Array of Bernreuter Scores
(Data in Table 5:2 Rearranged in Order of Size)

126	−13	−45	−79
116	−14	−45	−83
104	−17	−49	−86
64	−17	−49	−86
63	−18	−49	−87
63	−19	−50	−87
60	−21	−51	−91
53	−22	−51	−93
43	−22	−52	−93
42	−25	−53	−94
39	−27	−56	−95
30	−29	−57	−97
24	−29	−58	−101
8	−30	−60	−103
6	−30	−65	−104
5	−30	−65	−106
1	−31	−66	−108
1	−31	−67	−113
0	−33	−69	−120
−3	−37	−69	−124
−3	−37	−70	−131
−5	−39	−76	−132
−5	−40	−77	−137
−8	−41	−77	−154
−8	−42	−78	−165

Table 5:6. Array of Strong Interest Ratings
(Data in Table 5:3 Rearranged in Order of Rating)

A	B	B−	B−	C+	C+	C	C	C	C
A	B	B−	B−	C+	C+	C	C	C	C
B+	B	B−	C+	C+	C+	C	C	C	C
B+	B	B−	C+	C+	C+	C	C	C	C
B+	B	B−	C+	C+	C+	C	C	C	C
B+	B	B−	C+	C+	C+	C	C	C	C
B+	B	B−	C+	C+	C	C	C	C	C
B+	B	B−	C+	C+	C	C	C	C	C
B	B	B−	C+	C+	C	C	C	C	C
B	B	B−	C+	C+	C	C	C	C	C

character of the results, especially in Table 5:6, suggests a method of organizing variate data that not only is more satisfactory but is also more commonly used, viz., *the frequency distribution*. Since the ratings in Table 5:6 consist of only a few letters, the total number of cases in each category of A's, B+'s, B's, B−'s, C+'s, and C's can be quickly derived from the array to yield the frequency distribution shown in Table 5:7.

This arrangement obviously summarizes the data in Table 5:3 in a way greatly superior to the array shown in Table 5:6. However, as will be seen

Table 5:7. Frequency Distribution of Data in Table 5:6 (Physician—Strong Interest Inventory)

Interest Rating	Frequency
A	2
B+	6
B	12
B−	12
C+	24
C	44
	N = 100

in the next section, the procedures used in deriving this frequency distribution involve much more labor and time than is necessary. Instead of first ordering all the data into arrays, we can greatly simplify the procedure by *tallying* the original unordered group of data into appropriate classes (or class intervals).*

C. THE FREQUENCY DISTRIBUTION

The structure of a group of measures or ratings for a variable is readily revealed by the construction of a frequency distribution and a graph of the results. Such a procedure requires that the data be tallied into appropriate classes or class intervals.

The Class Interval

The first problem that ordinarily arises in the construction of a frequency distribution involves the selection of appropriate class intervals for the data. Sometimes the classes to be used are evident, and no further consideration is necessary. This was true of the data in Tables 5:3 and 5:6, whose frequency distribution was shown in Table 5:7.

More often, however, the class intervals are not so readily indicated by the original data. A case in point is the group of intelligence test data in Tables 5:1 and 5:4, the scores of which had a range of from 3 to 100. The Bernreuter data in Tables 5:2 and 5:5, with a range of from −165 to 126, are another example. For each group of data of these two variables, there were 100 cases. In order to have a frequency distribution that will give a picture of the whole which will be more meaningful and useful than the arrays in Tables 5:4 and 5:5, class intervals that include more than one score possibility must be set up. If this is not done, and the integral values of all possible scores within the range are taken as the class intervals, there will be too many null classes,

* The subdivisions of a quantitatively distributed variable are usually described as *classes*, or *class intervals*, whereas the subdivisions of a non-variable, or of a non-quantitative variable, are usually described as categories.

i.e., class intervals with no frequencies. Thus, for the Bernreuter data, there are 292 score possibilities, since the range is −165 to 126, and 165 + 126 + 1 (for zero) equals 292. But there were only 100 cases in the group and therefore each possible integral score value could not be represented by a frequency.

The usual procedure in selecting class intervals for a variable whose range is equal to or larger than the number of cases in a fairly good-sized group (N equal to or greater than 100) is to establish class intervals of a size that will yield from *12 to 20 classes* in the frequency distribution. There is, however, nothing magical in this particular choice. In practice, it is usually not necessary to have more than 20 class intervals unless the size of the groups is very large. On the other hand, if less than 12 class intervals are used, it may be necessary to make certain corrections in the results for computing some of the measures used in both descriptive and sampling statistics.*

Determining the Range or Size of a Class Interval

The easiest way to establish the range of class intervals so as to have from 12 to 20 intervals, each equal in size, is to divide the total number of different score possibilities (which is equal to the difference between the extreme values of the distribution as given by the range, plus one) by 12 or 15 or 20, or by any other number between 12 and 20, according to the number of intervals desired.

In the intelligence test scores in Table 5:1, the total number of score possibilities is 98, since the range of the scores is from 3 to and including 100. If a minimum of 12 class intervals is desired, 98 is divided by 12, and a rounded value of 8 is obtained. There will therefore be 12 or 13 class intervals with a range of 8 score units, each interval equal in size, for a frequency distribution of these data. In practice, however, class intervals for integral score values are more often taken for convenience as equal to 5 or 10 units. If class intervals equal to 10 score units are used for the intelligence test data in Table 5:1, there will be only 10 (or 11) classes. On the other hand, if class intervals equal to 5 score units are used, there will be 20 (or 21) classes. The choice will depend upon the general purpose underlying the statistical treatment of the original data in the investigation. If the research worker is mainly concerned in developing a frequency distribution to portray the structure of the group result as a whole, then the 10 or 11 class intervals of 10 units each will serve better than 20 or 21 intervals of 5 units each, since the total number of cases is only 100.

In the case of the Bernreuter data in Tables 5:2 and 5:5, the range was found to be −165 to 126, and the total number of different score possibilities was 292. Dividing 292 by 15 gives 19.5, and therefore intervals with a range of 20 score units will give approximately 15 classes. For convenience in notation

* Cf. Sheppard's correction for the standard deviation derived from broad classes, chap. 7.

and tabulation, class intervals with a range of more than 10 units are usually set up to the nearest multiple of 5, as for example 15, 20, 25, etc. This is the case for intervals up to 25 or 30 units. For intervals of greater size, ranges of 50, 75, or 100 units are usually employed.

The first thing to do, then, in making a frequency distribution is to lay out the range of score possibilities in successive class intervals of a convenient and appropriate size. If too many class intervals are used for groups of data with only 25 or 50 or even 100 cases, there will be too many null classes and the picture of the structure of the group as a whole will not be satisfactory. Too many class intervals used for the data in Tables 5:1 and 5:2 would be little improvement over the arrays shown in Tables 5:4 and 5:5. On the other hand, it is apparent that the interest data in Tables 5:3, 5:6, and 5:7 offer no initial problem of class arrangement, because the total number of different ratings is only *six* and hence there can be no more than six classes.

The Mathematical Limits of a Class Interval *

Whatever class intervals may be chosen, it is essential to know the precise mathematical limits of the successive class intervals used for a frequency distribution of a variable. This problem may not seem to be distinct from that of determining the size or range of a class interval. However, as soon as one begins to tabulate the original data into their respective class intervals, it is likely to become apparent. Furthermore, if any statistical computations are to be made from the tabulated data, it is essential to know the mathematical limits of each interval, because either the mathematical values of class-interval limits or the values of interval mid-points are necessary in computing most statistical measures of a variable.

The problem of the mathematical limits of class intervals will be illustrated by the tabulation of a group of age data. For convenience, let us assume that such data are to be tabulated in class intervals of one year each, and that the data range from 5 to 15 years. If the highest class interval is set up for those cases 15 years of age, the next highest class interval will include those 14 years of age, etc. If we assume further that such data form a continuum or continuous series of age possibilities, and if the first case to be tabulated is 14 years and 9 months old, the question arises as to which of these two classes shall be used for the tally. The answer depends upon what is taken as the class limits of each interval. Sometimes these limits are taken a half year below and a half year above the integral age value. In this case, the mathematical limits of the 15-year class interval will be 14.5 to, but not including, 15.5. Similarly, for the 14-year class interval, the limits will be 13.5 to 14.5, etc. In other words, the mathematical limits of such class intervals would be six months preceding and six months following each integral year

* For a more detailed discussion of this problem, see J. G. Peatman, "On the Meaning of a Test Score in Psychological Measurement," *American Journal of Orthopsychiatry*, 9:23–29, 1939.

value. With these limits, a person 14 years and 9 months of age would be tallied in the 15-year interval.

Age data, however, are often taken in such a way as to make erroneous the use of the class limits just described. Thus, if ages are reported as of each person's *last birthday*, such data will have to be tabulated with respect to class intervals whose lower limit is the integral years of birthday age. Those reporting an age of 15 years would thus be in a class interval whose limits begin at 15.0 years and range to, but do not include, 16.0 years. Similarly, an age of 14 years would be in the class interval ranging from 14.0 to 15.0 years.

If an investigator wishes to obtain age data which can be correctly tabulated in year intervals ranging from a half year below to a half year above the year age, he will ask individuals to give their age in years *as of their nearest birthday*, instead of their age *as of their last birthday*. Everyone reporting his age as 15 years would then be in the 14.5 to 15.5 range.

A Measure Occupies an Interval Whose Limits Extend Above and Below the Value of the Measure

This problem of establishing the mathematical limits of a class interval has been dealt with fairly systematically. However, in statistical literature two principles or methods are used. The difference in them can readily be illustrated by means of integral class intervals for measures which are themselves integral values. Consider, for example, the series of intelligence test scores in Table 5:1. If the size of each class interval is taken as 1, then each test score will have the range of a class interval. Some authors consider the lower limit of such an interval to be the value of the measure itself. Thus, the lower limit of the class interval for intelligence test scores of 100 will be 100. The upper limit will be 100.99^+ (to but not including 101.0). Other authors consider that the measure occupies an interval whose mid-value corresponds with the value of the measure itself. In this case, the lower limit for intelligence test scores of 100 would be 99.5 and the upper limit would be 100.499^+ (to but not including 100.5).

We shall employ the latter interpretation; that is, we shall consider the mathematical units of a class interval as equal to a half unit below and a half unit above the actual measures in the interval. Thus a reaction-time score measured to the one-hundredth of a second occupies a class interval that ranges a half hundredth below and a half hundredth above the score values: a score of .04 second occupies a unit interval with mathematical limits of .035 and $.0449^+$, or from .035 to $.045^-$ (i.e., as a limit).

This procedure not only is in agreement with general practice but has a more logical foundation than the first method. Any measure is subject to errors of observation. Such errors are as likely to affect a measure favorably (positively) as unfavorably (negatively). Therefore it is logical to interpret the measures obtained as lying, on the average, near the middle of unit class

intervals. The actual range of error may, of course, extend beyond the lower and upper limits of a unit interval. However, regardless of this possibility, when the mathematical limits of the interval are taken as below and above the value of an observed measure, the results will be more likely to agree with the facts than if the mathematical limits extend from unit value to unit value.

It is also to be observed that most of the measurements in psychology and related fields of investigation do not have any practical or useful meaning unless they are considered in relation to the other measures of the series of which they are a part. Since such measurements are *relative* values anyway, a difference of half an integral value makes no real difference. As T. L. Kelley pointed out more than twenty years ago in discussing this problem:

"Uniformity is needed, and it would be in harmony with well-nigh universal procedure in the physical and biological fields to consider a score of 10 as being also a class index, or midpoint of an interval. Should this lower the grade of a few million school children by one half a point, no harm would be done and the great advantage of having the recorded test scores exactly those to be used in calculating means, standard deviations, correlations, etc., and of having the recorded measures also the class indexes in graphs is attained." *

Finally, another advantage of this procedure for interpreting the mathematical limits of a class interval is the fact that a psychological test score, or any measure, is a value that is to be interpreted as *occupying an interval* rather than as coinciding with the value of a point on a scale of measures. An intelligence test score of 100 is to be interpreted as an index that occupies a class interval, say from 99.5 to 100.5^-, rather than as an index equal to the point value of 100. At the same time, for statistical purposes this index can be treated algebraically as 100 since 100 represents the mid-value of the interval.

There is only one real exception to the application of this *middle-of-the-interval interpretation of a measure.* The exception is simply stated. If a measure is originally derived in such a way that its value definitely signifies the lower limit of a unit interval, then the limits of the interval should be established to fit that fact. As already indicated, age measures as sometimes taken prove an exception to the general rule; that is, when individuals are asked to give their age in years *as of their last birthday*, the data obtained will form a series in which integral year values should be used as the lower limits of successive class intervals.

Mathematical vs. Written Interval Limits

In practice the actual mathematical limits of a class interval are not always expressly stated. Whether or not they are explicitly written, they are implied in all statistical computations. Thus, if a group of test scores is tabu-

* T. L. Kelley, *Statistical Method*, Macmillan, New York, 1923, pp. 12–13.

lated into class intervals, each with a range of five test score units, the lowest class interval is ordinarily written as 10 to 14, the next lowest as 15 to 19, etc. However, in line with the interpretation we have indicated, the mathematical limits of these intervals are 9.5 to but not including 14.5, 14.5 to but not including 19.5, etc. The use of score values rather than mathematical values for denoting class-interval limits has a twofold advantage. (1) The written notation itself is simpler; (2) the integral score limits do not suggest, as do the mathematical limits, that fractionate values actually occur in the test score data. This latter point is not unimportant. Whenever the most refined measurements are integral values, it is well to set up class intervals for frequency distributions that will not suggest a precision of measurement finer than the obtained integral values.

The Mid-Point Value of a Class Interval

The problem of determining the mid-value of any class interval is simplified by the adoption of a standard interpretation for the mathematical limits of class intervals. The mid-point of a class interval is obviously the value exactly midway between *the mathematical limits* of the interval. It can readily be obtained as the difference between the mathematical values of the upper and lower limits; this difference is divided in half and the result added to the lower limit value. Or half of this difference may be subtracted from the mathematical value of the upper limit of the interval. An even simpler method is to *add* the mathematical values of the lower and upper limits and divide this sum by 2. In other words, the value of the mid-point of an interval is equal to the average of the sum of the mathematical limits of the interval. Thus, for an interval with mathematical limits 9.5 and 14.5, the mid-point value is

$$\frac{9.5 + 14.5}{2} = \frac{24.0}{2} = 12.0$$

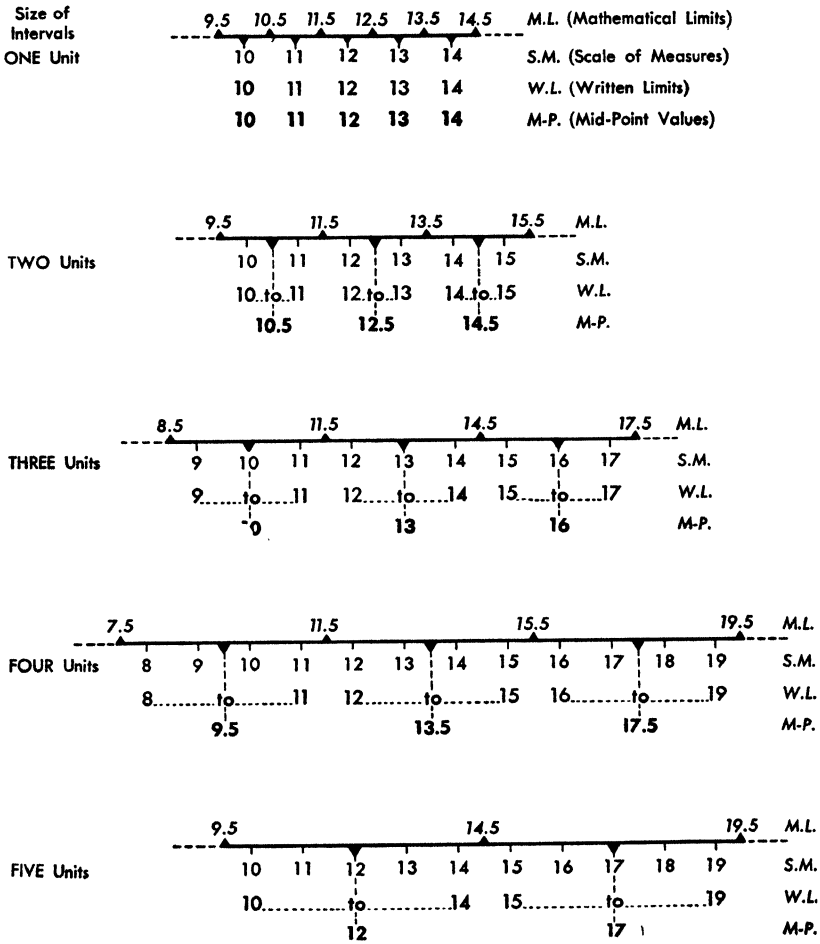
The application of the procedures described in the preceding paragraphs is illustrated in Table 5:8 and Fig. 5:1 for some commonly used class intervals.

Table 5:8. Mid-Point Values of Class Intervals of a Size Commonly Used for Research Results

Written Class-Interval Limits *	Mathematical Limits	Unit Size of Interval	Mid-Point Value
10	9.5 and 10.5^-	1 unit	10.0
10–11	9.5 " 11.5^-	2 units	10.5
9–11	8.5 " 11.5^-	3 units	10.0
8–11	7.5 " 11.5^-	4 units	9.5
10–14	9.5 " 14.5^-	5 units	12.0
10–19	9.5 " 19.5^-	10 units	14.5
25–49	24.5 " 49.5^-	25 units	37.0

* Note that the lower limits of these class intervals are written as values equal to multiples of the unit size of the interval, and that class intervals of one unit do not have written limits since the interval value is the mid-point of the interval.

Fig. 5:1. Illustration of Mathematical Limits, Written Limits, and Mid-Point Values of Varying Size Class Intervals for Scales of Measures



The Choice of the Scale Limits of Class Intervals

Once the *size* (or unit range) of the class intervals to be used for a frequency distribution has been decided upon, the next step consists in selecting the starting point of the highest (or lowest) class interval for the series of scores. If the range of a collection of scores is from 3 to 100, as in Table 5:1, and if the investigator has decided to use class intervals with a range of ten units, should he use a score of 3 to begin the lowest interval, or will some other number be more convenient? The answer is somewhat arbitrary, but the

ordinary procedure is usually systematized. If the original data form a series that has regularly spaced gaps which yield few or no frequencies, the class-interval limits should be chosen so that the frequencies which do occur lie near the mid-values of each successive interval. Such a series sometimes occurs in *percentage* school grades because teachers recognize (consciously or unconsciously) that reliable differentiations of pupils' achievement to within *one* per cent are hardly possible, and therefore they record grades at those values which are multiples of 5. In such a case, the class-interval limits should be chosen so that their mid-point values will be 95%, 90%, 85%, 80%, etc.

However, if the data of a series of measures are fairly well distributed throughout the scale, the lower limit value of each class interval is commonly set by taking its *written* value as a *multiple of the size of the interval.* Class intervals with a range of ten units would, under this procedure, have a written lower limit equal to 10 or a multiple of 10 (20, 30, etc.), or zero, if needed. Sometimes this systematization of the choice of written limits for class intervals is applied to the upper rather than the lower limit; that is, the written value of the *upper limit* of each interval is taken as a multiple of the size of the interval.

Class Intervals for the Intelligence Test Scores in Table 5:1

If either of the preceding principles is applied to the data in Table 5:1, the range of which was from 3 to 100, and if class intervals of ten units are used, the written value of the lower limit of the lowest class interval will be *zero*, or the written value of the upper limit of this class interval will be *10*. This interval will therefore be taken as 0 to 9, or as 1 to 10. Similarly, the written value of the lower limit of the highest class interval will be 100 (the highest score in the distribution), or the written value of the upper limit will also be 100. Thus, the highest class interval of ten units will be taken as 100 to 109 or as 91 to 100.

For these data it is better to take the upper (rather than the lower) limit of each class interval as a multiple of the size of the interval. If this is not done, and if the lower limit values are taken as multiples of ten, then the highest class interval will range from 100 to 109 and its mid-value of 104.5 will be considerably higher than the highest possible score (100) obtained in the distribution in Table 5:1. On the other hand, if the upper limit is used for the whole distribution, the class intervals for all score possibilities of the data in this table will be as follows:

91 to 100
81 to 90
71 to 80
61 to 70
51 to 60
41 to 50
31 to 40
21 to 30
11 to 20
1 to 10

Class Intervals for the Bernreuter Data in Table 5:2

Let us now consider the Bernreuter data in Table 5:2. As was previously indicated, the range of these data is −165 to 126, and the size of the interval to be used will be 20 units if approximately 15 class intervals are desired. If the written values of the lower limit of each interval are taken as a multiple of 20, the class intervals for all score possibilities will be as follows:

120 to	139
100 to	119
80 to	99
60 to	79
40 to	59
20 to	39
0 to	19
−20 to	−1
−40 to	−21
−60 to	−41
−80 to	−61
−100 to	−81
−120 to	−101
−140 to	−121
−160 to	−141
−180 to	−161

The Tally

With the class intervals chosen for the data in Tables 5:1 and 5:2, we are now ready to obtain the actual distribution of frequencies for each by means of a tally of each set of data. The tally procedure is identical with that described for categorical data in Chapter 2.

The simplest method is to check each score as it appears in the table, going down the columns or across the rows, and to tally each in its appropriate class interval. If the original data are on individual cards rather than in a table, each case, in turn, is tallied for the variable under consideration.*

The first score in Table 5:1 is 49. It will therefore be tallied in the class interval, 41 to 50. The next score, going down the column, is 52, and it will be in the interval, 51 to 60. Continuing with each score, we have the result shown in Table 5:9. The tally for the data in Table 5:2 is presented in Table 5:10.

* Note that when a correlation coefficient is to be computed for the relation between *two* variables, a cross-tabulation yielding a correlation tally is made, instead of separate frequency distributions for each variable. The two frequency distributions are then obtained from the correlation tally. (See chap. 9.)

Table 5:9. Tally of Intelligence Test Scores
(Data from Table 5:1)

Class Intervals	Tally of Frequencies
91 to 100	𝍸 |
81 to 90	𝍸
71 to 80	𝍸 𝍸 ⊏
61 to 70	𝍸 𝍸 𝍸
51 to 60	𝍸 𝍸 ⊏
41 to 50	𝍸 □
31 to 40	𝍸 𝍸 ⊏
21 to 30	𝍸 𝍸 ⌊
11 to 20	𝍸 𝍸 |
1 to 10	⊏

Table 5:10. Tally of Bernreuter Scores
(Data from Table 5:2)

Class Intervals	Tally of Frequencies
120 to 139	|
100 to 119	⌊
80 to 99	
60 to 79	□
40 to 59	⊏
20 to 39	⊏
0 to 19	𝍸 |
−20 to −1	𝍸 𝍸 ⌊
−40 to −21	𝍸 𝍸 𝍸 ⌊
−60 to −41	𝍸 𝍸 𝍸 |
−80 to −61	𝍸 𝍸 ⌊
−100 to −81	𝍸 𝍸 |
−120 to −101	𝍸 ⌊
−140 to −121	□
−160 to −141	|
−180 to −161	|

The *box method* of tallying, described briefly in Chapter 2, has been employed in Tables 5:9 and 5:10. This is in contrast to the older procedure for tallying in which each four successive tallies are denoted by vertical lines, the fifth tally being denoted by a slanting line through the four lines, thus:

𝍸 𝍸 ||

Either method is satisfactory, but the box method is preferable because there is likely to be less error when the frequencies for each class interval are counted.

The Frequency Distribution

The final frequency distribution is readily obtained from the tally by enumerating the number of tallies for each class interval. The distributions for the data in Tables 5:1 and 5:2 are presented in Tables 5:11 and 5:12.

Table 5:11. Frequency Distribution of the Intelligence Test Scores in Table 5:9

Class Intervals	Frequencies (*f*)
91 to 100	6
81 to 90	5
71 to 80	13
61 to 70	15
51 to 60	13
41 to 50	9
31 to 40	13
21 to 30	12
11 to 20	11
1 to 10	3
	$N = 100$

Table 5:12. Frequency Distribution of the Bernreuter Scores in Table 5:10

Class Intervals	Frequencies (*f*)
120 to 139	1
100 to 119	2
80 to 99	0
60 to 79	4
40 to 59	3
20 to 39	3
0 to 19	6
−20 to −1	12
−40 to −21	17
−60 to −41	16
−80 to −61	12
−100 to −81	11
−120 to −101	7
−140 to −121	4
−160 to −141	1
−180 to −161	1
	$N = 100$

In published reports the actual tally of original data is rarely included; rather, the distribution of a variable is usually presented as in these two tables. However, if the page is turned counterclockwise 90°degrees, the tally distributions in Tables 5:9 and 5:10 give a rough but graphic picture of the structure of each distribution, a picture not so readily conveyed by the frequency distributions in Tables 5:11 and 5:12.

D. THE HISTOGRAM AND THE FREQUENCY POLYGON

In order to give a more concrete picture of the structure of a group of variable data than is afforded by the frequency distribution alone, the results are portrayed graphically. Two types of graphs are used for frequency dis-

tributions; one is the frequency curve (line graph) and the other is the histogram. Each presents about the same picture, except that the curve tends to emphasize the continuity and general sweep of a distribution, whereas the histogram tends to emphasize distinctions from class interval to class interval. Which type of graph should be employed is for the most part a matter of personal preference unless one distribution is to be compared with another distribution which has already been graphed. In this case, the same type of graph should be employed for the second distribution.

Figs. 5:2 and 5:3 are histograms of the frequency distributions in Tables 5:11 and 5:12. Fig. 5:4 shows both a histogram and a frequency curve of the frequency distribution in Table 5:7. Figs. 5:5 and 5:6 are frequency curves of the histograms in Figs. 5:2 and 5:3.

The Histogram

The construction of a graph of a frequency distribution is greatly facilitated by the use of standard cross-section paper, either millimeter paper or paper ruled off in 20 units to the inch.

The distribution itself is plotted in the frame of reference of a geometric field, with two coordinate axes drawn at right angles to each other. The abscissa, or horizontal axis, is usually denoted as the *x-axis*. The ordinate, or vertical axis, is denoted as the *y-axis*. The two axes intersect at the *origin*. In a frequency distribution, the value of the *y*-axis at the origin is always equal to zero. This is because *frequencies* are always scaled on the vertical or *y*-axis, from zero to a value equal to the maximum frequencies obtained for the class intervals of the distribution. The *x*-axis, on the other hand, represents the scale of scores, or rating values. Since there is no true zero point in psychological scales, the presence of a zero for the *x*-axis depends on whether the particular series of measures being graphed includes a value of zero. The intelligence test scores in Fig. 5:2 are scaled from zero to 100, although actually there were no zeros in the distribution.

The Bernreuter scores include not only a zero but also negative numbers. This case is typical of the implications of a zero value on psychological scales. Zero is a number that serves to identify the *position* of an individual's performance or rating in a series. It does *not* signify a mathematical zero of *nothing*, i.e., in the case of the Bernreuter scale, it does not mean "no sociability." Similarly, the negative numbers in this distribution have no true algebraic significance. They serve rather to indicate *position* on a scale on which negative as well as positive numbers happen to be used. According to Bernreuter's *Manual for the Personality Inventory* from which these scores were derived, "Persons scoring high on this scale tend to be non-social, solitary, or independent. Those scoring low tend to be sociable and gregarious." Since any variable scaled on the *x*-axis should at the least signify for

the attribute or trait an *order* which ranges from *least to most*, the scale in Fig. 5:3 presumably may be interpreted as ranging from "least non-sociability" to "most non-sociability." Whether, in fact, such an interpretation is justified depends upon an empirical validation of the instrument (cf. Chapter 17, Section C).

Fig. 5:2. Histogram of Intelligence Test Scores—Frequency Distribution of Table 5:11

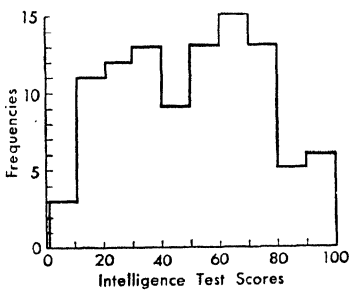
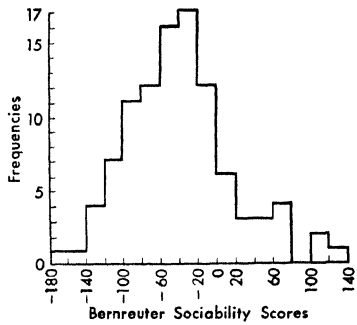


Fig. 5:3. Histogram of Bernreuter Scores—Frequency Distribution of Table 5:12

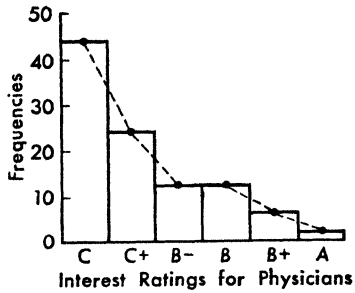


Scaling the y-axis and x-axis

The first problem in constructing a histogram is to mark off on graph paper the lengths of the x-axis and y-axis. There are no basic logical requirements to guide one here; what one does is rather a matter of general practice and convenience. Since the basic purpose of a histogram is to give a picture of the distribution, aesthetic considerations enter into the choice of procedure. Somewhat balanced proportions for the two scales are desirable. The scale on the x-axis is generally made somewhat longer than that on the y-axis, so as to give the effect of a figure which rests solidly on its base. If the ordinate scale is considerably longer than the abscissa scale, the effect is likely to be an unbalanced, top-heavy superstructure.

In practice, another consideration enters into the scaling of the two axes. For distributions of variables which tend to be of the normal bell-shaped type (cf. Fig. 1:1), the length of the x- and y-axes is chosen so as to be in a proportion of about 3 to $2\frac{1}{2}$. However, the distribution in Fig. 5:2 does not resemble a bell-shaped curve. It tends to be more rectangular, no doubt because of a relatively constant level of difficulty among items in the intelligence test. The scores of the Bernreuter Inventory in Fig. 5:3 are scaled on axes in the proportion of $3\frac{1}{4}$ to $2\frac{3}{4}$. This curve is more similar to the bell-shaped curve than is Fig. 5:2. The distribution of Bernreuter scores is definitely uni-modal near the center, and the frequencies decrease above and below the center roughly in the bilaterally symmetrical fashion characteristic of the

Fig. 5:4. Histogram of Strong Interest Ratings with Line Graph—Frequency Distribution of Table 5:7



normal bell-type distribution. That this bilateral decrease is not too marked is indicated by a slight piling up of frequencies between −100 and −30 and by a proportionate drop in frequencies for the upper half of the scale.

The distribution of Strong Interest ratings (for Physician) in Fig. 5:4 bears no resemblance to the "normal" distribution, but tends rather to be of the L-type. The two axes have nevertheless been scaled in a proportion roughly 3 to 2.

Drawing the Histogram

Having decided upon the lengths to use in scaling the x- and y-axes, we proceed to draw in the scales on the graph paper and to *plot* the actual frequencies for each class interval. In the histogram a semi-rectangular figure is used to indicate the relation between a given class interval and its frequencies. The *width* of the rectangle is equal to the width of the given class interval on the score scale. Technically the width should be taken as equal to the *mathematical limits* of the interval. In practice, however, the *written* score limits are used. Either procedure will of course give the same picture of the distribution. The only difference will be that the whole histogram will be shifted to the left by half a scale unit when the mathematical limits are used. In any event, when scores are integral values, as is the case in 5:2 and 5:3, the written scale for the x-axis is in terms of the integral values of either the lower or upper limits (not both) of successive intervals, rather than in terms of the mathematical limits.

Close inspection of Fig. 5:2 reveals that it has been drawn to the mathematical limits of each class interval. Thus the frequencies of the first interval at the lower (left) end of the scale are plotted for limits of 0.5 to 10.5, the integral limits having been taken as 1 to 10.

The *height* of the horizontal line drawn for each class interval is of course determined by the number of frequencies in the interval. If an interval has no frequencies, as in one case in Fig. 5:3, the graph of frequencies drops to the abscissa and a gap appears in the histogram.

Two additional points about the histogram should be noted. The first has to do with its actual construction. The frequencies of each class interval are often represented by a *closed* rectangle, as in Fig. 5:4. Whether the rectangles are closed, or open as in Figs. 5:2 and 5:3, is to a considerable extent a matter of personal preference. However, it is generally preferable not to close them in order that the *continuity* and *general structure* of the surface of the distribution of the variable will be readily perceived. From a logical point of view,

the closed rectangles suggest a non-continuous distribution such as is characteristic of the bar diagrams of categorical data (cf. Figs. 3:4, 3:5, and 3:6). Furthermore, in the case of variables the statistical treatment is developed on the assumption of a continuous, rather than a discrete or discontinuous, series of values. For this reason it is well to emphasize this continuity by not closing the rectangles.

The second point has to do with an assumption that is made when a histogram is used to portray the distribution of frequencies of a variable. The area under the surface of a histogram is an area of *frequencies* for the given scale of scores. This procedure assumes that the scores *are distributed evenly* throughout a given class interval, hence the horizontal line at the top of each semi-rectangle. Although this assumption may not be entirely supported by the ungrouped data for some distributions, it is adopted in the interest of statistical convenience. At the same time this assumption does not have to be closely fulfilled in practice in order for the results of the statistical treatment to have sufficient validity for ordinary psychological interpretation.

The Frequency Polygon, or Line Graph

The frequency distributions of variables are often shown more graphically by the frequency curve, or line graph, than by the histogram. A line graph has been superimposed on the histogram of the interest ratings in Fig. 5:1. Figs. 5:5 and 5:6 are frequency curves for the intelligence test and Bernreuter scores shown by histograms in Figs. 5:2 and 5:3.

Fig. 5:5. Frequency Curve (or Line Graph) of Intelligence Test Scores of Table 5:11

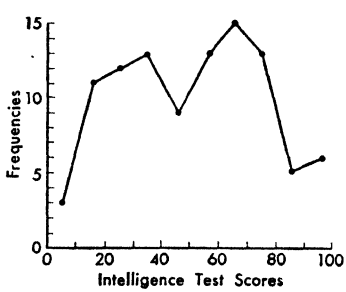
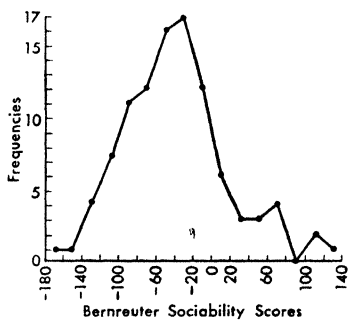


Fig. 5:6. Frequency Curve of Bernreuter Scores (Data of Table 5:12)



The graph of a frequency curve is prepared like a histogram except that the frequencies for each class interval are represented by points plotted with respect to the middle of each interval. The curve is then drawn by connecting the plotted points with *straight lines*.

We have seen that the mid-point values of class intervals can be obtained

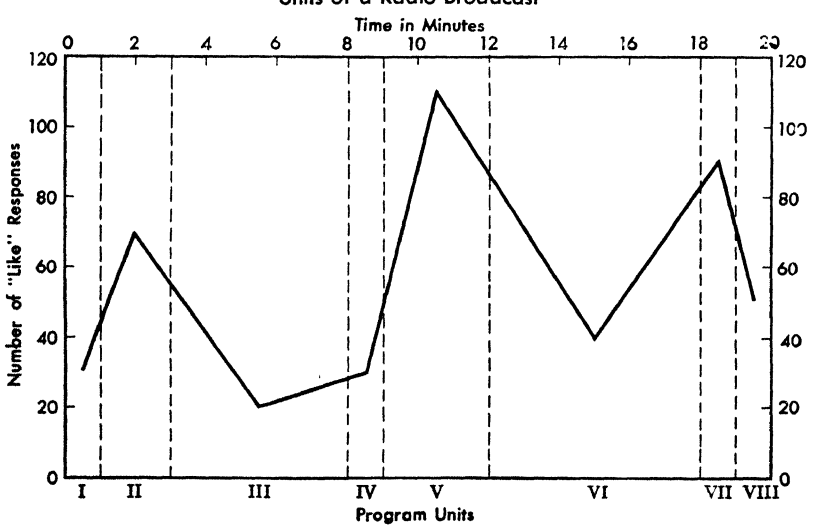
by finding the difference between the mathematical limits and adding one-half of this difference to the lower limit. It is often desirable to denote the score scale on the x-axis in terms of the values of the mid-points of each successive class interval rather than in terms of the values of successive lower limits. It of course makes no difference which values are denoted on the scale so long as the frequencies are plotted correctly in relation to the mid-point values of each interval. If the mid-points of class intervals are fractionate values, the successive lower integral limits of the class intervals are usually employed.

Comparative Usefulness of the Histogram and Frequency Polygon

An important advantage of the frequency curve over the histogram is the fact that the frequency distributions of several groups of data for a variable can be compared on the same graph more readily by means of a line graph or frequency curve. When such comparisons are to be made, the curves for each distribution should be clearly differentiated on the graph by means of different types of lines for each—a dotted line, a short bar line, a long bar line, as well as the solid line used in Figs. 5:5 and 5:6. When, however, the total number of frequencies of the several group results being compared differs markedly, line graphs of the *percentage frequency distributions* described in the next section are likely to be more satisfactory than line graphs of the original frequency distributions.

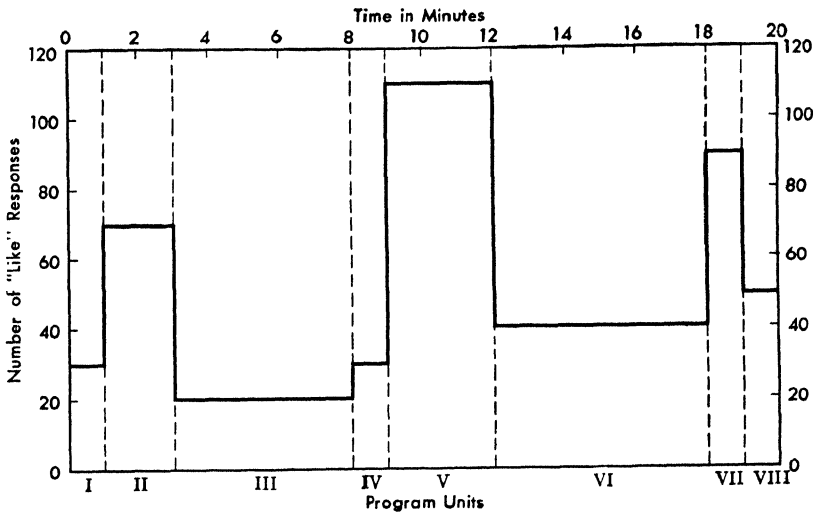
The histogram has an advantage over the line graph for series of data that

Fig. 5:7. Line Graph of Frequencies of "Like" Responses for the Successive Program Units of a Radio Broadcast



are grouped into class intervals unequal in size. Although such class intervals are rarely used for the frequency distributions of variables, they are often employed for the frequency data of a *time series* sequence. The superiority of the histogram over the line graph for such data is illustrated by Figs. 5:7 and 5:8, each of which depicts the trend of the same set of data.* The frequency scale (ordinate) represents the number of listeners expressing favorable

Fig. 5:8. Histogram of Frequencies of "Like" Responses for the Successive Program Units of a Radio Broadcast



attitudes toward the successive sequences on program units of a radio program. The program units of the broadcast are indicated on the horizontal scale (abscissa), each scaled according to the amount of broadcast time required for it. In studies of audience reaction the parts of a radio program should be divided on a functional or meaningful basis, rather than in terms of arbitrary and equally spaced time-interval units.

The line graph in Fig. 5:7 is inadequate because the level of the frequency of response for any particular part of the program is not clearly portrayed. Furthermore, a rising or declining trend of response *within* a program unit is suggested. Thus, the line graph is likely to imply that there was a higher frequency of favorable responses for the beginning and ending of the sixth program unit than for the middle of it. Actually, as indicated by the histogram of these same data, the information plotted is for the average number

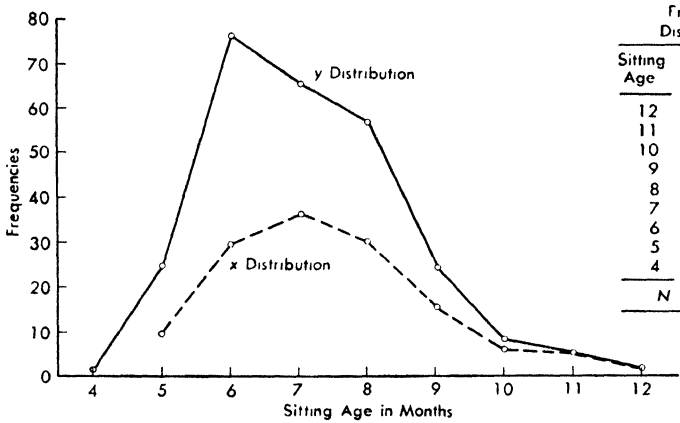
* These data were obtained from a Program Analyzer test of a radio broadcast. Cf. J. G. Peatman and Tore Hallonquist, *The Patterning of Listener Attitudes Toward Radio Broadcasts: Methods and Results*, Stanford Univ. Press, Stanford, 1945.

of reactions for each sequence *as a whole*. The rectangular character of the histogram indicates this fact and avoids the misleading suggestion given by the line graph. In addition, the histogram indicates more clearly the absolute as well as the relative length of each program unit.

E. THE PERCENTAGE FREQUENCY DISTRIBUTION AND POLYGON

The value of converting the frequencies of two or more distributions of a variable into *percentage frequencies*, so that the structure of their respective distributions may be compared more fairly, is well illustrated by Figs. 5:9 and 5:10, in which are compared the same two sets of data on the *sitting ages*

Fig. 5:9. Comparison of Two Distributions of Infants' Ages (in Months) of Beginning to Sit Alone



Sitting Age	Frequency Distributions f_x	f_y
12	1	1
11	5	5
10	6	8
9	15	24
8	30	57
7	36	65
6	29	76
5	9	24
4		1
	N = 131	261

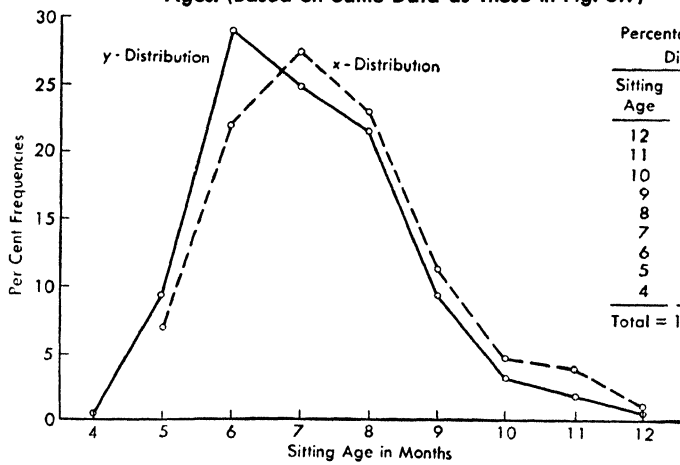
of two groups of infants.* The absolute frequencies in Fig. 5:9 yield two frequency curves whose form is apparently rather different—the one, peaked; the other, flat. However, as clearly revealed by the *percentage frequency distributions* in Fig. 5:10, the structure of the two sets of results is very similar. That this is the case is seen when the different N's of each group are taken into account by converting the frequencies of each class interval to their percentage of the total frequencies.

The construction of *percentage frequency curves* is identical with that of ordinary frequency curves except for the scaling of percentage values, instead of absolute frequencies, on the ordinate. The percentage values for each class interval of a distribution are most readily obtained by first computing to two or three decimal places the percentage value of one frequency. The product of the number of frequencies per interval and the percentage value

* Data from J. G. Peatman and R. A. Higgons, "Relation of Infants' Weight and Body Build to Locomotor Development," *American Journal of Orthopsychiatry*, **12**:234–240, 1942.

of a single frequency will give the desired percentage frequency value for the interval. Inasmuch as the percentage value of a single frequency is always the product of 100 and the reciprocal of N (the total sample), the necessary

Fig. 5:10. Comparison by Percentage Frequency Distributions of Infants' Sitting Ages. (Based on Same Data as Those in Fig. 5:9)



Percentage Frequency Distributions

Sitting Age	$\%_{f_x}$	$\%_{f_y}$
12	.8	.4
11	3.8	1 9
10	4.6	3 1
9	11.4	9.2
8	22.9	21.8
7	27.5	24 9
6	22.1	29.1
5	6 9	9.2
4		4
Total =	100 0	100 0

figures can be obtained from printed tables such as Barlow's.* Thus, N for the first sample, as indicated in Fig. 5:9, was 131; the reciprocal of N is

$$\frac{1}{N} = \frac{1}{131} = .00763$$

and 100(.00763) = 0.763, the percentage value of a single frequency. Similarly, for the second distribution, where $N = 261$:

$$\text{Percentage value of a single frequency} = \left(\frac{1}{N}\right) 100$$

$$100(\tfrac{1}{261}) = 100(.00383) = 0.383\%$$

For purposes of graphing, it is sufficient to carry out the percentage frequencies of each class interval to only one decimal place.

F. THE CUMULATIVE AND PERCENTAGE CUMULATIVE FREQUENCY DISTRIBUTION

The Cumulative Frequency Distribution

The comparison of two or more frequency distributions is also facilitated by the use of *cumulative frequency distributions*, especially *percentage cumulative frequency distributions.*

* *Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots and Reciprocals of All Integer Numbers up to 10,000*, Spon, Ltd., London. See also Table I, Appendix C, for reciprocals of all integer numbers up to 1000.

A cumulative frequency distribution is one for which the frequencies of adjoining class intervals, beginning at either end of the distribution, are added successively. In other words, the frequencies are cumulative from either end of the scale for successive class intervals. Ordinarily, as illustrated in the next to last column of Table 5:13, the frequencies are cumulated from the lower score values to the higher score values. The sum of the cumulative frequencies for the final class interval should, of course, always be equal to the number of cases (N) in the distribution.

Table 5:13. Cumulative and Percentage Cumulative Frequency Distributions of Infants' Sitting Ages

(Data from Fig. 5:9)

Sitting Age in Months (Age)	(Interval)	f	c.f.	Percentage c.f.
12	(11.5 to 12.5⁻)	1	131	100.0%
11	(10.5 to 11.5⁻)	5	130	99.2
10	(9.5 to 10.5⁻)	6	125	95.4
9	(8.5 to 9.5⁻)	15	119	90.8
8	(7.5 to 8.5⁻)	30	104	79.4
7	(6.5 to 7.5⁻)	36	74	56.5
6	(5.5 to 6.5⁻)	29	38	29.0
5	(4.5 to 5.5⁻)	9	9	6.9
		$N = 131$		

Percentage value of 1 frequency:

$$\frac{1}{N}(100) = \frac{1}{131}(100) = 0.763\%$$

In Table 5:13 the frequencies of the sitting-age scores of variable x in Fig. 5:9 are cumulated, beginning with the lowest class interval of 5 months. The 9 frequencies of this class interval are added to the 29 frequencies of the 6-month interval, giving a total of 38 cumulated frequencies for these two intervals. This signifies that 38 of the 131 infants had sitting-age scores of less than $6\frac{1}{2}$ months, since the upper mathematical limit of the 6-month class interval is 6.5. To these 38 cumulated frequencies are added the 36 frequencies of the next class interval, giving a total of 74 cumulated frequencies up to $7\frac{1}{2}$ months. This same procedure is carried out for the remaining class intervals, each in turn, the total cumulation equaling 131 frequencies for the final class interval, 12-months sitting age.

The Percentage Cumulative Frequency Distribution

As the name implies, a *percentage cumulative frequency distribution* represents the conversion of the frequencies of a cumulated distribution to percentage values, the total number of cases in the distribution being taken as

100%. The cumulative percentage values of each class interval for the sitting-age data are given in the last column of Table 5:13.

The simplest method of converting cumulative frequencies to percentage cumulative frequencies is first to compute the percentage value of one frequency for a given distribution. As already indicated for percentage frequency distributions, the percentage value of a frequency is always equal to

$$\frac{1}{N}(100)$$

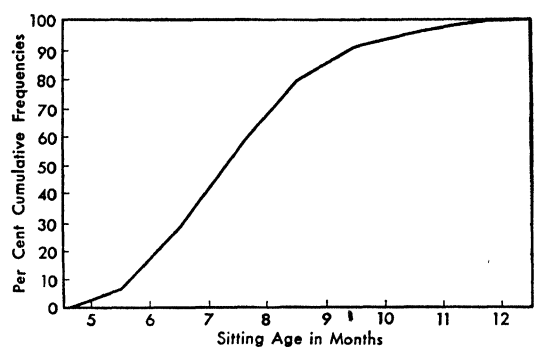
where N is the total number of cases in the distribution. The percentage value of one frequency for the distribution in Table 5:13 is 0.763%. The number of cumulated frequencies for each class interval is therefore multiplied by this value in order to secure the percentage values of the cumulated frequencies for each class interval.

If a percentage frequency distribution of a variable is already available, the percentage *cumulative* frequency distribution can of course be obtained by directly cumulating the *percentage* frequency values of successive intervals.

The Graphic Presentation of a Percentage Cumulative Distribution

A percentage cumulative frequency distribution is easily graphed and is illustrated in Fig. 5:11. The procedure for laying off the axes is similar to that used in graphing the percentage frequency distributions in Fig. 5:10. The vertical or ordinate scale is laid off in percentages beginning with zero and ending with 100, and the variable is scaled on the x-axis. However, when the frequencies have been cumulated from the lower end of the distribution, the percentage cumulated frequencies of a given class interval are always plotted at the upper *mathematical* limit of the interval, because the cumulated frequencies of a given interval are equal to the sum of all the frequencies of the distribution up *through* that interval.

Fig. 5:11. Cumulative Per Cent Frequency Curve. (Based on the Data in Table 5:13)



The Ogive

Sometimes the two scales are reversed. That is, the variable itself is laid off on the ordinate and the cumulated percentage frequencies are scaled on the

abscissa. When this is done, the resulting figure is called an *ogive*. Formerly the percentage cumulative frequency distribution was more often scaled to give an ogive than the curve in Fig. 5:11. A simple reversal of the axis position of the scales does not change the essential character of the graph information. However, since it is customary to plot all frequency distributions with the frequencies scaled on the ordinate axis, we shall follow this practice for the percentage cumulative frequency graph. The result serves research needs as effectively as the ogive, which is less convenient to use.

Usefulness of Percentage Cumulative Graph for Comparing Distributions

The percentage cumulative frequency distribution is of considerable value for *comparing* two or more distributions of a variable—not only as a whole, but also at any corresponding points. In Fig. 5:12, the two percentage frequency distributions in Fig. 5:10 have been plotted as percentage *cumulative* frequency curves. Although it is evident from an inspection of Fig. 5:10 that there is a tendency for the x-group of 131 infants to have older sitting ages than the y-group of 261 infants, that graph does not provide as satisfactory a basis for analyzing detailed differences throughout the range of the two distributions as does Fig. 5:12.

Fig. 5:12. Comparison by Cumulative Percentage Frequency Distributions of Infants' Sitting Ages. (Based on Same Data as Those in Fig. 5:9)

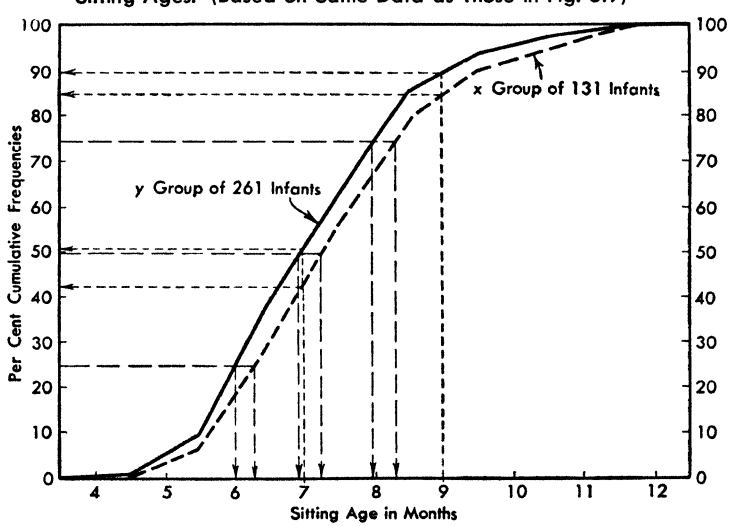


Fig. 5:12 is useful for comparing either *percentage frequencies* or *sitting ages*. For comparing percentage frequencies, horizontal lines are projected *from* three points (25%, 50%, and 75%) on the cumulative percentage scale to the

two curves and *down* to the *corresponding* sitting-age values for each group. Whereas the first 25% of the y-group had sitting ages of 6 months and less, the first 25% of the x-group had sitting ages as great as 6 months and 1 week. This tendency for the y-group to have an earlier sitting age is present throughout the distributions. Thus, 75% of the infants of the y-group were sitting alone by 8 months, but all 75% of the x-group had not attained this stage of development until about a week later. In fact, the difference between the two groups consistently averages about $\frac{1}{4}$ month sitting age, beginning at 6 months and continuing through the scale.

Comparisons can be made, on the other hand, *from* sitting-age values. For example, what percentage of the x- and y-groups was sitting alone by the age of 7 months? This question can be readily answered by projecting a perpendicular line from the 7-month point on the sitting-age scale to the two curves and then across to the corresponding percentage values of each group. Only 42% of the x-group were sitting unaided by 7 months, as compared to 50% of the y-group. At 9 months, 85% of the x-group and 90% of the y-group were sitting unaided.

EXERCISES

The following exercises are based upon the data in Table 5:14, which consists of three variables and the results on each variable for two groups of college students (100 in each group) composed of *college freshmen* on the one hand, and their *best friends* (among the freshman group) on the other.

The scores of each group are given by pairs, so that each college freshman's results are paired with those of his best friend. The pairs are numbered in column (1) from 1–100.

Variable G (columns 2 and 3) consists of the average grades made during the freshman year by the students in the two groups. (These scores are not the original percentage grades, but are measures converted to a scale with a range from zero to 99.)

The intelligence test variable, I.T. (columns 4 and 5), consists of the scores made by the students on an intelligence test administered when they entered college.

The third variable, A (columns 6 and 7), consists of the age of each student to the nearest year at the time of his admission to college.

1. Determine the range of the results of the three variables for each group of students.
2. Set up an array for the intelligence test scores of either student group.
3. Establish class intervals of an appropriate size for each of the three variables.
4. Differentiate the mathematical limits from the written limits for the class intervals of each of the three variables.
5. What are the mid-point values of the class intervals set up for each of the three variables?
6. Make a tally and frequency distribution of the results of the two groups for each of the three variables, and compare the college freshmen's results on each variable with those of their best friends.
7. Compare the results of the two groups for each of the three variables by means of a histogram for the freshmen and a frequency polygon for their best friends.

Table 5:14. Average Grades, Intelligence Test Scores, and Ages of 100 College Freshmen (F) and of Their Best Friends in the Freshman Class (B)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(1)	(2)	(3)	(4)	(5)	(6)	(7)
No.	Grade		I.T.		Age		No.	Grade		I.T.		Age	
	(F)	(B)	(F)	(B)	(F)	(B)		(F)	(B)	(F)	(B)	(F)	(B)
1	72	46	90	74	15	18	51	32	48	80	86	18	17
2	53	73	83	85	17	16	52	34	34	97	84	17	17
3	44	41	77	66	17	17	53	37	38	71	71	19	22
4	21	18	87	77	17	18	54	57	27	82	72	17	17
5	49	33	93	77	18	18	55	48	92	93	80	17	18
6	53	41	85	86	17	19	56	0	35	79	78	17	17
7	18	32	59	71	22	24	57	39	18	78	78	18	17
8	41	50	89	79	17	18	58	67	59	86	95	17	17
9	38	37	71	71	24	19	59	73	67	76	79	17	16
10	45	42	80	85	18	17	60	48	27	87	62	17	18
11	72	68	90	71	17	18	61	92	53	90	85	18	17
12	27	29	57	61	22	18	62	59	76	87	83	17	15
13	60	99	86	94	17	17	63	52	52	70	93	18	17
14	26	26	65	79	17	18	64	41	37	66	72	17	18
15	69	52	99	90	17	17	65	50	80	81	89	17	17
16	51	65	87	88	17	18	66	34	34	84	97	17	17
17	61	69	95	94	16	16	67	55	78	102	108	16	17
18	9	29	77	75	19	18	68	52	61	84	78	17	19
19	22	50	84	81	17	17	69	13	32	86	71	18	24
20	32	13	71	86	24	18	70	22	43	81	79	17	18
21	42	45	85	80	17	18	71	43	29	77	76	19	19
22	52	67	93	93	17	17	72	36	56	81	91	19	17
23	51	55	76	102	17	16	73	69	61	94	95	16	16
24	99	94	94	86	17	17	74	22	31	82	91	21	18
25	74	79	90	84	17	16	75	69	57	91	91	16	16
26	74	53	104	78	18	19	76	40	40	83	66	19	18
27	73	53	85	83	16	17	77	93	67	85	65	16	17
28	18	21	77	87	18	17	78	41	62	66	89	18	17
29	38	21	73	77	18	17	79	61	52	78	84	19	17
30	25	11	79	83	18	19	80	43	40	83	83	21	19
31	67	50	86	90	17	17	81	37	41	72	66	18	17
32	55	62	84	87	16	17	82	71	81	94	87	16	15
33	24	64	83	95	18	17	83	17	42	73	71	16	18
34	79	74	84	90	16	17	84	74	71	92	94	19	16
35	32	43	91	77	18	19	85	68	72	71	90	18	17
36	62	36	89	74	17	17	86	11	25	83	79	19	18
37	29	59	61	87	19	18	87	81	71	87	94	15	16
38	27	57	72	82	17	17	88	37	29	78	61	17	18
39	37	46	87	91	17	15	89	27	21	83	87	20	17
40	46	53	91	83	15	17	90	33	37	75	72	16	18
41	52	67	89	91	17	16	91	77	44	68	70	20	17
42	69	64	86	83	17	17	92	52	53	67	78	17	19
43	53	40	78	83	20	18	93	52	38	72	84	16	17
44	38	45	82	84	17	18	94	21	32	67	83	18	17
45	45	50	82	81	17	17	95	44	77	70	68	17	20
46	69	29	90	69	18	27	96	64	24	95	83	17	18
47	83	66	114	75	17	17	97	66	63	79	109	16	16
48	53	52	78	57	19	17	98	72	38	85	84	16	17
49	67	93	65	85	17	16	99	32	21	83	67	17	18
50	32	43	71	71	18	17	100	28	18	67	78	18	17

8. Take the first 50 cases of the freshman group for each of the three variables, make a percentage frequency distribution of these sub-groups, and compare and interpret the results with those for the total group by means of (a) percentage frequency polygons, and (b) cumulative percentage frequency distributions.

CHAPTER 6

The Centile Point Method for Variate Data

A. CENTILES AND THE DESCRIPTION OF VARIATE DATA

A centile point is a value on the score scale of a variable such that a given *percentage* of the frequencies of the distribution lies above the given point value and the remaining percentage of the frequencies lies below the given point value. A complete scale of centile values divides a distribution of frequencies into 100 equal parts, so that the total frequencies are divided into successive groups, each of which includes 1% of all the frequencies. A given centile point value always means exactly what it says, viz., that a certain percentage of the frequencies is located in the distribution above the given centile value, the remaining frequencies below. Thus, the 33rd centile of a distribution is always a value on the scale of measures such that 67% of the frequencies are above this value and 33% are below.

The purpose of any centile value is to provide a measure of a variable that will serve to summarize an aspect of the distribution. The centile method is a particularly valuable statistical technique because the basic interpretation of centile measures is not limited by the *form* of the distribution.

Centiles were originally called *percentiles*, and this term is still in considerable use. However, we have abbreviated the term to *centiles* for the sake of simplicity and of consistency with other commonly used measures which are based on the centile method, such as quartiles (never "per quartiles"), deciles (not "per deciles"), etc. We shall symbolize a centile point value by C and the appropriate numerical subscript to identify the particular value; thus, the 33rd centile point value is symbolized by C_{33}; the 56th centile by C_{56}, etc.

The basic assumption in using the centile method for describing the distribution of a variable is that the measures or scores of the distribution form a *continuous* series. This assumption, as we have seen, is inherent in the definition of a variable; that is, a variable yields a continuous series of measures ranging from least to most. It should be emphasized that in statistics this assumption is followed, even though the actual distribution of frequencies may be based on observations that are discrete in character, as for example, a distribution that gives for a city the number of children per family.

Centile Point Values vs. Centile Intervals

In computing and using centiles to summarize the data of a variable, it is important to distinguish clearly between a centile point value and a centile interval.

A Centile Point Value

A centile point is the value of a point on a scale of measures or scores which divides the *frequencies* of the distribution into two parts such that the sum of the two parts is equal to *all* the frequencies. Like a knife edge, it divides N, the total number of frequencies, into two parts. In practice, some frequencies of a distribution may have integral values corresponding to integral values of centiles. Such frequencies are ordinarily identified with the interval immediately above the given centile point value. Thus, if C_{50} is equal to 76.0, a score of 76 is located in the centile interval whose lower limit is 76.0. A frequency at the extreme upper range of scores is identified with the centile interval whose lower limit is C_{99}, because there can be no frequencies beyond C_{100}.

There are, by definition, 101 centile point values for any distribution. These values range from zero (C_0) to 100 (C_{100}). A full scale of centile point values, therefore, divides the *frequencies* of a distribution into 100 equal intervals, or parts. The 50th centile point value (C_{50}) divides the distribution of frequencies into two equal parts such that 50% of them are distributed above this centile point value and 50% are below. The 50th centile point value is commonly called the *median* of a distribution.

A Centile Interval

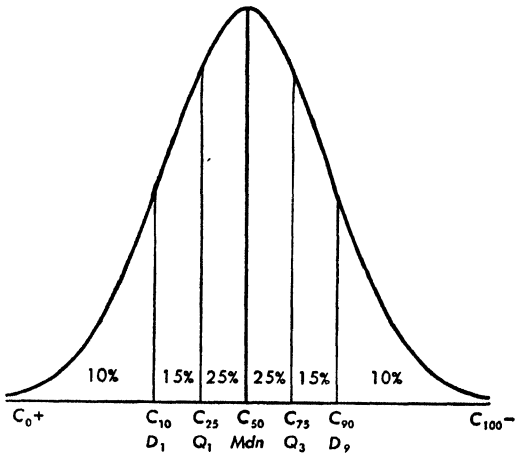
A centile interval is, by definition, the range between any two successive centile points of a distribution. It consequently includes 1% of the frequencies. The first centile interval of any distribution lies between C_0 and C_1—in other words, between the extreme lower range and the score value of the first centile. The 50th centile interval is the score range between C_{49} and C_{50}. The 100th centile interval is the range between C_{99} and C_{100}. Any centile interval of any distribution thus includes exactly 1% of the frequencies, although the actual score range of such intervals may vary considerably. For example, in a distribution similar to the normal frequency type, the score range of measures for centile intervals near the center of the distribution is much less than the range of centile intervals near either extreme of the distribution. This is illustrated in Fig. 6:1, in which centile intervals of a bell-shaped distribution are compared with centile intervals of a rectangular and a J-type distribution.

Quartiles, Terciles, Quintiles, Deciles, and Vigintiles

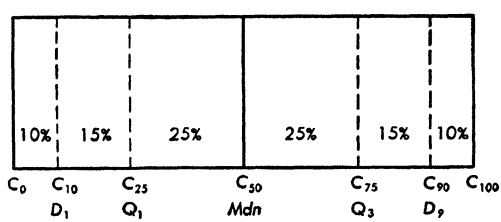
In practice, several centile point values are used for purposes of summarization. Thus, in addition to the median (C_{50}) already mentioned, C_{25} and C_{75} are used and described as the lower quartile (Q_1) and the upper quartile (Q_3) points within a distribution (Q_2 is the same as C_{50}, the median). The range between C_{25} and C_{75} is called the *inter-quartile range* and always includes the middle 50% of the frequencies of any distribution (see Fig. 6:1). Any one of the four quartile intervals includes 25% of the frequencies.

The range from C_{33} to C_{67} is called the *inter-tercile* range and includes the

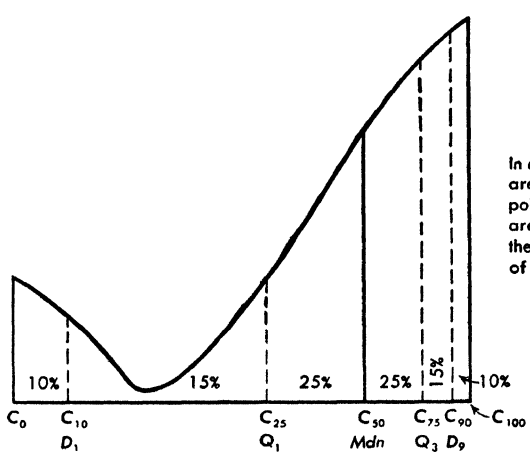
Fig. 6:1. Centile Intervals of a Normal Bell-Shaped Distribution Compared with Similar Intervals of a Rectangular and a J-Type Distribution. (All Three Distributions Have Similar Areas.)



Centile intervals of a normal, bell-shaped type of distribution are narrow in range at the center of the distribution because of the great concentration of frequencies around the median, or modal point.



In a rectangular type of distribution, all centile intervals are equal in size because the frequencies are uniformly distributed throughout the scale.



In a J-type distribution, centile intervals are unequal in size; futhermore, centile points above and below the median are not symmetrically distributed as in the bell-shaped and rectangular types of distributions.

middle *one-third* of the frequencies. C_{33} is ordinarily identified as the lower tercile (T_1) and C_{67} as the upper tercile (T_2). These tercile points, T_1 and T_2, thus divide a distribution of frequencies into three equal parts, whereas the quartile points divide it into four equal parts. C_{10} and C_{90} are also commonly used as centile point values for the summarization of a distribution. The range from C_{10} to C_{90} includes the middle 80% of the frequencies and is called the *D range* ($D = C_{90} - C_{10}$).

Vigintiles, by definition, are the centile point values of a distribution for successive intervals that include 5% of the frequencies. The first vigintile point value, Vn_1, is equal to C_5; $Vn_2 = C_{10}$; $Vn_3 = C_{15}$; $\cdots$ $Vn_{20} = C_{100}$. The first *vigintile interval* lies between Vn_0 and Vn_1 or C_0 and C_5; the twentieth vigintile interval, between Vn_{19} and Vn_{20} (or C_{95} and C_{100}).

Deciles give the point values of a distribution for successive intervals that include 10% of the frequencies. The first decile point value, D_1, is equal to C_{10}; $D_2 = C_{20}$; $D_3 = C_{30}$; $\cdots$ $D_{10} = C_{100}$. The first *decile interval* lies within the range of D_0 and D_1 (or C_0 and C_{10}); the second decile interval is D_1 to D_2 (or C_{10} to C_{20}); the tenth decile interval is D_9 to D_{10} (or C_{90} to C_{100}).

Distributions of frequencies are sometimes divided into *five* equal parts, instead of twenty, ten, four, or three; in this case the divisions are known as *quintile intervals*. The first quintile point value, Qn_1, is at C_{20}; $Qn_2 = C_{40}$; $\cdots$ $Qn_5 = C_{100}$. The first quintile interval lies within the range of Qn_0 and Qn_1 (or C_0 and C_{20}); the second quintile interval is Qn_1 to Qn_2 (or C_{20} to C_{40}); the fifth quintile interval is Qn_4 to Qn_5 (or C_{80} to C_{100}).

Tercile, quartile, and quintile divisions of a frequency distribution are extensively employed in psychological measurement, particularly in analyzing the functional implications of a test. Thus, a test is a useful instrument for measurement if a large proportion of those whose test scores are in the upper tercile also prove to do successful or satisfactory work in a given job, and if those whose test scores are in the lower tercile prove generally to do unsatisfactory work.

Summary of Some Commonly Used Centile Measures

Type	Number of Intervals	Measure of Dispersion
Centiles (*C*)	100	Range, C_0 to C_{100}
Vigintiles (*Vn*)	20	. . . . .
Deciles (*D*)	10	*D* range, D_1 to D_9
Quintiles (*Qn*)	5	. . . . .
Quartiles (*Q*)	4	Inter-quartile range, Q_1 to Q_3
Terciles (*T*)	3	Inter-tercile range, T_1 to T_2

In addition to these measures, the centile measures of *deviation*, viz., the quartile deviation and the tercile deviation, will be developed later in this chapter.

Comparative Implications of Centile Measures

Centile point values and the various statistical measures derived from them provide descriptive statistics that are applicable to any kind of variate distribution. However, caution is necessary in using centile values for summarizing a distribution because these values are computed with respect to the distribution of *frequencies*, rather than with respect to the unit values of the scores. Although the 50th centile interval, for example, always gives the range of 1% of the frequencies between C_{49} and C_{50} for any distribution, the scale values of scores will, as Fig. 6:1 showed, have different implications for different forms of distribution. For a normal distribution the score range of the 50th centile interval is relatively small as compared to the range of a centile interval at either extreme of the distribution. On the other hand, for a J-type distribution, the score range of centile intervals is smaller at the extremes than near the middle of the scale.

It is not misleading to compare the same centile intervals of two distributions having the same form; however, when two distributions differ markedly in form, such comparisons often lead to ambiguous or erroneous interpretations. This is why it is always well for the investigator to ascertain the forms of the distributions with which he is working, in order to know whether, in reality, they are comparable. The median, for example, is a measure of *central tendency* only when a distribution shows a central tendency, that is, when the greatest concentration of frequencies is near its center.

The Determination of Centiles

In practice, two methods are used for determining any centile value of a distribution. One is computational; the other, graphic. Both provide centile point values from which the ranges of centile intervals and other measures derived from centiles can be determined. For most practical purposes, the graphic method is as satisfactory as the computational method. We shall illustrate the graphic method first, since it serves the double purpose of yielding the desired centile determinations and of demonstrating what is basically involved in the centile method.

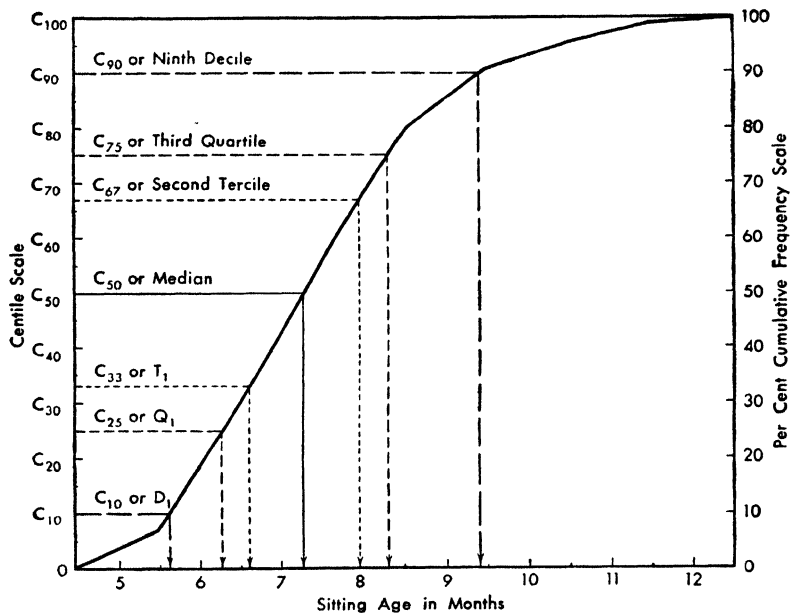
B. CENTILES BY THE GRAPHIC METHOD

The Centile Graph

The data in Table 5:13 were used for Fig. 5:11, showing the *percentage* cumulative frequency distribution. These data showed the distribution of the sitting ages of a group of 131 infants, i.e., the ages, in months, at which each infant was first able to support himself alone in a sitting position for at least one minute. Nine were able to do this at 5 months, and all but one could do it at the age of 12 months. The same data are used for Fig. 6:2, and from it any centile point values of a distribution can readily be estimated. By using the

centile method for bringing together additional summary facts about these data, we can obtain the sitting ages for any proportion of the group, as for example the range in sitting ages of the earliest 50%, or of the middle 50%, or of the last 10%, etc. In order to arrange the data of the frequency distribution in a form convenient for making the centile graph in Fig. 6:2, we obtain the cumulative frequencies and the percentage *cumulative* frequencies by the methods described in Chapter 5.

Fig. 6:2. The Centile Graph. (Based on the Sitting-Age Data of 131 Infants, Table 5:13)



In making a centile graph, the initial procedure is the same as for the percentage cumulative frequency curve in Figs. 5:11 and 5:12. In addition, as indicated in Fig. 6:2, the scale of centile values is laid off on the ordinate at the left side of the chart, with the percentage cumulative frequencies at corresponding ordinate points on the right side of the chart. As usual, the scale of *measures* (sitting age) is laid off on the abscissa at the bottom of the chart. The mid-point values of the class intervals, rather than the values of their limits, are usually noted on this scale.

In practice, the centile graph is drawn on millimeter paper so as to provide many subdivisions on both the centile and the score scales. Fairly accurate determinations of centile point values are thereby obtainable, especially if the entire graph is scaled on a large sheet of millimeter paper. In fact, with

extra large paper and consequently a graph that is very large and carefully and accurately drawn, estimates of centile values can be just as accurate for any practical purpose as *computed* values.

Determining the Score Values of Centiles from a Centile Graph

In order to make centile point estimates from a centile graph, the centile whose value is sought is first located on the ordinate scale. With this as the starting point, the scale value of the given centile is obtained on the abscissa by means of a vertical line projected down from a point on the percentage cumulative frequency curve, the point being exactly opposite the centile point whose value is to be determined. This is illustrated by the horizontal and vertical projections on Fig. 6:2. Thus, the position of C_{90}, the 90th centile (or ninth decile), on the ordinate centile scale is projected horizontally to the curve and a vertical line is dropped from this point to the scale of measures on the abscissa. Each projected line is thus drawn perpendicularly to its respective scale. The value of C_{90} is seen to be equal to 9.4 months sitting age. Similarly, the value of C_{75} is determined from the centile graph as equal to 8.3 months.

Table 6:1 brings together the estimates of the seven centile point values located on the centile graph in Fig. 6:2. These values enable the median, the inter-quartile range, the inter-tercile range, and the *D* range to be readily stated, thus providing useful descriptive information about a variate distribution.

Table 6:1. Determination of Centile Values from the Centile Graph in Fig. 6:2

Centile Point	Sitting Age
C_{90} (or D_9, the 9th decile)	9.4 months
C_{75} (or Q_3, the 3rd quartile)	8.3 months
C_{67} (or T_2, the 2nd tercile)	8.0 months
C_{50} (or *Mdn*, the median)	7.3 months
C_{33} (or T_1, the 1st tercile)	6.6 months
C_{25} (or Q_1, the 1st quartile)	6.3 months
C_{10} (or D_1, the 1st decile)	5.6 months

Median (C_{50}) = 7.3 months
Inter-quartile range (C_{25} to C_{75}) = 6.3 to 8.3 months
Inter-tercile range (C_{33} to C_{67}) = 6.6 to 8.0 months
D range (C_{10} to C_{90}) = 5.6 to 9.4 months

Vigintiles, quintiles, and any other centile values can be readily obtained from the centile graph in Fig. 6:2. If many such measures are needed, however, it is well to set up the graph with larger dimensions so that the values can be accurately read from it.

The foregoing procedure for determining the score value of any centile point can be reversed to yield the centile interval value for any score or meas-

ure on the abscissa scale. Thus, a sitting age of 5 months is in the 3rd centile interval; a sitting age of 8 months is in the 67th centile interval, etc.

C. THE COMPUTATION OF CENTILE VALUES

The *computation* of any centile value for a variable involves (1) *locating* the desired centile point value in the *distribution of frequencies*, and (2) determining the score value at the point thus located.

We shall illustrate these two steps for the data in Fig. 6:2, the original distribution of whose frequencies is shown in Table 6:2. As indicated in Chapter 5, the computation of centile values is simplified by using the *cumulated* frequency distribution. The usual cumulation of frequencies from the *lower* end of the distribution is presented in the next to the last column of Table 6:2. The last column shows the frequencies cumulated from the upper end in order to simplify *checking* all centile value computations made from the lower end of the distribution.

Table 6:2. Distribution of the Sitting Ages of 131 Infants (with Cumulative Frequency Distribution for Aid in Computing and Checking Centile Values)

Sitting Age in Months	Class-Interval Limits	f	c.f. (from 4.5 Months)	c.f. (from 12.5 Months)
12	11.5 to 12.5	1	131	1
11	10.5 to 11.5	5	130	6
10	9.5 to 10.5	6	125	12
9	8.5 to 9.5	15	119	27
8	7.5 to 8.5	30	104	57
7	6.5 to 7.5	36	74	93
6	5.5 to 6.5	29	38	122
5	4.5 to 5.5	9	9	131
		$N = 131$		

The Location of a Centile Point

A given centile point in a distribution of frequencies is located by computing its corresponding *percentage of* N, where N is, as usual, the total number of frequencies. Thus, C_{50} is located in the distribution in Table 6:2 as follows:

The percentage value of $C_{50} = \frac{50}{100}$, and $N = 131$.

Therefore $\frac{50}{100}(N) = \frac{1}{2}(131) = \frac{131}{2} = 65.5$

C_{50} is consequently located as a point value such that 65.5 of the frequencies are above this point and 65.5 are below this point. Frequencies are often fractionated in order to compute a centile point value (the distribution is assumed to be continuous).

We now need to determine the score value of the point value which is at the limit of the 65.5th frequency in the distribution of 131 frequencies.

Inspection of the cumulated frequencies in Table 6:2 reveals that this point value is in the class interval 6.5 to 7.5 months. This is the case since the upper limit of this interval includes 74 cases cumulated from the lower end of the distribution, whereas the upper limit of the preceding class interval (5.5 to 6.5 months) includes only 38 cases. Having thus located the class interval which contains the desired centile point value, we next proceed to *interpolate* its score value in the interval.

Interpolating the Score Value of a Centile Point

The *proportion* of the frequencies in the class interval, 6.5 to 7.5 months, that will include exactly 65.5 of the cumulated frequencies must first be determined. There are 36 frequencies in the class interval in which the 65.5th frequency is located; hence, the desired proportion is equal to

$$\frac{65.5 - 38}{36} = \frac{27.5}{36} = .76$$

where 38 is the total number of frequencies below the class interval 6.5 to 7.5 months.

This result, .76, represents the proportion of the frequencies in the lower part of the seven-month class interval which, together with the 38 frequencies cumulated through the six-month interval, will give a total of 65.5 frequencies. There are 38 frequencies below the seven-month interval. In order to reach 65.5, 27.5 more are needed. Since the seven-month interval has a total of 36 frequencies, $\frac{27.5}{36}$ gives the proper proportion of them, namely, .76.

The score value of C_{50}, thus located within the interval, will be equal to the lower mathematical limit of the seven-month interval, *plus* .76 of the range of the interval. Since the unit size of the interval is one month of sitting age,

$$C_{50} = 6.5 + .76(1.0) = 7.26, \text{ or } 7.3 \text{ months}$$

The value of C_{50} for this distribution is therefore 7.3 months sitting age. This is also the median value, since, as we have seen, the median of a distribution is located at C_{50}.

In interpolating a score value for a centile point, it is assumed that the frequencies are *uniformly* distributed throughout the interval in which the point is located.

An alternative procedure for the interpolation is as follows: The *score value* of *one* frequency, for the interval in which the centile point value is located, is first computed. In the above example, the score value of one frequency in the seven-month interval is equal to

$$\frac{1}{n_i}(i) = \frac{1}{36}(1.0) = .028$$

where n_i equals the number of frequencies in the interval and i is the unit size of the class interval—in this case, 1 month.

The score value of each frequency in this interval is therefore equal to .028 month of sitting age. Having already determined that the first 27.5 frequencies of this interval are needed in order to arrive at a point which will include the lowest 65.5 frequencies of the whole distribution, we therefore multiply these 27.5 frequencies by .028 and add the result to the lower mathematical limit value of the seven-month class interval. Thus,

$$C_{50} = 6.5 + 27.5(.028) = 7.27, \text{ or } 7.3 \text{ months}$$

This value should of course be the same, except for dropped decimals, as that obtained with the first interpolation procedure.

From the preceding development, the formula for the centile may be stated as follows: When the frequencies are cumulated from the *low-score end* of a distribution:

$$C_C = X_l + \left(\frac{pN - f_b}{f_i}\right) i \qquad \begin{matrix}[6:1] \\ \text{Any centile } C\end{matrix}$$

where X_l is equal to the mathematical value of the *lower* limit of the interval in which the desired centile value is located; p is the proportion of the distribution needed for any particular centile value, as for example, $p = \frac{33}{100}$ when C_{33} is desired; N is the total number of frequencies in the distribution; f_b is the number of frequencies *below* the lower limit, X_l; f_i is the number of frequencies in the *interval* in which the centile is located; and i is the size of the class interval.

Checking the Computed Centile Value

It is important in statistical work not only to check computations, but, if possible, to employ a method of checking which is relatively independent of the particular steps used in making the original computations. Such a method is readily found by working from the upper end of the distribution. The centile point value of C_{50} will be located so as to include the same number of frequencies from the upper end of the distribution as from the lower end, since the value of C_{50} is located so as to divide the frequencies into halves.

Inspection of the last column of Table 6:2, in which the 131 frequencies have been cumulated from the upper end of the distribution to facilitate checking, reveals that 57 of the cumulated frequencies lie above the lower limit of the eight-month class interval. We need, therefore, to go to the next lower interval, viz., seven months, to find the point value for the 8.5 additional frequencies ($65.5 - 57 = 8.5$) needed to give a total of 65.5. The score value of this point is then interpolated as before. Thus,

$$\frac{65.5 - 57}{36} = \frac{8.5}{36} = .24$$

This interpolated value of .24 is multiplied by the range value of the interval and the product is *subtracted* from the *upper* mathematical limit of the seven-month interval in order to give the value of C_{50} on the score scale. Thus,

$$C_{50} = 7.5 - .24(1.0) = 7.26, \text{ or } 7.3 \text{ months}$$

If the second method of interpolation is used, we again determine the percentage value of *one* frequency in the seven-month interval and obtain the following:

$$C_{50} = 7.5 - 8.5(.028) = 7.26, \text{ or } 7.3 \text{ months}$$

The computations for the other six centile values previously estimated from the centile graph in Fig. 6:2 are presented in Table 6:3. It will be observed that the principles of computation just described for C_{50} apply for any centile value. Thus, the score value of C_{90} is determined by first locating its position in the distribution of frequencies: 9/10 of 131 = 117.9. Hence the value of C_{90} is such as to divide the frequencies into two parts, with 117.9 below C_{90} and the remainder (131 − 117.9 = 13.1) above C_{90}. The value of C_{90} is next found to be located (from the cumulative frequency distribution in Table 6:2) in the nine-month class interval (8.5 to 9.5). This interval has 15 frequencies and the interpolated value of C_{90} is computed to be 9.4 months of sitting age.

Table 6:3. The Computation of Centile Point Values
(For the Sitting-Age Data in Table 6:2)

C	Division of Frequencies (Cumulated from the Lower End of Distribution)	Location of Interval	Interpolated Value Within Class Interval	Value of Centile (in Months)
C_{90}	$\frac{90}{100}N = \frac{9}{10}(131) = 117.9$	8.5–9.5	$\frac{117.9 - 104}{15} = .93(1.0) = .93$	8.5 + .93 = 9.43 [or 9.4]
C_{75}	$\frac{75}{100}N = \frac{3}{4}(131) = 98.25$	7.5–8.5	$\frac{98.25 - 74}{30} = .81(1.0) = .81$	7.5 + .81 = 8.31 [or 8.3]
C_{67}	$\frac{67}{100}N = \frac{67}{100}(131) = 87.77$	7.5–8.5	$\frac{87.77 - 74}{30} = .46(1.0) = .46$	7.5 + .46 = 7.96 [or 8.0]
C_{50}	$\frac{50}{100}N = \frac{1}{2}(131) = 65.5$	6.5–7.5	$\frac{65.5 - 38}{36} = .76(1.0) = .76$	6.5 + .76 = 7.26 [or 7.3]
C_{33}	$\frac{33}{100}N = \frac{33}{100}(131) = 43.23$	6.5–7.5	$\frac{43.23 - 38}{36} = .15(1.0) = .15$	6.5 + .15 = 6.65 [or 6.6]
C_{25}	$\frac{25}{100}N = \frac{1}{4}(131) = 32.75$	5.5–6.5	$\frac{32.75 - 9}{29} = .82(1.0) = .82$	5.5 + .82 = 6.32 [or 6.3]
C_{10}	$\frac{10}{100}N = \frac{1}{10}(131) = 13.1$	5.5–6.5	$\frac{13.1 - 9}{29} = .14(1.0) = .14$	5.5 + .14 = 5.64 [or 5.6]

The procedure for checking any centile is also based on the same principles as were described for checking C_{50}. Thus, to check C_{90}, this centile is located from the upper end of the distribution. The point from this end that will

exactly correspond with C_{90}, taken from the lower end of the distribution, will of course be 10 per cent of the way down in the distribution of frequencies. Thus, 10/100 of 131 = 13.1, and this frequency is seen (from the last column in Table 6:2) to be located in the interval 8.5 to 9.5 months. There are 12 frequencies above the upper limits of this interval, and 15 within the interval. Hence, interpolating for C_{90} and *subtracting* the result from the upper limit of the interval, we have:

$$C_{90} = 9.5 - \frac{13.1 - 12}{15}(1\,0) = 9.5 - .07 = 9.43, \text{ or } 9.4 \text{ months}$$

The formula for checking any centile value, with the frequencies cumulated from the *high-score end* of the distribution, is as follows:

$$C_C = X_u - \left(\frac{pN - f_a}{f_i}\right) i \qquad [6:1a]$$

Any centile (check formula)

where X_u is equal to the mathematical value of the *upper* limit of the interval in which the desired centile value is located; f_a is the number of frequencies *above* the upper limit, X_u; and p, N, f_i, and i are the same as for Formula 6:1.

Comparison of Estimated and Computed Centile Values

If centile values are estimated from a carefully drawn centile graph, the results should be the same, or at least approximately the same, as computed centile values. In Table 6:4 the results as *estimated* from the centile graph in Fig. 6:2 are compared with the *computed* centile values in Table 6:3. It will be observed that the estimated and computed values in Table 6:4 are the same in all cases to within one-tenth of a month sitting age. For all practical purposes the results are "identical."

Table 6:4. Comparison of Estimated and Computed Centile Values for the Sitting-Age Data of 131 Infants

Centile Values	Sitting Age	
	Estimates from Centile Graph (Fig. 6:2, Table 6:1)	Computed Values (Table 6:3)
C_{90} D_9	9.4 months	9.4 months
C_{75} Q_3	8.3 "	8.3 "
C_{67} T_2	8.0 "	8.0 "
C_{50} *Mdn*	7.3 "	7.3 "
C_{33} T_1	6.6 "	6.6 "
C_{25} Q_1	6.3 "	6.3 "
C_{10} D_1	5.6 "	5.6 "

When many centile values are to be determined for a distribution, the centile graph is thus a labor-saving device that can yield as satisfactory

results as computed values. Furthermore, it has an advantage over the computational procedure in that it affords a picture of the *trend* of the results. Finally, once a centile graph is made, any centile values which may be needed later for comparative or other purposes can readily be read from the graph.

D. CENTILE MEASURES

The data of a variate distribution can be *summarized* by various measures based on point centile values. These centile measures, which have already been referred to as the median, terciles, quartiles, quintiles, deciles, and vigintiles, will now be described more fully.

The Median (A Measure of Central Tendency?)

The median, by definition,* is a centile point value in a scale of scores or measures such that the total distribution of frequencies is divided into two equal parts at that point. In other words, 50% of the frequencies are above the score value of the median and 50% are below it. The median always signifies exactly this, and is therefore equal to C_{50} (the value of the 50th centile point). Furthermore, for variables which show a *uni-modal* tendency near the center of the distribution, the median serves as a useful measure of central tendency.

Sometimes variables yield distributions which are uni-modal but at the same time skewed; that is, instead of being bilaterally symmetrical from the modal part of the distribution, the scores spread out much farther at one end than at the other. In such cases the median, as a measure of central tendency, usually provides a more typical score than does the arithmetic mean. This point will be developed further in the next chapter.

The formula for the median, which states the operations needed to determine the value of C_{50}, and is a special case of Formula 6:1, is as follows:

$$Mdn = X_l + \left(\frac{N/2 - f_b}{f_i}\right) i \qquad [6:1b]$$

Median (*Mdn*), special case of Formula 6:1

where X_l is equal to the mathematical value of the lower limit of the interval in which the median is located; N is the total number of frequencies in the distribution; f_b is the number of frequencies *below* the lower limit, X_l; f_i is the number of frequencies in the *interval* in which the median is located; and i is the size of the class interval.

---

* The median is also sometimes defined as the *mid-score* or *mid-case* of an array of measures. Although this concept provides a quick measure of the median when N is odd, it does not fit the point value concept of centiles as developed in this chapter. For example, it would lead to difficulties, in dealing with the data of distributions which are assumed to be continuous, to derive centiles, vigintiles, deciles, etc., in terms of the mid-score concept of the median. Hence we shall not be concerned with this concept; rather we shall compute the median, the value of C_{50}, as any other centile measure is computed.

The D Range—A Measure of Dispersion

The pairs of centile point values which are ordinarily used together to summarize the dispersion or spread of scores in a distribution were shown in Fig. 6:2 by similar types of vertical lines projected from the curve to the sitting-age scale. Thus, D_1 (or C_{10}) and D_9 (or C_{90}) are used to give the D range. It includes the range of the middle 80 per cent of the frequencies:

$$D = C_{90} - C_{10} \qquad [6:2]$$

D range

As was indicated in Table 6:1, the D range for the data in Fig. 6:2 was found to be 5.7 to 9.4. Thus, the middle 80% of the distribution of 131 infants have sitting ages ranging from 5.7 to 9.4 months. $D = 9.4 - 5.7 = 3.7$ months.

The Quartile Deviation—A Measure of Deviation or Variability

C_{25} and C_{75} have long been used to give the range of the middle 50% of the frequencies of a distribution. Since this is the range between the first and third quartile points (Q_1 and Q_3), it is usually called the inter-quartile range. It is ordinarily an even more stable part of a distribution than the D range.

When distributions tend to be bilaterally symmetrical with respect to the median, the inter-quartile range provides the basis for a very useful measure of variability, viz., *Q.D.*, the quartile deviation. *Q.D.* is equal to one-half the inter-quartile range:

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{C_{75} - C_{25}}{2} \qquad [6:3]$$

Quartile deviation (*Q.D.*)

The quartile deviation for the data in Fig. 6:2 is computed as follows:

$$Q.D. = \frac{C_{75} - C_{25}}{2} = \frac{8.3 - 6.3}{2} = \frac{2.0}{2} = 1.0 \text{ month}$$

For this distribution the median (C_{50}) happens to be exactly midway on the score scale between the values of the first and third quartiles (C_{25} and C_{75}). As indicated in Table 6:1, the median is 7.3, whereas the first quartile is 6.3 and the third quartile is 8.3. Consequently, in using the median *as a point of reference* to summarize the *deviational tendency* of this distribution of sitting-age data, we are warranted in stating the following:

$$Mdn \pm Q.D. = 7.3 \pm 1.0 = 6.3 \text{ to } 8.3 \text{ months}$$

The quartile deviation itself gives the range of 25% of the frequencies above or below this median, but the *median plus and minus the quartile deviation* gives exactly the range of the middle 50% of the cases. This relation of *Mdn* to *Q.D.* is of course misleading if the point values of the inter-quartile range are not bilaterally symmetrical with respect to the median. The use of the quartile deviation with the median to summarize the deviational tendency of a distribution is sound only when the *median plus and minus* Q.D. gives a range in score values that corresponds closely to the actual point values of

the inter-quartile range as given directly by C_{25} and C_{75}. Therefore, when distributions do not tend to be bilaterally symmetrical with respect to the median, the dispersion of the middle 50% of the frequencies should generally be summarized by citing the actual values of C_{25} and C_{75} rather than by computing the quartile deviation and using it with the median.

The Tercile Deviation—A Measure of Deviation or Variability

It is often useful to divide distributions of frequencies into three equal parts; hence, the tercile range. The first tercile interval is the range from the lowest score values of a distribution to the point value of C_{33}. This latter centile value is the first tercile point value (T_1); the second tercile point (T_2) is at C_{67}; and the third tercile value corresponds to C_{100}. The inter-tercile range is equal to the middle tercile interval T_1 to T_2 (or C_{33} to C_{67}) and for the data in Fig. 6:2 is 6.6 to 8.0. Hence, the middle 33% of the distribution of measures for 131 infants range from 6.6 to 8.0 months of sitting age.

A measure of deviation analogous to the quartile deviation has been also developed for the tercile range. The principle for its computation is the same; hence *T.D.*, the tercile deviation, may be symbolized as follows:

$$T.D. = \frac{T_2 - T_1}{2} = \frac{C_{67} - C_{33}}{2} \qquad [6:4]$$

Tercile deviation (*T.D.*)

The tercile deviation is thus equal to half the inter-tercile range, just as the quartile deviation is equal to half the inter-quartile range. For the sitting-age data in Fig. 6:2,

$$T.D. = \frac{8.0 - 6.6}{2} = \frac{1.4}{2} = .70 \text{ month}$$

The tercile deviation gives the range of one-sixth of the frequencies above or below the median, provided, of course, the value of the median is midway between the point values of T_1 and T_2. When this is the case, the median *plus and minus* T.D. gives the range of the middle one-third of the measures of a distribution. For the data in Fig. 6:2,

$$Mdn \pm T.D. = 7.3 \pm .70 = 6.6 \text{ to } 8.0 \text{ months}$$

These values correspond to the actual C_{33} and C_{67} point values; consequently, it is appropriate to use the median as a point of reference for a tercile measure of deviation to describe the results of this particular distribution.

Tercile divisions of a distribution have in recent years been employed by some investigators in order to set up a threefold differentiation of criterion scores used in the validity analysis of test results: average group (T_1 to T_2), above average group (T_2 and above), and below average group (below T_1). For such a purpose this is often a better division than quartiles in which an "average" group is taken as lying within the limits of the inter-quartile range (Q_1 to Q_3).

E. THE USE OF CENTILES FOR COMPARING THE RESULTS OF TWO OR MORE DISTRIBUTIONS OF A VARIABLE

Centiles are valuable not only for describing and summarizing important details of a distribution but also for *comparing* two or more distributions of a variable. For the latter purpose, either graphs or tables (or both) may be employed.

We shall present both a graph and a table to compare *teachers' salaries* as distributed for three school systems, X, Y, and Z, located in areas suburban to New York City.* The salary distributions for each of these systems are presented in Table 6:5. The number of teachers reported for school system X was 164; for school system Y, 151; for school system Z, 114. Both the original

Table 6:5. Distributions of Teachers' Salaries for Three Public School Systems in New York State

Annual Salaries	School System X			School System Y			School System Z		
	f	c.f.	% c.f.	f	c.f.	% c.f.	f	c.f.	% c.f.
$4301 to $4500							3	114	100.0
$4101 to $4300							4	111	97.4
$3901 to $4100				9	151	100.0	3	107	93.9
$3701 to $3900				11	142	94.7	18	104	91.2
$3501 to $3700	1	164	100.0	15	131	87.3	12	86	75.4
$3301 to $3500	16	163	99.4	10	116	77.3	18	74	64.9
$3101 to $3300	19	147	89.6	23	106	70.7	16	56	49.1
$2901 to $3100	8	128	78.0	27	83	55.3	9	40	35.1
$2701 to $2900	22	120	73.2	19	56	37.3	8	31	27.2
$2501 to $2700	57	98	59.8	12	37	24.7	16	23	20.2
$2301 to $2500	17	41	25.0	10	25	16.7	1	7	6.1
$2101 to $2300	13	24	14.6	6	15	10.0	5	6	5.3
$1901 to $2100	4	11	6.7	8	9	6.0	1	1	0.8
$1701 to $1900	3	7	4.3	1	1	0.7			
$1501 to $1700	3	4	2.4						
$1301 to $1500	1	1	0.6						
	N = 164			N = 151			N = 114		

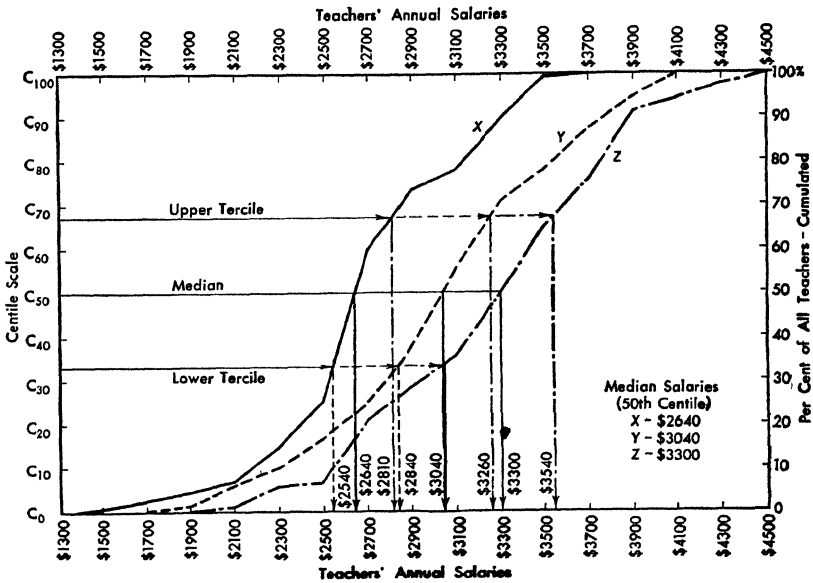
frequency distributions and the percentage cumulative frequency distributions are given in this table. These data provide the basis for the centile graphs shown in Fig. 6:3. We shall proceed to interpret the results of these graphs and then present a tabular comparison of teachers' salaries for each of the three school systems (Table 6:6).

The variable, teachers' salaries, is scaled at the bottom of Fig. 6:3; they range from $1300 to $4500 per year. The percentage cumulative frequency

* These data for teachers' salaries are based on figures assembled for the school year 1943–1944, and made available by Mr. Vernon G. Smith, Superintendent of Schools, Scarsdale, New York.

curves for each of the three schools are plotted according to the procedure already described for centile graphs.

Fig. 6:3. Comparison of Teachers' Salaries for Three School Systems in Areas Suburban to New York City for 1943–1944; Principals' Salaries Excluded



Only the upper and lower tercile values for each of the three school systems are shown in the figure. However, centile graphs are very convenient for the comparison of two or more distributions at any point thereof. Once the graphs are made, they can be referred to at any time, and any desired values quickly derived. Furthermore, they can be read from either direction. That is, the percentage of teachers above or below a certain salary, as well as the salary value of any given centile point, can be readily determined. For example, Fig. 6:3 reveals that none of the teachers in system X received an annual salary in excess of $3700, and, on the other hand, that none of the teachers in system Z were paid less than $1900 a year.

The median salaries for each system are readily obtained by projecting a horizontal line from the 50% point on the ordinate scale across to each of the three curves and dropping vertical lines at the three points of intersection to the base line. These medians are approximately $2640, $3040, and $3300, for X, Y, and Z respectively. The difference between the median salaries of X and Z is considerable, amounting to $660. However, there are also considerable differences at most points throughout the three distributions. Thus, examination of the lower and upper tercile values reveals that two-thirds of

the teachers in system X received an annual salary of *less than* \$2810 (T_2), whereas more than two-thirds of the teachers in both systems Y and Z received annual salaries *in excess* of \$2810. One-third of the teachers in system Y received salaries greater than \$3260 (T_2), and one-third of the teachers in system Z received salaries greater than \$3540 (T_2). On the other hand, one-third of the teachers in system X received salaries of less than \$2540 (T_1), whereas only 18% of the teachers in system Y and only 9% of the teachers in system Z received less than this amount.

Inspection of the slopes of the centile curves for each of the school systems is especially revealing. The steeper the slope, the less the dispersion or spread of salaries and the greater the concentration of cases. Thus, for school system X, the slope of the curve is very steep for salaries between \$2500 and \$2700. Within these relatively narrow limits, the salaries of 35% of the teachers in this system are located, since the centile point value of a salary of \$2500 is approximately C_{25}, and the centile point value of a salary of \$2700 is approximately C_{60}. The difference between C_{60} and C_{25} is equivalent to 35% of the cases. On the other hand, there is no such sharp and extended rise in the slope of the curves for systems Y and Z. This means that the salaries of the teachers in these systems were spread much more evenly through the scale than were the salaries paid in X.

As for *the tabular comparisons:* Four sets of centile values, commonly used for comparative purposes, are presented in Table 6:6. All have been determined directly from the centile graphs in Fig. 6:3, rather than computed from the frequency distributions shown in Table 6:5. They are the *D* range, the inter-tercile range, the inter-quartile range, and, finally, the median with the quartile and tercile deviations. Not all of these four sets of data are necessary for the tabular comparison of two or more distributions. Often only the median and quartile deviations are reported. Whether or not the *D* range and the inter-tercile range are also used depends on the nature of the data being compared, as well as upon the detail necessary for the report.

The *D* range, as we have seen, gives the dispersion of the middle 80% of the cases in a distribution. In all three school systems compared in Table 6:6, this range is greater than \$1000, being nearly \$1500 for the teachers in system Y. It will be observed, furthermore, that there is a difference of \$600 between the upper limits (C_{90}) of the *D* range for school systems X and Z. Ten per cent of the teachers in system Z received salaries in excess of \$3900, whereas the upper 10% in system X were paid between \$3300 (C_{90}) and \$3700 (upper limit).

The inter-tercile range was referred to in the description of Fig. 6:3. The dispersion of the middle one-third of the teachers' salaries varies from \$270 for X to \$500 for Z. Two-thirds of the teachers in system X received salaries of less than \$2810 (T_2), whereas two-thirds of the teachers in system Z received salaries of more than \$3040 (T_1).

The inter-quartile range for system X is \$500, being the same as the inter-

Table 6:6. Comparison of Teachers' Salaries for Three Public School Systems
(Determined from Centile Graphs in Fig. 6:3)

Centile Values	School Systems		
	X Salaries	Y Salaries	Z Salaries
D_9 (C_{90})	\$3300	\$3775	\$3900
D_1 (C_{10})	\$2175	\$2300	\$2550
D range	\$2175 to \$3300 = \$1125	\$2300 to \$3775 = \$1475	\$2550 to \$3900 = \$1350
T_2 (C_{67})	\$2810	\$3260	\$3540
T_1 (C_{33})	\$2540	\$2840	\$3040
Inter-tercile range	\$2540 to \$2810 = \$ 270	\$2840 to \$3260 = \$ 420	\$3040 to \$3540 = \$ 500
Q_3 (C_{75})	\$3000	\$3425	\$3700
Q_1 (C_{25})	\$2500	\$2725	\$2840
Inter-quartile range	\$2500 to \$3000 = \$ 500	\$2725 to \$3425 = \$ 700	\$2840 to \$3700 = \$ 860
Median (C_{50})	\$2640	\$3040	\$3300
Q.D.	\$ 250	\$ 350	\$ 430
T.D.	\$ 135	\$ 210	\$ 250

tercile range of the salaries in system Z. The middle 50% of the teachers' salaries for the latter system varied from \$2840 to \$3700, a range of \$860. Seventy-five per cent of the teachers in system X were paid less than \$3000 (Q_3), whereas 75% of the teachers of system Z received salaries of more than \$2840 (Q_1).

The median salaries in Table 6:6 indicate that half the teachers in system X received salaries of less than \$2640, whereas half the teachers in system Y received salaries in excess of \$3040, and half those in system Z received salaries in excess of \$3300.

The values of the quartile deviations and tercile deviations are equal to half the inter-quartile and inter-tercile ranges in the upper part of the table. If the medians for each of the distributions are approximately midway between their respective upper and lower quartile and tercile points, these ranges can be tersely summarized in terms of quartile deviation and tercile deviation. Let us examine first the results for school system X.

The median $\pm$ the quartile deviation of system X is equal to \$2640 $\pm$ \$250, which gives a range of \$2390 to \$2890. The actual lower and upper quartile point values for this distribution are \$2500 and \$3000. The median is not midway between these quartile points, and hence the results can be described more accurately in terms of the actual quartile points rather than in terms of the quartile deviation.

In system Y, the median ± the quartile deviation is equal to \$3040 ± \$350, which gives a range of \$2690 to \$3390. Since the actual quartile point values for this distribution are \$2725 and \$3425, the use of the latter, rather than the quartile deviation, to describe the result is again preferable.

In system Z, the median ± the quartile deviation is equal to \$3300 ± \$430, which gives a range of \$2870 to \$3730. The lower and upper quartile point values for this distribution are \$2840 and \$3700. Although system Z thus gives the closest correspondence of values, the quartile point values, rather than the quartile deviation, would be preferred in describing the results of this distribution.

The above procedure for describing the results, in which the three distributions are compared in terms of their medians and quartile deviations, is less satisfactory than a direct comparison of the quartile points and medians. This is true because the distributions are not bilaterally symmetrical with respect to their medians. On the other hand, inspection of *the tercile deviations* indicates that these measures *satisfactorily describe the variability* of the middle one-third of the cases in each distribution, with respect to the medians of each. Thus, the median ± the tercile deviation for system X is equal to \$2640 ± \$135. This gives a range of \$2505 to \$2775. The actual lower and upper tercile values are \$2540 and \$2810. In school system Y, the median ± the tercile deviation is equal to \$3040 ± \$210, which gives a range of \$2830 to \$3250. These values practically coincide with the actual lower and upper tercile values of \$2840 and \$3260. In system Z, the median ± the tercile deviation is equal to \$3300 ± \$250. This gives a range of from \$3050 to \$3550. The actual lower and upper tercile values are again practically the same, viz., \$3040 and \$3540.

Thus, the tercile deviation taken in relation to its median provides a satisfactory description of the mid-variability of each distribution, whereas for these particular distributions the quartile deviations are less satisfactory than the quartile points themselves. It is therefore apparent that quartile deviations and tercile deviations should not be used to summarize and compare the variability of distributions unless examination of the results shows that the median plus and minus either of these measures of deviation gives ranges that closely correspond to the actual quartile or tercile point values.

F. THE USE OF THE CENTILE METHOD FOR COMPARING THE RESULTS OF TWO OR MORE VARIABLES

The comparisons of teachers' salaries in three different school systems by means of centile graphs illustrate the detail in which the centile method can be used to compare two or more different distributions of the same variable. The same method is also useful for comparing two or more different variables, provided of course there is a logical basis for considering them

together. The procedure is aptly illustrated by the following market research data on two aspects of people's habits, viz., rising and breakfasting.*

Two samples totaling "6705 representative families" of New York City were personally interviewed by field interviewers of Crossley, Inc., for radio Station WOR of New York. About one-half of the total group was asked:

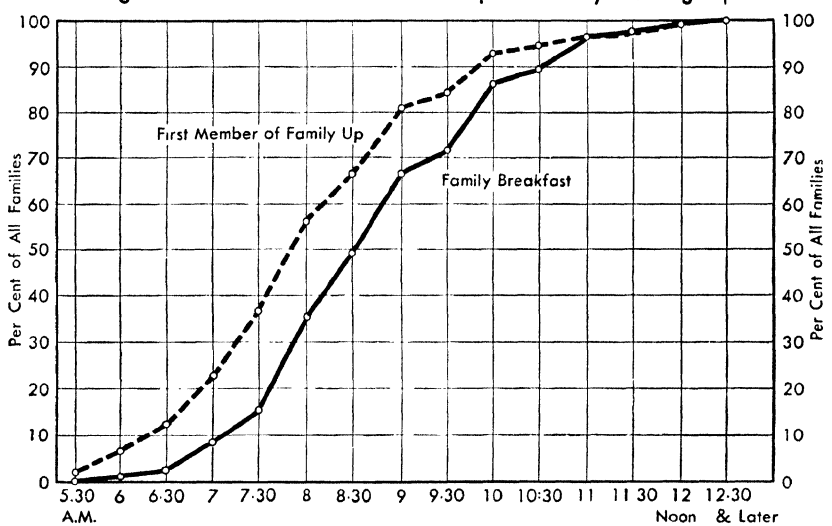
> "What was the earliest time last Sunday that a member of your family was up?"

The other half was asked:

> "What time did the family have breakfast last Sunday?"

The results of this market survey are summarized in Table 6:7, and the behavior of the two groups, each for its respective variable, is compared by the centile graphs in Fig. 6:4. The data are somewhat surprising, for it had been thought that most New Yorkers sleep much later Sunday mornings.

Fig. 6:4. "When New Yorkers Get Up on Sunday Mornings" †



† This figure reproduced by courtesy of Station WOR, New York, and the editors of *Broadcasting Magazine*.

The survey results indicate that the median first-riser of a family is up before 8 o'clock; two-thirds of the first-risers are up by 8:30; and 4 out of every 5 are up by 9 o'clock. Furthermore, by 8:30, half the families had eaten or were eating breakfast, and by 9 o'clock two-thirds of them had eaten or were eating. Less than 5% arose or breakfasted after 11 o'clock.

* Ray Lyon, "New Yorkers Early Risers," *Broadcasting Magazine*, March 26, 1945, p. 34.

Table 6:7. "When New Yorkers Get Up on Sunday Mornings"

Time of Morning	First Person Up f	Family Breakfast f	First Person Up % c.f.	Family Breakfast % c.f.
12:30 and later	9	20	100%	100%
12:00 noon	74	59	99.7	99.4
11:30	9	34	97.5	97.7
11:00	93	261	97.2	96.7
10:30	48	98	94.3	89.0
10:00	287	510	92.9	86.2
9:30	109	160	84.2	71.2
9:00	465	565	80.9	66.5
8:30	328	490	66.8	49.9
8:00	653	690	56.9	35.5
7:30	460	219	37.1	15.3
7:00	423	217	23.1	8.9
6:30	121	46	10.3	2.5
6:00	164	39	6.6	1.1
5:30 A.M. and earlier	54	0	1.6	0.0
	N = 3297	3408	100%	100%

The steepest part of the "family breakfast" curve in Fig. 6:1 would be the optimum time for breakfast radio programs, for this is the period when the greatest number is eating breakfast. It begins at 7:30 and lasts until 9.

EXERCISES

1. Why is the assumption of a continuously distributed variable essential to the use of the centile point method?
2. What are the implications of centile point values and centile intervals?
3. For any distribution, are there more frequencies between C_{85} and C_{95} than between C_{45} and C_{55}? Why?
4. What are the centile point limits of:
   a. the inter-quartile range
   b. the inter-tercile range
   c. the 8th decile interval
   d. the 3rd quartile interval
   e. the 3rd tercile interval
   f. the D range
   g. the range
   h. the 12th vigintile interval
   i. the 1st quintile interval
5. What proportion of the frequencies of a distribution lie within the limits of:
   a. the D range
   b. the inter-quartile range
   c. the inter-tercile range

6. Under what circumstances can and cannot a distribution be adequately summarized by:
   a. the median $\pm$ the quartile deviation
   b. the median $\pm$ the tercile deviation
7. In what centile interval do the following measures lie:
   a. the median
   b. the lower quartile (Q_1)
   c. the upper tercile (T_2)
   d. the 8th decile (D_8)
8. For each of the variables in Table 6:7, summarize and interpret the results in terms of the following centile measures, both by estimates from a centile graph and by computed centile values:
   a. median
   b. tercile deviation
   c. quartile deviation
   d. the D range
9. Determine whether the variation characteristics of each variable in Table 6:7 can be adequately summarized by:
   a. the median $\pm$ the quartile deviation
   b. the median $\pm$ the tercile deviation
10. Compare each of the three variables in Table 5:14 in terms of centile values derived from a centile graph, and interpret the results for the following:
    a. average grades of college freshmen and of their best friends
    b. intelligence test scores of college freshmen and of their best friends
    c. ages of college freshmen and of their best friends

CHAPTER 7

The Mean and Standard Deviation

A. THE METHOD OF MOMENTS FOR VARIATE DATA

We shall present in this chapter another statistical method that is widely used for the summarization and comparison of variate data. In contrast to the centile method developed in the preceding chapter, the statistical measures now to be discussed are based on deviations from the arithmetic mean of the distribution rather than on the location of frequencies at various points on a scale of measures.

The chief measures to be obtained are the arithmetic mean (M) and the standard deviation (σ). The method of their computation is essentially algebraic and is often described as the method of moments.*

Both the arithmetic mean and the standard deviation are widely used in sampling and analytical statistics, as well as in descriptive statistics. Another measure of deviation is the average deviation ($A.D.$). Since it is not used so frequently, it will be described, for reference purposes, at the end of this chapter.

Basic Symbols

Henceforth, it will be convenient to employ the following commonly used symbols for the basic measures and procedures of the method of moments:

1. A measure or score of a variate is symbolized by the capital letter X. If two or more distributions are compared, the measures of each are symbolized by different capital letters, for example, by X, Y, and Z, or A, B, C, $\cdots Z$.
2. Any particular measures of a variate are symbolized by numerical subscripts to the capital letter, as for example, X_1, X_2, $\cdots X_n$.

* The arithmetic mean of a distribution is a moment of the first order, and a standard deviation is the square root of the moment of the second order taken with respect to the mean.

The term "moment" as used in statistics has been taken from physics, where a moment is a measure of a force with respect to the tendency of the force to produce rotation. The strength of this tendency varies according to the amount of force and the distance from the point at which the force is applied. In a frequency distribution, the arithmetic mean is taken as the origin, and the frequencies of each class interval are taken as the forces at distances in terms of x_1, x_2, x_3 . . . x_n. The mean is the moment of the first order; it is symbolized by μ_1 and is equal to $\Sigma fx/N$. The moment of the second order is symbolized by μ_2 and is equal to $\Sigma f(x^2)/N$.

3. The *arithmetic summation* of a series of measures is symbolized by S.
4. The *algebraic summation* of a series of measures is symbolized by the Greek letter for capital S, viz., Σ.
5. The mean is symbolized by M. The median, as we have seen, is represented by *Mdn*. The mode (point or interval with the greatest concentration of frequencies) is symbolized by *Mo*. Thus these three symbols are clearly differentiated from one another.
6. A deviation is symbolized by the small letter corresponding to the capital letter used for the original measures or scores. Thus, x symbolizes the deviate value of X; y, the deviate value of Y.
7. Generally, and unless specifically indicated otherwise, a deviation, x, is always taken as the difference between an original measure, X, and the mean of the distribution from which the measure is derived. Hence,

$$x = X - M_x$$

   Deviations of measures greater in value than the mean are *positive;* of measures less in value, negative. The latter are symbolized by a *minus* sign.
8. Any particular deviations of a variate are symbolized by numerical subscripts, as for example, $x_1, x_2 \cdots x_n$.
9. Small letters are also used to symbolize a variable. Thus, a single variable is usually designated as variable x; a second variable as variable y; a third, as variable z. If there are more than three variables (but less than 27) in an investigation, they may be symbolized by letters from the beginning of the alphabet: variable a, variable b, etc., or each variable may be numbered in succession and designated as variable 1, variable 2, $\cdots$ etc.

Unfortunately, the symbols of statistics are not universally uniform. However, the preceding symbols are commonly used in statistical work developed for and applied to the biological and social sciences.

B. THE MEAN

Definition

The simplest and most common definition of the mean is that it is the *sum* of all the measures of a distribution divided by the total *number* (N) of measures. The mean is an average, and is often referred to as the arithmetic mean to distinguish it from other averages. However, the *sum* of all the measures is basically an algebraic rather than an arithmetical sum. Either method of summing will, of course, yield the same result if all the measures of a distribution are positive (or if all are negative) numbers, but not if the distributions include both negative and positive numbers.

The mathematical definition of the mean is as follows: A mean is a number such that the algebraic sum of the deviations of all measures from that number

is equal to zero. It is a measure for a distribution such that the sum of the positive deviations exactly equals the sum of the negative deviations. Thus, the mean of the following three numbers, 10, 6, and −10, is equal to:

$$\frac{X_1 + X_2 + X_3}{N} = \frac{10 + 6 + (-10)}{3} = \frac{6}{3} = 2.0$$

This result yields a number, 2.0, such that the algebraic sum of the deviations of each measure from 2.0 is zero. Thus, where $x = X - M_x$:

$$x_1 + x_2 + x_3 = (10 - 2) + (6 - 2) + (-10 - 2) = 8 + 4 + (-12) = 0$$

The mean is a measure that summarizes an essential aspect, the average, of a variable distribution. It is one of the most important measures in statistical theory, for it provides the investigator with a number which represents the average value or *size* of all the different measures of a distribution. The mean is a measure of the *central tendency* of distributions that tend to be of the bell-shaped type. Since the frequencies are more concentrated near the central part of a bell-shaped distribution than at any other part, the mean is an index of the most *typical* measure of such a series. However, the mean is a satisfactory measure of central tendency only when the distribution of measures from which it is obtained tends to be uni-modal and bilaterally symmetrical, i.e., with a concentration of measures around the mid-point of the distribution and with a decrease in the measures above and below the mid-point. Obviously, the mean is not a measure of central tendency for distributions that tend to be U-shaped or J-shaped or rectangular (see Fig. 6:1) and hence have no central tendency. Nevertheless, from a mathematical point of view, the mean is a general measure and can be used to obtain a *summary statistic* for any kind of distribution.

The point to be remembered is that most statistical measures are obtained not only for the purpose of summarizing certain facts about a distribution of measures, but also for comparative purposes. The mean of one distribution is compared with that of another; but unless the two distributions are similar in form, the comparisons are likely to be misleading. To compare the mean of a J-type distribution with the mean of a normal-type distribution is an example; only for the latter would the mean be a measure of central tendency and represent the most typical score. But it would not be misleading to compare the arithmetic means of two J-shaped distributions provided it were clear that in such cases the mean is not a measure of central tendency or representative of the typical score. The possible implications of the mean are usually clarified by knowing the form of the distribution from which it is derived. It is because so many variables in the biological and social sciences yield uni-modal distributions of the bell-shaped type that the mean is often described as a measure of central tendency.

In practice, there are three methods for computing the mean:

I. A long method, with the data not organized into a frequency distribution.

II. A long method, with the data organized into a frequency distribution.
III. A short method, with the data grouped as for Method II.

Method I: The Mean from Unordered Data

When an adding machine is available, Method I is the simplest of the three methods, because it involves merely the summation of all the measures of a distribution and the division of the sum by the total number of measures. However, this method has a disadvantage in that it is often difficult, if not impossible, to ascertain the form of a distribution, especially one with a large range, unless the data are tallied into a frequency distribution with class intervals greater than the original units of measurements. In other words, it is often necessary to group the data into the class intervals of a frequency distribution in order to determine whether, in fact, the variable manifests a central tendency. Unless the original data of a group of measurements are ordered into a frequency distribution, inappropriate measures may be used to describe and summarize the results. We have already emphasized the importance, in statistical practice, of first describing the data of a variable by a frequency distribution. Methods II and III have therefore been developed for this arrangement of the data.

The data presented in Table 7:1 were obtained from 30 subjects in an experiment on *personal tempo*. Each measurement represents the metronome rate that each subject judged was the tempo he most preferred. There are thus 30 measures, one for each subject. The data in the table are arranged in three columns and are in the order in which they were originally obtained.

Table 7:1. Arithmetic Mean—Long Method with Ungrouped Data
(Data: Preferred Metronome Rates as Judged by Subjects in an Experiment on *Personal Tempo*) *

Subject No.	Score	Subject No.	Score	Subject No.	Score
1	146	11	72	21	126
2	180	12	72	22	176
3	60	13	126	23	112
4	104	14	152	24	120
5	108	15	116	25	122
6	132	16	144	26	96
7	152	17	172	27	120
8	116	18	126	28	132
9	76	19	130	29	108
10	180	20	150	30	104

$\Sigma = 3730;\ N = 30$

$$\text{Arithmetic Mean} = \frac{3730}{30} = 124.3$$

* These figures represent the tempo of the metronome beat most preferred by each of 30 subjects. The scores are the numbers of the usual metronome scale. (From Honors Research Project in Psychology at The City College of New York, by Bernard Steinzor.)

The sum of the 30 measures is 3730. The mean of this group of measures is therefore 3730 divided by 30 (the number of measures), or 124.3. Thus,

$$M = \frac{3730}{30} = 124.3$$

This method of computing the mean may be symbolized as follows:

$$M = \frac{\Sigma X}{N} \qquad [7:1]$$

Arithmetic mean (M), from ungrouped data

where ΣX is the sum of all the measures and N is the total number of measures.

Method I has the advantage of quick computation when the total number of measures is small or when an adding machine is available. This method is widely employed in machine computations. However, in using it, the investigator should set up, independently, a frequency distribution so that he will know the implications of the average he obtains. Even careful study of Table 7:1 does not reveal at all clearly whether the mean, 124.3, is *typical*, in the sense that many of the 30 cases cluster around it.

Method II: The Mean—Long Method with Data Grouped into a Frequency Distribution

The data in Table 7:1 have been rearranged in Table 7:2 into a frequency distribution with class intervals of 20 units. There are seven class intervals, and their frequencies vary from 1 to 9. Since the measures tend to be concentrated near the middle intervals of the distribution, the mean in this case is a measure of *central tendency*.

Table 7:2. Arithmetic Mean—Long Method with Grouped Data (Data from Table 7:1)

Class Intervals	f	Mid-Pt.	f(Mid-Pt.)
180 and above	2	189.5	379.0
160–179	2	169.5	339.0
140–159	5	149.5	747.5
120–139	9	129.5	1165.5
100–119	7	109.5	766.5
80– 99	1	89.5	89.5
60– 79	4	69.5	278.0
	$N = 30$		$\Sigma = 3765.0$

Range = 60 to 180; $N = 30$; $\Sigma = 3765$

$$\text{Arithmetic Mean} = \frac{3765.0}{30} = 125.5$$

The computation of the mean by Method II is illustrated in Table 7:2. The procedure may be summarized as follows:

1. Set up appropriate class intervals and make a frequency distribution of the original data.

2. Determine the mid-point values of each class interval.
3. Multiply the mid-point value of each class interval by the number of frequencies or cases in the interval. (These products appear in the last column of the table.)
4. Obtain the algebraic sum of all these products.
5. Divide this sum by the number of measures (N). The quotient thus obtained is the mean.

As indicated in the table, the sum is 3765.0. This value divided by 30 gives a mean equal to 125.5. The procedure for Method II may be symbolized as follows:

$$M = \frac{\Sigma(fX_{mp})}{N}, \text{ or } \frac{\Sigma(fX)}{N} \qquad [7:2]$$

Arithmetic mean (M), from data grouped into class intervals

where X_{mp} is the value of the mid-point of each class interval, and f is the number of frequencies.

The means obtained by the two methods have different values although they are derived from the same data. With Method I, the mean is 124.3, whereas with Method II it is slightly larger, 125.5. Such a difference is to be expected; in fact, it would be most unlikely for the two means to have exactly the same value. This is the case because Method II assumes that the mid-point value of any class interval is a representative value for all the scores in that interval. Obviously there will always be class intervals for which the mid-point values do not coincide with an exact average of the measures in the interval. But for large distributions the assumption is practical, and such discrepancies between the mean values obtained with the two methods are not likely to be serious. However, for distributions with few frequencies, sizable differences may occur if only a few broad class intervals are used.

From a mathematical point of view, the mean obtained by Method I might be described as more exact than that obtained by Method II. However, from the point of view of sampling statistics, a *sample mean* obtained by Method II is likely to be just as representative of the *population mean* as the mean obtained by Method I; it may even be more representative. It is because of this consideration in particular that discrepancies in the results obtained with the two methods are usually judged to have little or no importance.

Method III: The Mean—Short Method with Grouped Data

Once a frequency distribution of the data of a variable has been made, Method III is the easiest method of computing the mean. It is called a *short method* because the arithmetical computations are considerably simplified, which means that the method not only requires less time but also is less subject to computational errors than is Method II.

The principles underlying the short method can best be described by means of Tables 7:3, 7:4, and 7:5.

A common device is employed in Table 7:3 to simplify the computations, namely, the subtraction of a constant amount from all measures. Since the lower limit of the class interval with the smallest value is 60, we could subtract 60 from the mid-point of each class interval in the distribution. However, the procedure is simplified even more if we subtract 69.5 from such mid-

Table 7:3. Arithmetic Mean—First Step in the Short Method
(Data from Table 7:2)

Class Intervals	f	Mid-Pt.	(Mid-Pt.) − 69.5	f(Mid-Pt. − 69.5)
180 and above	2	189.5	120	240
160–179	2	169.5	100	200
140–159	5	149.5	80	400
120–139	9	129.5	60	540
100–119	7	109.5	40	280
80– 99	1	89.5	20	20
60– 79	4	69.5	0	0
	N = 30			Σ = 1680

$$\text{Arithmetic Mean} = \frac{1680}{30} + 69.5 = 56.0 + 69.5 = 125.5$$

points, since 69.5 is the *mid-point* value of the lowest interval. The differences between the mid-point values of each interval and 69.5 are given in the fourth column of the table. Since they vary only from 120 to zero, they are obviously simpler to work with than the original mid-point values. The products of these residual mid-point values and the frequencies for each interval are shown in the last column.

The sum of the products of the last column is 1680. Dividing this by 30, the number of cases, gives a mean of 56.0. This is the mean of the values in the distribution of measures after 69.5 was subtracted from the mid-point value of each class interval. The mean of the original measures will consequently be obtained by adding 69.5 to the result. Thus,

$$56.0 + 69.5 = 125.5$$

Inasmuch as the frequency distribution used in Tables 7:2 and 7:3 had identical class intervals, the means obtained by Methods II and III should be and are equal.

The procedure illustrated in Table 7:3 can be further simplified by *dividing* the residual mid-point values shown in the fourth column by a constant. Since all the mid-point values are multiples of 20 (the size of the class interval), the constant to be used is obviously 20. The quotients resulting from this division are given in the fifth column of Table 7:4. These new mid-point values are now in the simplest arithmetic terms; they range from 6 to zero. They are then multiplied by the number of frequencies or cases in the respec-

tive class intervals. The products obtained appear in the last column of Table 7:4. The sum of these products, 84, is then divided by 30, the number of cases. The result, 2.8, is an average, being the mean of the measures of the distribution with 69.5 subtracted from the mid-point value of each interval and with the reduced mid-point values divided by 20.

Table 7:4 Arithmetic Mean—Short Method Continued
(Data from Table 7:2)

Class Intervals	f	Mid-Pt.	(Mid-Pt.) − 69.5	$\frac{\text{Mid-Pt.} - 69.5}{20}$	$f\left(\frac{\text{Mid-Pt.} - 69.5}{20}\right)$
180 and above	2	189.5	120	6	12
160–179	2	169.5	100	5	10
140–159	5	149.5	80	4	20
120–139	9	129.5	60	3	27
100–119	7	109.5	40	2	14
80– 99	1	89.5	20	1	1
60– 79	4	69.5	0	0	0
	N = 30				Σ = 84

$$\text{Arithmetic Mean} = \left(\frac{84}{30}\right)20 + 69.5 = 125.5$$

The values derived from this simplified procedure now need to be compensated for by *multiplying* in what was excluded by division, and by *adding* what was subtracted. Since 20 was taken as a constant divisor, the mean of the sum of the products in the last column, 2.8, is multiplied by 20. This gives 56.0, and to it, as in Table 7:3, is added 69.5 (the amount subtracted). This gives a mean of 125.5 for the distribution, a value which coincides (as it should) with that obtained in Table 7:3.

The procedure illustrated in Table 7:4 is basically the simplest method for the data of frequency distributions. The procedure used in Table 7:5 is essentially the same as the preceding, except that the amount subtracted is taken nearer the center of the distribution rather than at one end. Although in the present example this procedure does not materially simplify the computations, the labor involved in computing is usually greatly reduced when the distributions have 15 or 20 or more class intervals.

The short method used in Table 7:5 is usually described as follows:

1. The amount subtracted is called the *guessed mean* and is symbolized by *G.M.*
2. The *divisor* is always taken as equal to the size of the class interval and is symbolized by *i.*
3. The deviations (mid-points) for each class interval are symbolized by x' after the subtraction and division of the preceding two steps.

4. As usual, frequencies are symbolized by f, and the number of cases in the distribution by N.

It makes no difference in the result whether the guessed mean (the constant which is subtracted) is taken exactly at the middle class interval of the distribution. We shall see that the procedure used in Table 7:4, with the guessed mean (69.5) taken at the lowest class interval, yields a result identical with that obtained when the guessed mean is taken at the middle interval of the distribution. The most practical procedure is to take it *near* the middle of the distribution and at a class interval with a great many frequencies. In Table 7:5

Table 7:5. Arithmetic Mean—Short Method Continued
(Data from Table 7:2)

Class Intervals	f	Mid-Pt.	(Mid-Pt.) − G.M.	$x' = \frac{\text{Mid-Pt.} - \text{G.M.}}{i}$	fx'
180 and above	2	189.5	60	3	6
160–179	2	169.5	40	2	4
140–159	5	149.5	20	1	5
120–139	9	129.5	0	0	0
100–119	7	109.5	−20	−1	−7
80– 99	1	89.5	−40	−2	−2
60– 79	4	69.5	−60	−3	−12
	$N = 30$				$\Sigma = -6$

$$\text{G.M.} = 129.5;\ i = 20;\ c = \frac{\Sigma(fx')}{N} = \frac{-6}{30} = -.2$$

$$\text{Arithmetic Mean} = \text{G.M.} + ic = 129.5 + 20(-.2) = 125.5$$

it has been taken at the fourth and middle class interval, whose mid-point value is 129.5. This, then, is *G.M.*, the guessed mean. The differences are shown in the fourth column of the table and the results of dividing them by i, the size of the class interval, which is 20, are shown in the fifth column. These quotients are the x' values for each interval. In the last column, the frequencies of each class interval have been multiplied by the x' values. The algebraic sum of the resulting products, −6, is then averaged by dividing by the number of cases. This average for the data in Table 7:5 is equal to −.2. The operation is symbolized as follows:

$$\frac{\Sigma(fx')}{N}$$

Most authors describe this average as the *correction* and symbolize it by c. Actually, of course, this is an average of the residual values of the distribution, with *G.M.* and i taken out. The corrections are in reality the values of

G.M. and *i*. However, we shall follow the usual practice and denote the average of the residual values as *c*. Thus,

$$c = \frac{\Sigma(fx')}{N}$$

The procedures used in the short method may therefore be symbolized as follows:

$$M = G.M. + i\frac{\Sigma(fx')}{N}, \text{ or } G.M. + ic \qquad [7:3]$$

Arithmetic mean (*M*), from a guessed mean, *G.M.*

Substituting the values obtained in Table 7:5 in Formula 7:3, we have

$$\begin{aligned} M &= 129.5 + 20(-.2) \\ &= 129.5 - 4.0 \\ &= 125.5 \end{aligned}$$

This value of the mean checks (as it should) with the results obtained in Tables 7:3 and 7:4.

In practice, the procedure used in Table 7:5 is simplified by omitting the third and fourth columns. This has been done in Table 7:6, which therefore illustrates the final table for the simplified procedure used in Method III. This procedure is further illustrated in Table 7:7 for another group of data which includes both negative and positive numbers. This distribution is somewhat skewed, i.e., not bilaterally symmetrical with respect to the mean, but it is uni-modal and the skewness is so slight that it does not affect the usefulness of the mean as a measure of central tendency.

Table 7:6. Arithmetic Mean—Final Table for Short Method
(Data from Table 7:2)

Class Intervals	*f*	*x'*	*fx'*
180 and above	2	3	6
160–179	2	2	4
140–159	5	1	5
120–139	9	0	0
100–119	7	−1	−7
80–99	1	−2	−2
60–79	4	−3	−12
	$N = 30$		$\Sigma = -6$

$G.M. = 129.5;\ i = 20;\ c = \frac{-6}{30} = -.2$

Arithmetic Mean $= G.M. + ic = 129.5 + 20(-.2) = 125.5$

Table 7:7. Arithmetic Mean—Short Method
(Data: Bernreuter Personality Inventory Scores for 100 College Freshmen)

Class Intervals	f	x'	fx'
100 to 129	2	5	10
70 to 99	1	4	4
40 to 69	8	3	24
10 to 39	4	2	8
−20 to 9	16	1	16
−50 to −21	24	0	0
−80 to −51	23	−1	−23
−110 to −81	14	−2	−28
−140 to −111	7	−3	−21
−170 to −141	1	−4	−4
	$N = 100$		$\Sigma = -14$

G.M. $= -35.5$; $i = 30$; $c = \frac{-14}{100} = -.14$

Arithmetic Mean $= -35.5 + 30(-.14) = -39.7$

C. THE STANDARD DEVIATION

Definition

The standard deviation is the most universally used measure of variability. It is defined as the square root of the mean of the squared deviations of all measures in a distribution. This definition, as in the case of most statistical definitions, summarizes the operations involved in computing the measure. However, it should be noted that the standard deviation is an *average* measure. It is the square root of the second-order moment.

In order to describe a distribution, it should be obvious by now that a measure of *variability* is needed in addition to the mean. Two distributions of a variable may have similar means but differ markedly in the extent of scatter of their respective measures about the means.

With respect to sampling theory and problems of analytical statistics, the standard deviation is widely employed as the *standard* measure of variability, or deviational tendency. It is generally the most reliable of all the measures of deviational tendency.* The standard deviation, when derived from distributions of the normal probability type, is widely employed in psychological measurement since it is especially relevant to the development and interpretation of *Standard scores* (see Chapter 8). In fact, the standard deviation has come to be a *relative* unit of measurement (differentiation) for most psychological scales of abilities.

* The meaning of the statistical concept of reliability is developed in some detail in Chap. 17, Section B. At this point, it is sufficient to point out that in sampling theory any measure of a distribution derived from a random sample, or series of random samples, is the more reliable, the less it differs in value from the value of that measure for the population or universe as a whole.

The three methods used for computing the standard deviation have their analogues from the computation of the mean. Thus,

I. A long method, ungrouped data.
II. A long method, grouped data.
III. A short method, grouped data.
IIIa. A short method, ungrouped data.

Method III is by far the easiest and the least subject to computational errors, as the following examples will show.

Method I: Standard Deviation from Ungrouped Data

The disadvantages of the first method for computing the standard deviation are similar to those of Method I for the mean; in addition, the computations are unnecessary, and even more laborious. The method is illustrated in Table 7:8.

Table 7:8. Standard Deviation—Long Method for Ungrouped Data
(Arithmetic Mean = 124.3; Data from Table 7:1)

Subject Number	Score (X)	Deviation (x)	Deviation Squared (x^2)	Subject Number	Score (X)	Deviation (x)	Deviation Squared (x^2)
1	146	21.7	470.89	16	144	19.7	388.09
2	180	55.7	3,102.49	17	172	47.7	2,275.29
3	60	−64.3	4,134.49	18	126	1.7	2.89
4	104	−20.3	412.09	19	130	5.7	32.49
5	108	−16.3	265.69	20	150	25.7	660.49
6	132	7.7	59.29	21	126	1.7	2.89
7	152	27.7	767.29	22	176	51.7	2,672.89
8	116	−8.3	68.89	23	112	−12.3	151.29
9	76	−48.3	2,332.89	24	120	−4.3	18.49
10	180	55.7	3,102.49	25	122	−2.3	5.29
11	72	−52.3	2,735.29	26	96	−28.3	800.89
12	72	−52.3	2,735.29	27	120	−4.3	18.49
13	126	1.7	2.89	28	132	7.7	59.29
14	152	27.7	767.29	29	108	−16.3	265.69
15	116	−8.3	68.89	30	104	−20.3	412.09
			21,026.15				7,766.55

$$\Sigma(x^2) = 21{,}026.15 + 7{,}766.55 = 28{,}792.7$$

$$\text{Standard Deviation} = \sqrt{\frac{\Sigma(x^2)}{N}} = \sqrt{\frac{28{,}792.7}{30}} = 30.98, \text{ or } 31.0$$

The mean must first be computed. The mean for the data in Table 7:1, which have been used in Table 7:8, was found to be 124.3. The deviations (x) are next obtained:

$$x = X - M_x$$

each deviation being the difference between a measure (X) and the mean (M_x) of the distribution. The direction of the differences is indicated by the use of negative signs for negative deviations. The deviation for each of the 30 cases is given in the x columns of Table 7:8.

The third step consists in squaring each deviation; these computations are facilitated by a table of squares.* The square of each deviation is given in the x^2 columns of the table. It is usually sufficient in descriptive statistics to carry the squares to two decimal places for data originally obtained in integral values. Summing the x^2 values for each measure of the distribution completes the preliminary computations necessary for calculating the standard deviation by this method. As indicated at the bottom of the table, the average of the sum of the x^2's is obtained and the square root of this average is computed. For these data, the standard deviation is found to be 31.0. This measure is therefore the square root of the mean of the squared deviations (taken from the mean). In practice, the standard deviation is symbolized by σ (Greek *sigma*).

The preceding computations may be symbolized as follows:

$$\sigma = \sqrt{\frac{\Sigma(x^2)}{N}} \qquad [7:4]$$

Standard deviation (σ), for ungrouped data

Variance

The square of σ, viz., σ^2, is called the measure of a distribution's *variance*, and is used in many problems of sampling and analytical statistics. The *variance* of a distribution is the mean of the squared deviations and is, as was earlier indicated, the second moment, μ_2, from the mean:

$$\sigma^2 = \Sigma(x^2)/N \qquad [7:4a]$$

Variance

Method II: Standard Deviation—Long Method with Grouped Data

The data in Table 7:2 are used to illustrate Method II for computing σ. This method likewise involves unnecessarily laborious computations and consequently is rarely used. Method III, the short method, gives exactly the same result as Method II (except for the possible effect of dropped decimals on the result).

The steps of this method, shown in Table 7:9, can be summarized as follows:

1. The mean is first obtained from the frequency distribution of grouped data. (The mean for the frequency distribution of Table 7:9 was computed in Table 7:2 and was found to be 125.5.)
2. The deviations, x (the differences between each mid-point and the mean), are next computed for each class interval.
3. These deviations are then squared to give x^2.

---

* See Table I, Appendix C for squares of integers from 1 to 1000.

Table 7:9. Standard Deviation—Long Method with Grouped Data
(Arithmetic Mean = 125.5; Data from Table 7:2)

Class Intervals	f	Mid-Pt.	x	x^2	$f(x^2)$
180 and above	2	189.5	64.0	4,096.0	8,192.0
160–179	2	169.5	44.0	1,936.0	3,872.0
140–159	5	149.5	24.0	576.0	2,880.0
120–139	9	129.5	4.0	16.0	144.0
100–119	7	109.5	−16.0	256.0	1,792.0
80– 99	1	89.5	−36.0	1,296.0	1,296.0
60– 79	4	69.5	−56.0	3,136.0	12,544.0
	$N = 30$				$\Sigma = 30{,}720.0$

x = Mid-Pt. − Mean = Mid-Pt. − 125.5

$$\sigma = \sqrt{\frac{\Sigma f(x^2)}{N}} = \sqrt{\frac{30{,}720.0}{30}} = \sqrt{1{,}024.0} = 32.0$$

4. The x^2's are multiplied by f, the frequencies of their respective class intervals. (These values are given in the last column of Table 7:9.)
5. The $f(x^2)$'s are added to obtain the sum of all squared deviations: $\Sigma f(x^2)$.
6. This sum is averaged; i.e., $\Sigma f(x^2)$ is divided by N to obtain the mean of the squared deviations.
7. The standard deviation, σ, is then obtained by extracting the square root of the mean of the squared deviations.

The preceding steps in the computation of σ by Method II may be symbolized as follows:

$$\sigma = \sqrt{\frac{\Sigma f(x^2)}{N}} \qquad [7:5]$$

Standard deviation (σ), for variate data grouped into class intervals

For the data in Table 7:9, σ is equal to 32.0, which varies only slightly from the value obtained for the ungrouped data in Table 7:8. As previously indicated for the mean, such a discrepancy is to be expected in the results obtained with Method I, for which the data are ungrouped, and those obtained with Method II, for which the data are grouped into a frequency distribution.

Method III: Standard Deviation—Short Method with Grouped Data

The simplest procedure for computing the standard deviation is Method III. The arithmetic is simple, especially when compared with that required for the first two methods.

The procedure is illustrated in Table 7:10, the data and computations for the mean being taken from Table 7:6. In fact, only one additional column of

Table 7:10. Standard Deviation—Short Method with Grouped Data and Charlier's Check on Computations

(Data from Table 7:6)

Class Intervals	f	x'	$f(x')$	$f(x'^2)$	(Check) $f(x'+1)^2$
180 and above	2	3	6	18	32
160–179	2	2	4	8	18
140–159	5	1	5	5	20
120–139	9	0	0	0	9
100–119	7	−1	−7	7	0
80– 99	1	−2	−2	4	1
60– 79	4	−3	−12	36	16
	$N = 30$		$\Sigma = -6$	$\Sigma = 78$	$\Sigma = 96$

$$i = 20; \quad c = \frac{-6}{30} = -.2$$

$$\sigma = i\sqrt{\frac{\Sigma f(x'^2)}{N} - c^2} = 20\sqrt{\frac{78}{30} - .04} = 20\sqrt{2.6 - .04} = 20\sqrt{2.56}$$

$$= 20(1.6) = 32.0$$

Check: $\Sigma f(x'+1)^2 = \Sigma f(x'^2) + 2\Sigma(fx') + N$

$96 = 78 + 2(-6) + 30$

$96 = 96$

computations is required by this method. After the $f(x')$ values are obtained for each class interval, these values are multiplied by the corresponding x' values to obtain $f(x'^2)$. Algebraically,

$$x'(fx') = f(x'^2)$$

and it is therefore not necessary to have a separate column of x'^2 values to be multiplied by f.

Summing the $f(x'^2)$ column gives, for the total distribution, the products of the frequencies and the squared deviations (taken from the guessed mean, with the size of the intervals, i, excluded by division). This sum, 78, is given at the bottom of the next to the last column of the table. The summation may be symbolized as follows:

$$\Sigma f(x'^2)$$

As in the case of the short method of computing the mean, it is now necessary to restore to this result what was eliminated by subtraction and division. The correction, c, used for the mean, is *squared* and *subtracted* from $\frac{\Sigma f(x'^2)}{N}$. This value, c^2, is always subtracted (regardless of whether the guessed mean is greater or less than the actual value of the mean), because the average of the squared deviations from a guessed mean not equal to the actual mean will always be too large, but never too small. This is the case because all deviations are squared and are therefore *positive* values.

After c^2 is subtracted from the average of the squared deviations, the square root of this corrected average is next obtained. This root value is then multiplied by i, the size of the class interval, to give the standard deviation for the distribution of original scores. For the data in Table 7:10, σ equals 32.0, exactly the same value as was obtained by Method II for this distribution of grouped data in Table 7:9.

The preceding operation for the computation of σ by the short method may be symbolized as follows:

$$\sigma = i\sqrt{\frac{\Sigma f(x'^2)}{N} - \left(\frac{\Sigma f(x')}{N}\right)^2} \quad \text{or} \quad i\sqrt{\frac{\Sigma f(x'^2)}{N} - c^2} \qquad [7:6]$$

Standard deviation, σ, short method from guessed mean

Charlier's Check

It is well always to have an independent method by which to check a series of arithmetic or algebraic computations. If no simple checking methods are available, the original operations may need to be repeated. However, *Charlier's check* is convenient in checking the basic operations required by the short method for computing the mean and standard deviation.

The use of this check is illustrated in the last column and at the bottom of Table 7:10. First, 1 is added to the unit deviations shown in column x'; the resulting sums for each class interval are then squared and multiplied by their respective frequencies. Thus, x' for the highest class interval in the table is 3. Adding 1 to 3 gives 4, and squaring this gives 16. The number of frequencies for this class interval is 2; hence the product for the check column is $2(16) = 32$.

As indicated at the bottom of the table, the sum of the check column, $f(x' + 1)^2$, is equal to the sum of the $f(x'^2)$ column *plus* twice the sum of the $f(x')$ column *plus* N, the total number of frequencies.

Charlier's check tests the accuracy of all the computations *within* the table which yield the *basic sums* needed for the final computation of M and σ, but it does not check the accuracy of the latter two values.

Method IIIa: Standard Deviation—Short Method with Ungrouped Data

Method III can also be used for computing the standard deviation for a set of original *ungrouped* data of a variable whose measures are positive integral values (or can readily be treated as such).

We have not called this procedure a fourth method because it is only a special case of Method III. It is widely employed in machine computations of the mean and standard deviations. Its disadvantage lies in the fact that, since a frequency distribution is unnecessary, the investigator is likely to obtain and interpret his results for a variable without concerning himself with the *form* of the distribution from which the mean and σ are obtained.

This neglect may lead to serious errors in interpreting both the mean and standard deviations in the case of distributions that depart radically from the normal, bell-shaped type. This is especially true for extremely skewed, uni-modal distributions as well as the U- and J-types.

Table 7:11. Standard Deviation—Short Method with Ungrouped Data (Data from Table 7:1)

Subject No.	X (x')	X^2 (x'^2)	Subject No.	X (x')	X^2 (x'^2)
1	146	21,316	16	144	20,736
2	180	32,400	17	172	29,584
3	60	3,600	18	126	15,876
4	104	10,816	19	130	16,900
5	108	11,664	20	150	22,500
6	132	17,424	21	126	15,876
7	152	23,104	22	176	30,976
8	116	13,456	23	112	12,544
9	76	5,776	24	120	14,400
10	180	32,400	25	122	14,884
11	72	5,184	26	96	9,216
12	72	5,184	27	120	14,400
13	126	15,876	28	132	17,424
14	152	23,104	29	108	11,664
15	116	13,456	30	104	10,816
	$\Sigma = 1{,}792$	$\Sigma = 234{,}760$		$\Sigma = 1{,}938$	$\Sigma = 257{,}796$

$$\Sigma X = 1792 + 1938 = 3730. \quad \Sigma(X^2) = 234{,}760 + 257{,}796 = 492{,}556$$

$$\text{G.M.} = 0;\ i = 1;\ c = \frac{\Sigma x'}{N} = \frac{\Sigma X}{N} = \text{Arithmetic Mean } (M) = 3730/30 = 124.3$$

$$\sigma = i\sqrt{\frac{\Sigma(x'^2)}{N} - c^2} = 1.0\sqrt{\frac{\Sigma(X^2)}{N} - M^2} = \sqrt{\frac{492{,}556}{30} - (124.3)^2}$$

$$= \sqrt{16{,}418.53 - 15{,}450.49} = \sqrt{968.04} = 31.1$$

The computation procedure, illustrated in Table 7:11 with the data of Table 7:1, may be summarized as follows:

1. Each original score (X) is taken as a deviation from a guessed mean equal to *zero*. Hence,
$$X - G.M. = X - 0 = X, \text{ and } X = x'$$

2. The sum of the original scores (X) divided by the number of cases (N) gives the mean. This is equal to c, the correction:
$$\frac{\Sigma X}{N} = \frac{\Sigma x'}{N}, \text{ since } X = x'$$
Since the original scores are integral values, i, the size of the "class intervals" for such ungrouped data, is equal to 1.0.

3. Each original score is squared to give the square of the deviations:

$$X^2 = x'^2$$

4. The squared deviations are summed to give $\Sigma(X^2)$, which equals $\Sigma(x'^2)$.
5. The correction squared (which in this case is the mean squared) is subtracted from the mean of the sum of the squared deviations.
6. The square root of the result obtained in the preceding step is the desired value of σ.

The value of σ obtained by the short method in Table 7:11 is similar to that obtained by the long method with ungrouped data in Table 7:8. The two should be identical, except for dropped decimals, inasmuch as the data in both cases were obtained from the same arrangement of measures, i.e., they were *ungrouped.*

The preceding operations may be symbolized by Formula 7:6, since Method IIIa is simply a special case of Method III. However, they are often symbolized in terms of the original measures, as follows:

$$\sigma = i\sqrt{\frac{\Sigma(X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2} \qquad [7:6a]$$

Standard deviation, σ, special case of formula 7:6

where X equals the values of the original positive integral measures, and i, the size of the "intervals," is therefore equal to 1.0.

Sheppard's Correction * for σ

When a variable has only a few broad classes, a mathematical error arises in computing the standard deviation for distributions of the normal bell-shaped type, because, as was indicated in Chapter 5, the mid-points of each class interval do not coincide with the actual means of the cases within the intervals. When many class intervals are used, the difference is negligible; however, when there are less than ten or twelve, it is sometimes worth while to correct for the constant error that arises. Such a correction was developed by Sheppard. It is easy to apply, and hence its use may well be considered when distributions are of the normal bell-shaped type but with few class intervals.

We shall illustrate its use with the personal tempo data in Table 7:10, even though this distribution is not particularly bell-shaped. Only 7 class intervals were used for the distribution of these scores because of the few cases (only 30), rather than because of any limitations inherent in the measures themselves. Actually the range of scores was from 60 to 180, and many more class intervals could have been used if there had been enough frequencies to warrant smaller intervals.

* W. F. Sheppard, "The Calculation of the Moments of a Frequency Distribution," *Biometrika,* 5:450–459, 1907.

The correction itself is a constant, equal to 1/12, or .0833. This constant is *subtracted* from the average of the squared unit-deviations from the actual rather than the guessed mean, as follows:

$$\sigma_{corrected} = i\sqrt{\sigma_{u\,d}^2 - .0833} \qquad [7:7]$$

Standard deviation, σ, with Sheppard's correction for broad classes

For the data in Table 7:10:

$$\sigma = 20\sqrt{2.56} = 20(1.6) = 32.0$$
$$\sigma_{cor.} = 20\sqrt{2.56 - .0833} = 20(1.57) = 31.4$$

Thus the corrected standard deviation for the distribution of these personal tempo scores is 31.4 instead of 32.0. Mathematically this difference is noticeable, but psychologically it is not very important. That is to say, a difference of six-tenths of a unit on the metronome scale on which these personal tempo scores were based makes little or no difference in the psychological interpretation of the result. Furthermore, in sampling statistics the error arising from broad classes is often small compared with errors of sampling and measurement. Sheppard's correction is an unnecessary, mathematical over-refinement in such situations.

It will be observed that this correction of .0833 is always *subtracted* from the average of the squared unit-deviations from the mean because the constant error arising from broad classes increases rather than decreases the size of the deviations: In a normal, bell-shaped distribution, there are more cases between the mid-point of a class interval and the limit of the interval that is nearer the mean of the distribution, than between the mid-point and the other limit of the interval. The correction is not necessary for the arithmetic mean because the errors for intervals above the mean will tend to cancel out the errors for intervals below the mean.

D. THE AVERAGE DEVIATION

The *average deviation*, *A.D.*, sometimes referred to as the *mean deviation*, is another statistical measure that is used to summarize the deviational tendency of the measures of a variable. However, it is not used as commonly in statistical practice as is the standard deviation. For one thing, the average deviation does not exist from an algebraic point of view, since the algebraic sum of deviations from the mean is equal to zero. From the point of view of sampling, furthermore, the standard deviation is a more reliable measure of deviational tendency than is the average deviation. Nevertheless, because of its occasional use the computation of the average deviation will be briefly described.

Definition

The average deviation is the arithmetic mean of the differences between the measures of a distribution and the mean of that distribution. These differ-

ences (or deviations) are summed without regard to the direction of the differences from the mean of the distribution. If X symbolizes any measure of a distribution, and x *symbolizes a deviation,* the average deviation is equal to the following:

$$A.D. = \frac{S(X - M)}{N}, \text{ or } \frac{S(x)}{N} \qquad [7:8]$$

Average deviation, ungrouped data (*A.D.*)

where S symbolizes the *operation* of summing the deviations obtained, the direction of the differences being neglected.

Two methods for computing the average deviation will be described: Method I, the average deviation from ungrouped data; and Method II, the average deviation from grouped data.

Method I: Average Deviation—Ungrouped Data

Method I is illustrated in Table 7:12, in which the data from Table 7:1 have been used. We saw in Table 7:1 that the mean for the 30 personal tempo scores was 124.3. As indicated in the third column of Table 7:12, the difference (x) between each score and the mean is obtained without regard to sign. The sum of these differences is 718.0. Therefore,

$$A.D. = \frac{S(x)}{N} = \frac{718.0}{30} = 23.9$$

Table 7:12. Average Deviation—Ungrouped Data
(Arithmetic Mean = 124.3; Data from Table 7:1)

Subject No.	X (Score)	x^* $(X - M)$	Subject No.	X (Score)	x^* $(X - M)$
1	146	21.7	16	144	19.7
2	180	55.7	17	172	47.7
3	60	64.3	18	126	1.7
4	104	20.3	19	130	5.7
5	108	16.3	20	150	25.7
6	132	7.7	21	126	1.7
7	152	27.7	22	176	51.7
8	116	8.3	23	112	12.3
9	76	48.3	24	120	4.3
10	180	55.7	25	122	2.3
11	72	52.3	26	96	28.3
12	72	52.3	27	120	4.3
13	126	1.7	28	132	7.7
14	152	27.7	29	108	16.3
15	116	8.3	30	104	20.3
		$S = 468.3$			$S = 249.7$

$$A.D. = \frac{S(x)}{N} = \frac{468.3 + 249.7}{30} = \frac{718.0}{30} = 23.93$$

* Ordinarily x is always summed with regard to sign. However, as indicated in the text, the sign is disregarded in obtaining the average deviation. The operation of summing is therefore symbolized by S rather than the Greek symbol Σ.

Method II: Average Deviation — Grouped Data

The data in Table 7:3 used to compute the arithmetic mean are utilized in computing the average deviation by Method II, shown in Table 7:13. Arithmetically, the procedure is the same as that followed in Table 7:12, except that the original scores are grouped into 7 class intervals and the mid-point value of each interval is taken as representative of the scores in each interval. For these grouped data, the arithmetic mean was found to be 125.5. This value is therefore subtracted from the mid-point value of each class interval. These differences (deviations) are then multiplied by the frequencies for each interval, giving the results shown in the last column of Table 7:13. The sum of these products, 744.0, divided by N, the number of cases in the distribution, gives a mean deviation equal to 24.8.

Table 7:13. Average Deviation—Grouped Data
(Arithmetic Mean = 125.5; Data from Table 7:2)

Class Intervals	f	Mid-Pt.	x (Mid-Pt.) − M	f(x)
180 and above	2	189.5	64.0	128.0
160–179	2	169.5	44.0	88.0
140–159	5	149.5	24.0	120.0
120–139	9	129.5	4.0	36.0
100–119	7	109.5	16.0	112.0
80– 99	1	89.5	36.0	36.0
60– 79	4	69.5	56.0	224.0
	N = 30			S = 744.0

$$\text{Average Deviation} = \frac{Sf(x)}{N} = \frac{744.0}{30} = 24.8$$

Formula 7:8 (the $A.D.$) thus becomes as follows for the data of a variate grouped into class intervals:

$$A.D. = \frac{Sf(x)}{N} \qquad [7:9]$$

A.D. for variate data grouped into class intervals

As in computing the arithmetic mean and the standard deviation from ungrouped and grouped data, there is a slight difference in the values of the mean deviations shown in Tables 7:12 and 7:13. As previously indicated, this is to be expected because the mid-points of all the class intervals do not ordinarily give the exact average of the original scores within the intervals.

It will be observed that the $A.D.$ is always less in value than the standard deviation. For the data in Table 7:13, $A.D.$ was 24.8, whereas in Table 7:9 σ for the same data was found to be 32.0.

E. THE COEFFICIENT OF RELATIVE VARIATION

Karl Pearson developed a measure for comparing the *relative* variability of two or more variates whose means are dissimilar. The measure is known as the Coefficient of Relative Variation, symbolized by V, and is stated as follows:

$$V = \frac{100\sigma}{M} \qquad [7:10]$$

Pearson's Coefficient of Relative Variation

where the numerator is the standard deviation of the distribution and the denominator is the arithmetic mean. V measures, in percentages, the ratio of the standard deviation of a distribution to its mean. Thus, if V equals 20%, this signifies that the standard deviation is 20% as large as the mean of the distribution from which it is derived.

If two variates have unequal standard deviations but their means are the same, the Coefficient of Relative Variation is unnecessary, because the deviations of both variates are relative to the same point of reference, i.e., the same mean. On the other hand, the standard deviation of two distributions may be the same, but their means dissimilar. In such cases, the variability of the distribution with the smaller mean is relatively greater than the variability of the other distribution.

Pearson developed this coefficient primarily for physical measures which have a true zero point on the scale of measures, and it is quite satisfactory for such variates. In the case of psychological measurements, however, test variables do not have a true zero point and consequently the Coefficient of Relative Variation can be logically used only when two or more distributions for the same variate are compared. Thus, it is sound to compare the relative variation of a group of boys with that of a group of girls for measures derived from the same test. But it is not sound to compare the relative variation of the results made by a group on one test with the results made by the same or a different group on a different test. This is true because the means of two different psychological variates are not fixed, but are rather a function of the characteristics of the construction of the test, as well as of the ability of the individuals taking it. This difference can be illustrated by the following example:

A group of 100 boys is given two different tests, A and B. The results are summarized as follows:

	N	M	σ	V
Test A	100	50	10	100(10)/50 = 20.0%
Test B	100	60	10	100(10)/60 = 16.7%

If these results could be taken at their face value, it could be concluded that the group was relatively more variable on Test A (V = 20.0%) than on Test B (V = 16.7%). These results, however, cannot be interpreted lit-

erally, because the variates are different. If, for example, there had been 20 additional easy items on Test A which all the boys in the group would have answered correctly, the results would have been as follows:

	N	M	σ	V
Test A	100	70	10	100(10)/70 = 14.3%
Test B	100	60	10	100(10)/60 = 16.7%

V_A is now seen to be 14.3% instead of 20%. But actually the variability of the group has not changed at all. Twenty points have been added to each individual score because of the inclusion of the 20 easy items, and the mean is consequently 70 instead of 50. This increase in the mean therefore reduces the size of V as indicated above.

The Coefficient of Relative Variation is thus seen to be a capricious index of relative variability for psychological variables unless two or more groups are compared with respect to the same variate. In such comparisons, the scale of measures is the same and consequently the result is relatively unaffected by the particular character of the test itself. Consider, for example, the results of Test A given to a group of 100 boys and 100 girls:

	N	M	σ	V
Boys	100	50	10	100(10)/50 = 20.0%
Girls	100	40	8	100(8)/40 = 20.0%

Thus the "absolute" variability (in terms of σ) of the group of boys is 25% greater than that of the girls, since $100(10 - 8)/8 = 25\%$. The relative variability of the two groups, however, is the same, 20% in both cases.

EXERCISES

1. Under what circumstances are the mean, median, and mode of a distribution always equal in value?
2. What is the essential difference between the mean and the median of a distribution?
3. Under what circumstances is the mean a measure of central tendency of a distribution?
4. What is the essential difference between the standard deviation and the quartile deviation?
5. What is the essential difference between the standard deviation and the average deviation?
6. Under what circumstances can and cannot Pearson's coefficient, V, be used to compare the relative variations of two or more distributions?
7. Using the data in Table 5:14:
   a. Compute the means and standard deviation of the *ages* of the freshmen and their best friends by Method I for the first 25 subjects (long method, data not grouped in the frequency distribution).
   b. Compute the means and standard deviations of the *grade averages* of the freshmen and their best friends by means of Method II (long method, data grouped).

c. Compute the means and standard deviations of the *intelligence test* scores of the freshmen and their best friends by Method III (short method with data grouped).

8. Compare and interpret the results for college freshmen and their best friends obtained in the preceding exercise.
9. Compute the mean and standard deviation for both variables in Table 6:7.
10. In the preceding exercise, are the means in both cases as adequate as the medians for the purpose of summarizing the "central tendency" of each distribution?
11. Apply Sheppard's correction to the standard deviations for the age variable for the first 25 cases of both college freshmen and their best friends (Table 5:14), and compare these corrected values with those obtained in Exercise 7a.
12. Compare the relative variation of (a) the grade scores, (b) the intelligence test scores, and (c) the ages of the college freshmen and their best friends (Table 5:14).
13. For the data of Exercise 12, can the relative variability of the college freshmen's average grades be compared with that of their intelligence test scores? Why?

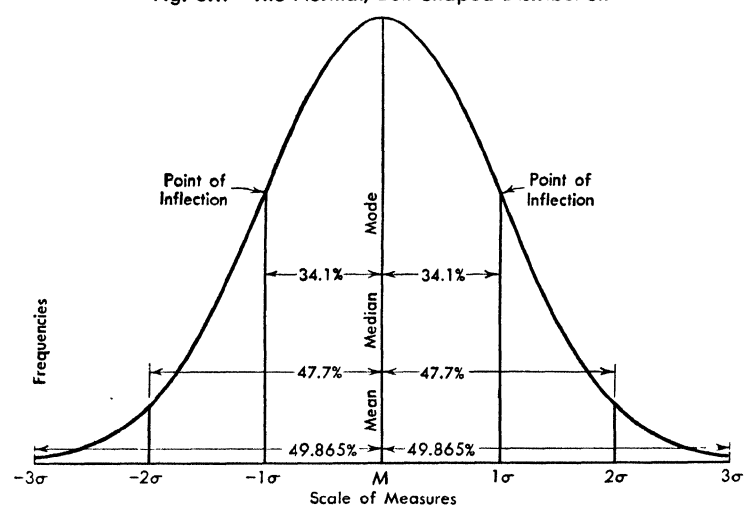
CHAPTER 8

Comparative Implications of the Normal, Bell-Shaped Curve

A. IMPLICATIONS OF M, AND σ FOR NORMAL, BELL-SHAPED DISTRIBUTIONS

The mean and standard deviation have important theoretical as well as practical implications when they are derived from a distribution that tends to be of the normal, bell-shaped type. Although knowledge of their mathe-

Fig. 8:1. The Normal, Bell-Shaped Distribution



matical properties, under such circumstances, is indispensable for the problems of sampling and analytical statistics, awareness of some of the basic implications of M and σ is also relevant so that the meaning of these measures may be broadened for purely descriptive problems. We shall therefore describe some of the properties of the normal, bell-shaped distribution, the general form of which is shown in Fig. 8:1. By describing the normal distribution with respect to the first and second moments, we can at the same time ascertain some of the "normal" implications of the mean and standard deviation.

The Mean as Point of Reference

The *mean* is taken as the fundamental point of reference. All deviations are computed from it, and the algebraic sum of these deviations equals zero.

The Mean as a Fulcrum

The mean is at a point on the scale that cuts in half both the total *weights* of the measures and the total *number* of frequencies. Whereas the median divides a distribution of *frequencies* into two equal halves, regardless of the values or weights of the measures, the mean does this and more. It is analogous to a *fulcrum*, for if a normal distribution were balanced on a knife-edge at a point on the abscissa scale corresponding to the value of the mean, the distribution would be in perfect equilibrium.

The Median and Mean

The value of the median coincides with the value of the mean when a distribution is bilaterally symmetrical, as is the case for the normal distribution.

Uni-Modality and the Mode

The normal distribution has one modal point; in other words, it is a unimodal distribution with the greatest number of frequencies at the mean. Hence, the value of the mode * coincides with the value of the mean and median.

Bilateral Symmetry

The normal distribution is bilaterally symmetrical with respect to the mean. Not only do the sums of all the positive and negative deviations equal each other, and therefore summate to zero, but the algebraic sum of any part of the deviations is equal to zero, regardless of the deviate distances above and below the mean, so long as the two distances are taken equally. In other words, the frequencies above and below the mean decrease at a uniform rate with each successive interval (of whatever size) above and below the mean. The slope of the curve is always the same at equal distances above and below the mean.

Points of Inflection and σ

The standard deviation is the standard measure of variability for the normal distribution. As the second moment with respect to the mean, σ measures a

* The mode is defined as the value of the most frequently occurring measure in a distribution, or, better, as the mid-point value of the class interval with the greatest number of frequencies. It is sometimes used as a third type of measure of central tendency for distributions of the normal, bell-shaped type. However, it is also useful in describing the modal intervals of U-shaped (two modes) and J-shaped (one major mode and one minor mode) types of distributions.

range above and below the mean that is exactly equal to that range taken with respect to the *points of inflection* * on each side of the peak of the curve. That is, if lines perpendicular to the abscissa are dropped from the point of inflection on each side of the curve, they coincide with perpendicular lines drawn to the curve from the abscissa at points exactly one standard deviation above and below the mean.

Thus, the range of the standard deviation above and below the mean marks the points on the curve at which the rate of decrease in frequencies changes. Between the mean and one standard deviation distance ($M \pm 1\sigma$), the rate of decrease accelerates, whereas beyond $M \pm 1\sigma$ this rate decelerates. In other words, the slope of the curve is convex between $M \pm 1\sigma$.

Asymptotic Character of the Normal Curve

When the normal curve is considered as representing a distribution of frequencies, the number of frequencies is necessarily *infinite*, because otherwise the surface of the curve would not be perfectly smooth and continuous. Now, as the distance of deviations from the mean is increased, the proportion of frequencies decreases. However, no matter how great a distance from the mean is taken, the frequencies never equal zero. In other words, the tails of the normal curve never reach the base line (abscissa) but are asymptotic with respect to the x-axis.

The Practical Limits Equal $M \pm 3.0\sigma$

For practical purposes, on the other hand, the limits of the frequencies of empirical distributions that are of the normal type rarely exceed a distance greater than 3.0 standard deviation units from the mean. As indicated in Table 8:1, shortly to be discussed, 49.865% of the area (or frequencies) of the normal distribution lies between the mean and 3.0σ. Therefore, 2(49.865) or 99.73% of the frequencies lie within the limits of $M \pm 3.0\sigma$.

The proportion of the area between the mean and 5.0σ is seen in Table 8:1 to be 49.99997133. The proportion of frequencies for an infinite distance beyond 5.0σ is equal to 50.0 − 49.99997133, which is only 0.00002867%. When both tails of the distribution are considered, only 2(0.00002867)% of the frequencies lie beyond the limits of $M \pm 5.0\sigma$. This amounts to less than 6 hundred-thousandths of 1% (0.00006 of 1%).

σ as the Standard Measure of Variability

A deviation, as we have seen, is symbolized by x, x being equal to $X - M_x$. Inasmuch as deviations for the normal distributions are measured in terms of σ, any such deviate distance is symbolized as $\frac{x}{\sigma}$, or $\frac{X - M_x}{\sigma_x}$. Hence by

* The point at which a concave downward portion of a curve meets a concave upward portion. Cf. Fig. 8:1.

this ratio it is possible to denote any measure of a distribution in terms of a standard unit of differentiation or variability. Thus, any measure which is at a point on the scale one standard deviation above the mean has a value in standard deviation units of 1.0. For if $X = M_x + 1\sigma_x$,

then
$$\frac{X - M_x}{\sigma_x} = \frac{(M_x + 1\sigma_x) - M_x}{\sigma_x} = \frac{\sigma_x}{\sigma_x} = 1.0$$

Measures as z Scores

Deviations in units of the standard deviation are symbolized by z, and are called z scores. Thus,

$$z_x = \frac{X - M_x}{\sigma_x}, \quad \text{or} \quad z = \frac{X - M}{\sigma} \qquad [8:1]$$

z score

When the subscripts are omitted, it is understood that the score (X), the mean (M), and the standard deviation (σ) are all derived from the same distribution of a variable. We said earlier that particular measures of a distribution are symbolized by numerical subscripts; similarly, z scores for particular measures are symbolized by the same subscripts. Thus, the value of a score, X_1, converted to its deviate distance in σ units from the mean, is symbolized by z_{x_1} and is equal to

$$z_{x_1} = \frac{X_1 - M_x}{\sigma_x}$$

Measures below the mean have *negative* z score values and are always so labeled. Measures above the mean are positive and their z score values are usually written without the plus sign.

z Scores Signify Relative Position in a Series *

The concept of the *positional meaning* of a measure is of extreme significance to the concept of measurement itself in psychology and related fields. This is the case because (1) original scores or measures usually have little or no meaning until considered in relation to the distributions from which they are derived, and (2) measures of ability, attitudes, interests, etc., are not additive as are units of the c.g.s. system in physical measurements; at best, psychological measures denote the *position* of an individual in a distribution. The functional implications of a given position in a scale are a problem for empirical determination.

z scores, or their derivatives, are universally used to express the *relative position* of an original measure in the series of measures or distribution from which it is derived. Consequently, z scores and their derivatives are valuable for comparing the *relative position* of measures of different variables.

* Cf. J. G. Peatman, "On the Meaning of a Test Score in Psychological Measurement," *American Journal of Orthopsychiatry*, 9:23–47, 1939; especially pp. 29 ff.

If a person receives a score of 90 on Test x and a score of 55 on Test y, these values of 90 and 55 are not directly comparable, because they are obtained from two different variables whose units of measurement are not the same. However, they can be made comparable in respect to their *relative position* in each distribution if they are converted to z score values and thereby expressed in terms of their standard deviate distance from their respective means. If the mean of the x variable equals 70 and its standard deviation is 20, then

$$z_{x_1} = \frac{90 - 70}{20} = 1.0$$

And if the mean of the y variable equals 50 and its standard deviation is 5.0, z_{y_1} is also equal to 1.0, since $\frac{55 - 50}{5} = 1.0$. In both cases, therefore, the measures are in scale positions that are one standard deviation above their respective means, despite the fact that the original value of one measure is 90 and that of the other is 55.

In other words, if measures of two or more different distributions are located at the same abscissa points on normal distributions, they all have the same z score values, regardless of the magnitude of their original values.

Centile Implications of Standard Measures

The relative scale position of measures can be expressed, as just indicated, in terms of their standard deviate distance above or below the mean. Their *position* can also be interpreted in terms of centile values. This is done by differentiating the normal distribution into successive deviate distances from the mean and determining the *proportion* or *percentage* of frequencies found in the interval between the mean and any given distance. The unit of differentiation used for this purpose is again the standard deviation, and the deviate distances are therefore $\frac{x}{\sigma}$ values, i.e., z scores.

Table 8:1 presents the differentiation of the total area (or frequencies) of a normal distribution into fractional parts for deviate distances above or below the mean, ranging from a distance of z = zero (the mean) to a distance of $z = 5.0$. The proportion of the area (or frequencies) for any deviate distance from the mean is given in *percentages* in the body of the table.

If an original score has a z value of 1.3, it is above the mean (since it is positive); and it is at a point on the scale of measures such that 40.32% of the area (or frequencies) lies between this point and the mean. This is illustrated in Fig. 8:2.

If 40.32% of the frequencies are between the mean and a score that has a z value of 1.3, then 90.32% of the frequencies of the distribution will be below this point, and 9.68% will be above it. In other words, an original score whose

Table 8:1. Fractional Parts of the Total Area Under the Normal Probability Curve, Corresponding to Distances on the Base Line Between the Mean and Successive Points Laid Off from the Mean in Units of Standard Deviation

Example: Between the mean and a point 1.3σ, *i.e.*, $\left(\frac{x}{\sigma} = 1.3\right)$, lies 40.32% of the entire area under the curve, or 40.32% of the frequencies.

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	00.00	00.40	00.80	01.20	01.60	01.99	02.39	02.79	03.19	03.59
0.1	03.98	04.38	04.78	05.17	05.57	05.96	06.36	06.75	07.14	07.53
0.2	07.93	08.32	08.71	09.10	09.48	09.87	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	32.89	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	43.82	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	49.865									
3.5	49.97674									
4.0	49.99683									
4.5	49.99966									
5.0	49.99997133									

deviate distance in a distribution is 1.3 standard deviation units above the mean is at a point that cuts the distribution into two parts such that the lower part includes slightly more than 90% of the frequencies, and the upper part includes the remainder.

Fig. 8:2. The Location on a Normal, Bell-Shaped Distribution of a Measure 1.3 Standard Deviation Units above the Mean

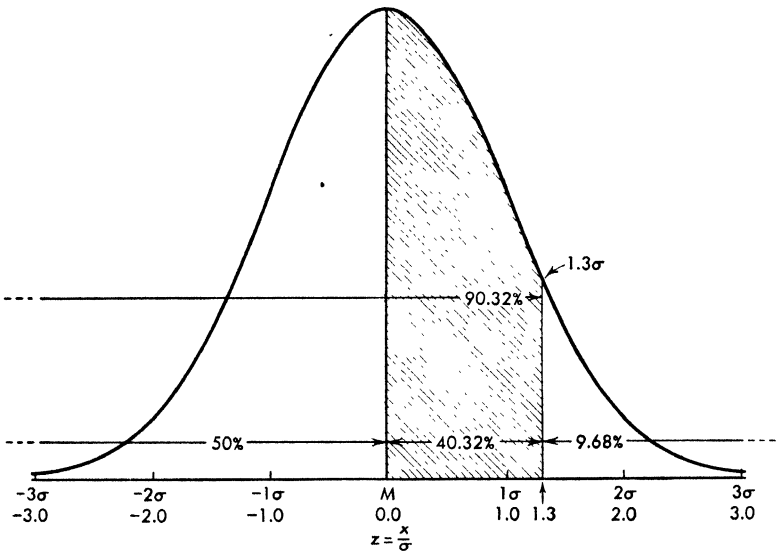
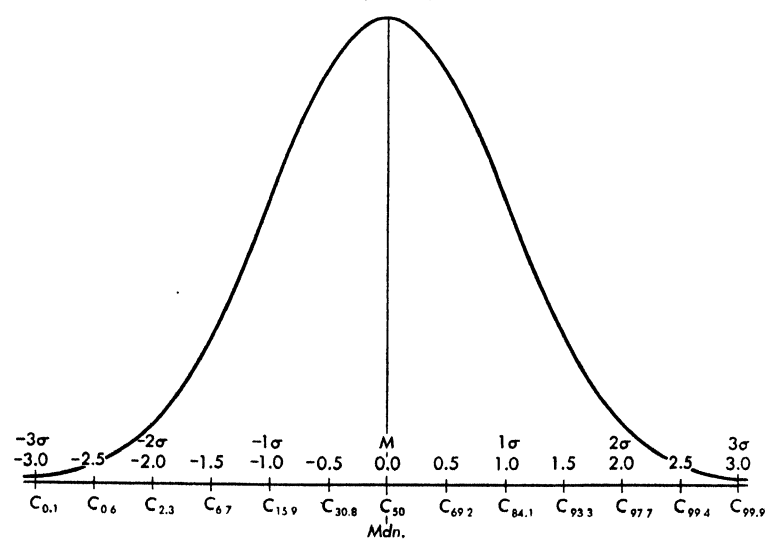


Fig. 8:3. Centile Implications of Standard Deviation Distances and z Score Units of the Normal, Bell-Shaped Distribution



Corresponding Centile Point Values for σ and z Score Values

We saw in Chapter 6 that a measure that lies in this position on the centile scale is in the 91st centile interval. Therefore, an original measure with a z score value of 1.3 lies in the 91st centile interval of the normal distribution.

The centile interval locations of z score values for a normal distribution are summarized in Table 8:2, and the relationship between centile point values and z score differentiations for such a distribution is illustrated in Fig. 8:3. Thus, as indicated in the table, all z scores equal to or greater than 2.33 are in the interval of C_{100}; all z scores of -2.33 or less are in the first centile interval, C_1. The mean ($z = 0.0$) is in the 51st centile interval (as is the median), in accordance with the principle that measures whose value corre-

Table 8:2. The Centile Intervals of Original Measures Converted to z Score Values

(Assuming Normal Variability)

z Scores	Centile Interval	z Scores	Centile Interval
5.0		0.0	C_{51}
4.0		−0.1	C_{47}
3.0	C_{100}	−0.2	C_{43}
2.5		−0.3	C_{39}
2.33		−0.4	C_{35}
2.3	C_{99}	−0.5	C_{31}
2.2	C_{99}	−0.6	C_{28}
2.1	C_{99}	−0.7	C_{25}
2.0	C_{98}	−0.8	C_{22}
1.9	C_{98}	−0.9	C_{19}
1.8	C_{97}	−1.0	C_{16}
1.7	C_{96}	−1.1	C_{14}
1.6	C_{95}	−1.2	C_{12}
1.5	C_{94}	−1.3	C_{10}
1.4	C_{92}	−1.4	C_9
1.3	C_{91}	−1.5	C_7
1.2	C_{89}	−1.6	C_6
1.1	C_{87}	−1.7	C_5
1.0	C_{85}	−1.8	C_4
0.9	C_{82}	−1.9	C_3
0.8	C_{79}	−2.0	C_3
0.7	C_{76}	−2.1	C_2
0.6	C_{73}	−2.2	C_2
0.5	C_{70}	−2.3	C_2
0.4	C_{66}		
0.3	C_{62}	−2.33	
0.2	C_{58}	−2.5	
0.1	C_{54}	−3.0	C_1
		−4.0	
		−5.0	

sponds with the lower limit of an interval lie within that interval. Actually, of course, the point value of the mean lies exactly at the mid-point of the scale.

Commonly used points of reference for z score values between -3.0 and 3.0 are noted on the normal distribution shown in Fig. 8:3. Thus, measures that are one standard deviation above the mean are in the 85th centile interval, because the point centile value at $z = 1.0$ is 84.1. It follows that 84% of the frequencies of a normal distribution are below $z = 1.0$. Therefore, a measure whose position on the scale is one standard deviation above the mean exceeds in value 84% of the measures of the distribution.

A measure that is one standard deviation *below* the mean lies in the 16th centile interval because the centile point value of $z = -1.0$ is 15.9. Therefore, nearly 16% of the measures of a normal distribution are lower in value than one whose z score equivalent is -1.0. The range of $M \pm 1.0\sigma$ includes approximately the middle 68% of the frequencies. Although this follows from the preceding discussion, it can perhaps be computed more readily from the data in Table 8:1 than observed from Fig. 8:3. The percentage of the frequencies between M and 1.0σ is 34.13 in Table 8:1; twice this amount therefore gives the percentage of frequencies lying between $M \pm 1.0\sigma$.

At 2.0σ, the centile point value is given in Fig. 8:3 as 97.7. A z score of 2.0 therefore lies in the 98th centile interval, and about 2% of all the measures of a normal distribution are greater than one whose value is 2.0 standard deviation units above the mean. At -2.0σ, on the other hand, the centile point value is 2.3, and therefore a z score of -2.0 lies in the 3rd centile interval. Approximately 98% of all the measures of the distribution are greater than one whose value is -2.0 standard deviation units below the mean. The range of $M \pm 2.0\sigma$ includes approximately the middle 95% of the measures of a normal distribution. ($M + 2.0\sigma$, according to Table 8:1, equals 47.72% and twice this figure for $M \pm 2.0\sigma$ is 95.44%.)

Measures whose z values are 0.5 and -0.5 lie in the 70th and 31st centile intervals respectively. About 30% of the measures of a normal distribution are thus greater in value than a measure one-half a standard deviation above the mean, and about 30% are less in value than a measure one-half a standard deviation below the mean. The range of $M \pm 0.5\sigma$ thus includes approximately the middle 40% of the frequencies. ($M + 0.5\sigma$, according to Table 8:1, equals 19.15%; twice this figure for $M \pm 0.5\sigma$ is 38.30%.)

Measures whose z values are 1.5 and -1.5 lie in the 94th and 7th centile intervals respectively. About 93% of the measures of a normal distribution are thus less in value than a measure $1\frac{1}{2}$ standard deviations above the mean, and about 93% are greater in value than a measure the same distance below the mean. The range of $M \pm 1.5\sigma$ includes approximately the middle 85% of the frequencies. ($M + 1.5\sigma$, according to Table 8:1, equals 43.32%; twice this figure for $M \pm 1.5\sigma$ is 86.64%.)

Summary of Commonly Used Measures of Dispersion About the Mean

The percentages of the total frequencies of a normal distribution that are included within the limits of various z score values are summarized in Table 8:3. These percentages, which are taken from Table 8:1, are commonly used reference values for the normal distribution. The student will be wise to memorize Table 8:3 because a ready knowledge of these dispersions facilitates the interpretation of the mean and standard deviation of distributions that are of the normal type.

Table 8:3. The Dispersion of Frequencies About the Mean of a Normal Distribution

Range in Terms of M and σ	Per Cent of Frequencies Included Within the Range (Rounded to Nearest Unit Value)
$M +$ or $- 0.5\sigma$	19%
$M +$ and $- 0.5\sigma$	38%
$M +$ or $- 0.6745\sigma$	25%
$M +$ and $- 0.6745\sigma$	50%
$M +$ or $- 1.0\sigma$	34%
$M +$ and $- 1.0\sigma$	68%
$M +$ or $- 1.5\sigma$	43%
$M +$ and $- 1.5\sigma$	87%
$M +$ or $- 2.0\sigma$	48%
$M +$ and $- 2.0\sigma$	95%
$M +$ or $- 2.5\sigma$	49%
$M +$ and $- 2.5\sigma$	99%
$M +$ or $- 3.0\sigma$	$50^{-}\%$
$M +$ and $- 3.0\sigma$	$100^{-}\%$

The range of $M \pm 0.6745\sigma$ has been included in this table because it defines the range of the middle 50% of the frequencies of a normal distribution and is the basis for the *probable error* (*P.E.*) in sampling distributions (cf. Chapter 13, Section E). The *P.E.* of a measure whose sampling distribution is of the normal bell-shaped type is always 0.6745σ.

The Normal Probability Curve

The curve of normal probability is the normal, bell-shaped distribution shown in Fig. 8:1. It is often referred to as the normal curve of error, because random errors of measurement commonly yield distributions of this type.

At this point it is sufficient to note that the mean and standard deviation, as the first and second moments, are the standard measures for probability and error distributions of the normal type. Extensive use will be made later of the properties and implications of normal probability in the development of Tests of Significance in the problems of analytical and sampling statistics.

The Formula for the Normal Curve

The equation for the normal distribution, differentiated in terms of standard deviation units, is as follows:

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}} \qquad [8:2]$$

Normal probability function in terms of σ

This is the normal probability function. In plotting a normal distribution for N frequencies, it is not necessary to work directly from this function; rather, one can utilize tables of ordinate values (y) for different values of x which have been developed for distributions whose total area is taken as unity (cf. Table 1, Appendix B).

Relationship Between Various Measures of Variability in a Normal Distribution

We indicated earlier that σ is greater than $A.D.$ for any distribution. Both these measures of variability are larger than the quartile deviation and the tercile deviation, but smaller than the D range. When a distribution can be assumed to be normal, the relation between these various measures is as follows:

$$\begin{aligned} \sigma &= 2.317T.D. = 1.483Q &&= 1.253A.D. &&= .390D \\ A.D. &= 1.849T.D. = 1.183Q &&= .798\sigma &&= .311D \\ Q &= 1.563T.D. = .845A.D. &&= .6745\sigma &&= .263D \end{aligned}$$

B. THE USE OF z SCORES AND STANDARD SCORES FOR COMPARATIVE PURPOSES

We have seen that the standard deviation has come to be used as the standard measure of variability of the normal, bell-shaped type of distribution. Original scores of a variable can readily be expressed in units of σ by conversion to z scores, where

$$z_x = \frac{X - M_x}{\sigma_x}$$

And, as indicated in Fig. 8:3, the positional implications of z scores can be interpreted in terms of centiles. However, z scores are somewhat inconvenient to use, particularly with machine methods, because of the presence of negative numbers. All original scores *below* the mean will have negative z score values. In order to obviate the use of negative numbers, a variety of con-

version scales have been developed, the most satisfactory being the Standard score scale.*

Standard Scores (S)

Original measures or scores of a variable are readily converted into Standard scores by adding 5.0 to each of the z score values of the originals. Thus,

$$S = 5.0 + z_x$$

or

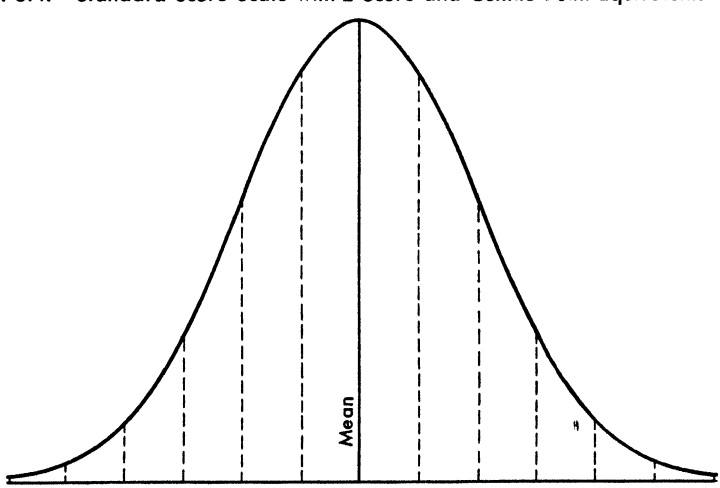
$$S = 5.0 + \frac{X - M_x}{\sigma_x} \qquad [8:3]$$

Standard score

where X is the original score and M and σ are the mean and standard deviation of the variable or distribution of which the original score is a member.

The relation of Standard scores to z scores and centile ranks is shown in Fig. 8:4. It is to be emphasized that the normal, bell-shaped distribution is the basis for interpreting a Standard score scale as a yardstick for the differentiation of the measures of a variable. Under such circumstances Standard scores are a convenient and appropriate device for (1) the development of tables of norms, as in Table 8:4, and (2) the comparison of one individual's scores on different tests or variables, as in the profile charts shown in Figs. 8:6 and 8:7.

Fig. 8:4. Standard Score Scale with z Score and Centile Point Equivalents



S	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0
z	-3.0		-2.0		-1.0		0		1.0		2.0		3.0
C	0.1		2.3		15.9		50.0		84.1		97.7		99.9

* Although z scores themselves are sometimes referred to as Standard scores, we shall limit the use of the latter term to the conversion scale in which the mean is taken as equal to 5.0 and, as for z scores, the standard deviation remains 1.0. Cf. W. V. Bingham, *Aptitudes and Aptitude Testing*, Harper, New York, 1937, chap. 19.

For practically all distributions of test scores, a Standard score range of from 2.0 to 8.0 is adequate. In fact, for many distributions, like that in Table 8:4, the actual range of scores is likely to be less than this range.

Standard scores are usually written to one decimal place. This means that for a table of norms for a psychological test, there can be as many as 60 differentiations within the limits of 2.0 and 8.0. A scale with 60 intervals is more than adequate for any test; in fact, most tests are not sufficiently reliable to warrant the use of so many intervals. In most cases, 10 or 20 intervals are adequate.

Standard Score Norms

The development of test norms in terms of Standard scores and their centile equivalents is well illustrated by Table 8:4 for the Bennett Test of

Table 8:4. Standard Score Norms, for the Bennett Mechanical Comprehension Test, Form AA, Candidates for Policeman and Fireman Positions

Method I

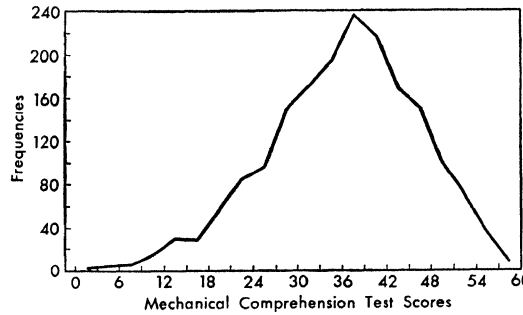
Original Score	Standard Score	Centile Interval
56	7.0	98
51	6.5	94
46	6.0	85
41	5.5	70
36	5.0	51
31	4.5	31
26	4.0	16
20	3.5	7
15	3.0	3
10	2.5	1

Method II

Original Scores	Standard Score Interval	Centile Limits of Interval
56 and above	7.0 and above	97.7 to 100
51 to 55	6.5 to 6.9	93.3 to 97.6
46 to 50	6.0 to 6.4	84.1 to 93.2
41 to 45	5.5 to 5.9	69.2 to 84.0
36 to 40	5.0 to 5.4	50.0 to 69.1
31 to 35	4.5 to 4.9	30.8 to 49.9
26 to 30	4.0 to 4.4	15.9 to 30.7
20 to 25	3.5 to 3.9	6.7 to 15.8
15 to 19	3.0 to 3.4	2.3 to 6.6
10 to 14	2.5 to 2.9	0.6 to 2.2
9 and below	less than 2.5	0 to 0.5

Mechanical Comprehension. The particular norms in this table were developed for use with fireman and policeman candidates, and the distribution of scores from which the norms were obtained is shown in Fig. 8:5.

Fig. 8:5. Bennett's Distribution of Mechanical Comprehension Test Scores for 1838 Policemen and Firemen*



*The original data, which were used also in developing Table 8:4, were furnished by The Psychological Corporation, New York, through the courtesy of Dr. George K. Bennett.

The standardizing group, i.e., the group to whom the test was administered for the purpose of developing norms, consisted of 1838 policemen and firemen. Their mean score on the test was 35.6 and the standard deviation of the distribution was 10.1. On the assumption that the distribution in Fig. 8:5 is sufficiently close to the normal, bell-shaped type, the Standard scale norms in Table 8:4 were developed as follows:

$$S = 5.0 + z = 5.0 + \frac{X - M_x}{\sigma_x} = 5.0 + \frac{X - 35.6}{10.1}$$

To find X, an original score value for any value of S:

$$\begin{aligned} X &= M_x - 5.0\sigma_x + S\sigma_x \\ &= 35.6 - 5.0(10.1) + S(10.1) \\ &= 10.1S - 14.9 \end{aligned} \qquad [8:4]$$

To find the original score value of any Standard score of a distribution

Thus, with this formula for X and the particular mean and σ values of the standardizing group of 1838 scores, any original score value for a given value of S can be computed, as follows:

For $S = 5.0$: $X = 10.1(5.0) - 14.9 = 35.6$ (the mean)
For $S = 3.0$: $X = 10.1(3.0) - 14.9 = 15.4$ (or 15)
For $S = 6.5$: $X = 10.1(6.5) - 14.9 = 50.75$ (or 51)

Other values of X for given value of S are computed in the same way.

Two methods for presenting the Standard score norms and their centile equivalents are shown in Table 8:4 for the Bennett Mechanical Comprehension Test, Form AA. Method I is perhaps used more generally than Method II. The latter is simply a clarification of what the norms derived by Method I imply. Thus, according to Method I, an original score of 36 is equal to a Standard score of 5.0, and an original score of 41 is equal to a Standard score

of 5.5. If a person receives a score of 38 on the test, his Standard score lies in the interval between 5.0 and 5.5. This is shown clearly by Method II, in which the scores are given by class intervals.

The Standard score test norms in this table are set up in successive class intervals, each of which is equal to one-half standard deviation. This division of the scale would theoretically yield 12 intervals between Standard score values of 2.0 and 8.0. However, as indicated in Fig. 8:5, the actual distribution of the 1838 scores is not bilaterally symmetrical. That is, the tails of the distribution are not equidistant from the modal point of the curve; rather, the distribution tends to be *skewed* in the direction of the lower scores. The mean of 35.6 is consequently just below the modal interval, which ranges from 36 to 38. Furthermore, the maximum possible score that can be made on this test is 60, and therefore the tail at the right, or toward the higher scores, could not extend relatively as far from the mean as the tail at the lower end of the distribution. The actual range of scores of the distribution is thus less than the theoretical Standard score range of from 2.0 to 8.0. Eleven Standard score intervals, rather than 12, suffice to give a table of norms in terms of the actual data obtained from this group of policemen and firemen.*

The Standard Score Profile Chart or Psychograph

Centile vs. Standard Score Scales of Test Difficulty

Centile scores alone can be used for the development of norms. They are particularly desirable in comparing test results for distributions all of which do not tend to be of the normal bell-shaped type. When, however, distributions whose scores are to be compared are of the normal type, the Standard score scale is preferable to the centile scale because of the great concentration of cases at the center of the distributions. That is, the centile scale is too likely to suggest that differences in test difficulty are the same throughout the scale—that the difference in difficulty between C_{50} and C_{60} is the same as that between C_{80} and C_{90}. Actually, for most tests, a subject needs to achieve success on considerably fewer items to raise his place in the centile scale from C_{50} to C_{60} than to raise it from C_{80} to C_{90}.

Another advantage of a Standard score scale over a centile scale is the fact that it gives more consideration to the difficulty of the test. This does not mean that the standard deviation itself is a basis for differentiating test scores on a scale that yields equal units of difficulty. In other words, it does

* It should be noted that Bennett's norms for policeman and fireman candidates, published with the Mechanical Comprehension Test, are presented in terms of centiles (20 intervals), rather than in terms of Standard scores. However, Standard scores can be used for somewhat skewed distributions such as that in Fig. 8:5, if their divergence from the normal, bell-shaped type of distribution is no greater than might be expected on the basis of chance. See chap. 15, Section A, for a statistical test of the possible significance of the divergence of a uni-modal, skewed distribution from the normal bell-shaped type.

not follow that any two standard deviation intervals are equal in difficulty; rather, the Standard score scale is more closely related to the differences in difficulty than is the centile scale. However, the spacing of centile intervals on a scale can be adjusted to correspond to intervals based on the standard deviation. This adjustment is illustrated in the profile charts in Fig. 8:6 and 8:7.

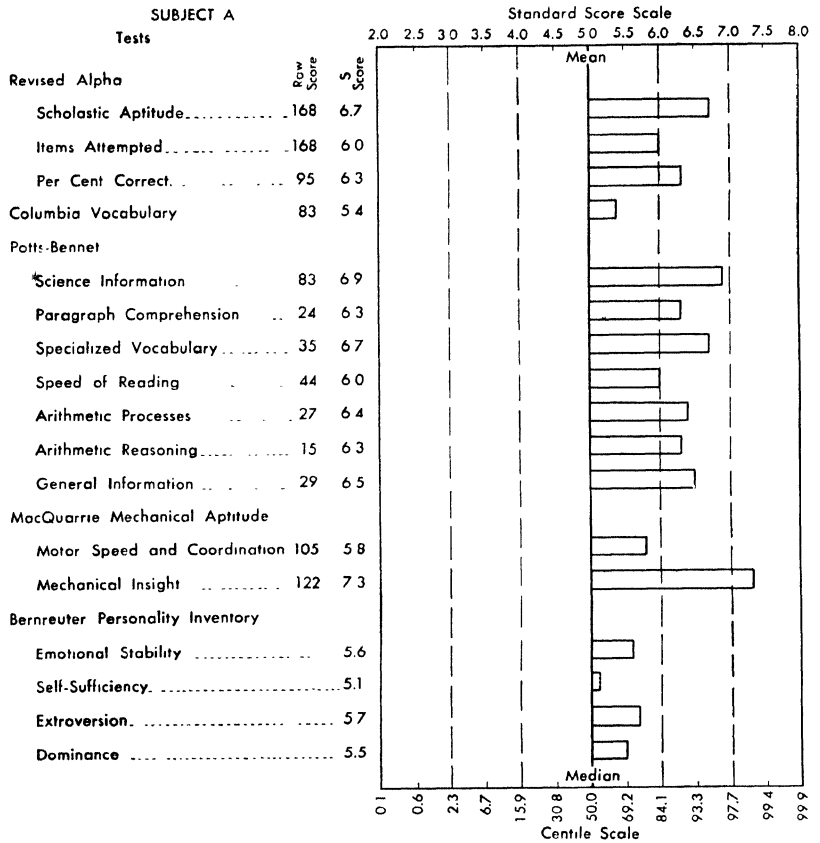
Prerequisites for a Profile Chart

The purpose of a profile chart is to provide a graphic device by which a person's placement on one test can be directly compared with his placement on other tests. Thus, a profile chart is a means of summarizing and comparing one person's results on a battery of tests. It shows at a glance whether he did well on all the tests, or poorly, or whether his performance was scattered. One assumption basic to the development of a profile chart cannot be overemphasized: The distributions for each test whose scores are compared must be derived from the *same* group of individuals (or, in the case of sampling statistics, from samples of the same type or kind of population). In other words, a profile chart is developed on the basis of norms that have been established for each test whose scores are to be compared, and the norms in turn must be obtained from the same group of subjects.

The importance of this basic assumption can be brought out by the following: Mr. Jones is given two different tests, and his performance on the first is compared with his performance on the second by means of the norms provided with each test. According to these norms, he has a Standard score of 6.0 on the first test and a Standard score of 5.0 on the second. However, the norms for the second test were developed from the test results of a restricted group of people who were above average in ability, whereas the norms for the first test were developed from the test results of a group that was not restricted in the general range of ability. Mr. Jones' Standard scores of 6.0 and 5.0 are therefore not on comparable scales, and they should not be plotted on the same profile chart. His Standard score of 5.0 on the second test would undoubtedly have been higher had it been based on norms derived from the test results of the first group whose range of ability was not so restricted.

The essential characteristic of the profile chart thus is the fact that it should provide a standardized matrix that can be used for comparing an individual's placement or position on two or more tests. Unless the norms for each test result are derived from the same group or type of population, ambiguous, if not bizarre, interpretations will result. Obviously the Standard score of one test, based on adult norms, cannot be compared with the Standard score of another test, based on norms for 10-year-olds. Nor should norms derived from, say, a group of college graduates be used on the same profile chart with those derived from adults generally. The *point of reference* on the chart is the center vertical line that represents the mean (Standard score of 5.0)

Fig. 8:6. Individual Psychograph or Profile Chart *



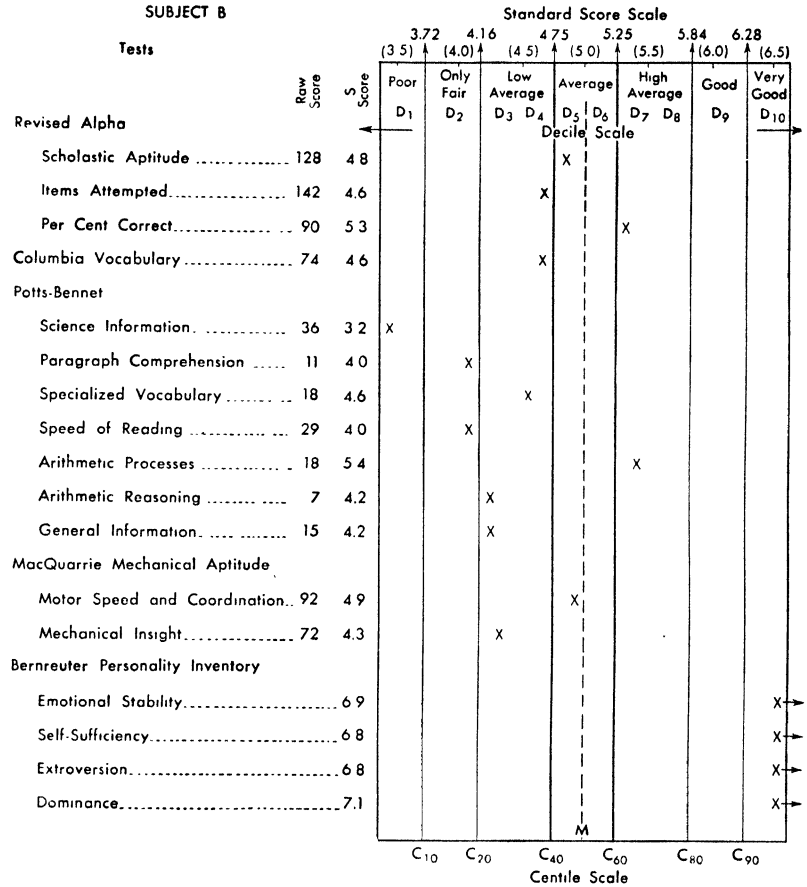
* Data for this person, courtesy of Dr. George K. Bennett, The Psychological Corporation, New York.

of each test whose scores are to be compared. Only if the means for each test are derived from the same group will they be comparable. Similarly, the *scale of differentiation* on a profile chart is established in terms of the standard deviation of each test. Only if the standard deviations of each test are derived from the same group, whose results approximate normal, bell-shaped distributions, will they be satisfactory for comparisons made in terms of Standard scores.

The Construction of a Profile Chart

Two variations in the Standard score profile chart are shown in Figs. 8:6 and 8:7. Both are similar in that the tests whose scores are compared are listed *down* the chart on the left-hand side; vertical lines are employed to

Fig. 8:7. Individual Psychograph or Profile Chart *



* Data for this person, courtesy of Dr. George K. Bennett, The Psychological Corporation, New York.

mark different points on the test scales; and the basis for these differentiations of the test results is the Standard score scale. In Fig. 8:6 the practical range of possible test results from a Standard score of 2.0 to one of 8.0 is scaled across the top of the chart in equally spaced intervals of one standard deviation each; and rectangles are employed to represent the subject's placement, or relative position, on each of the 17 test variables. Fig. 8:7, on the other hand, is so constructed that the D range (from C_{10} to C_{90}) for the middle 80% of the cases is enlarged; descriptive terms are employed for four decile intervals (two at each extreme) and for three quintile intervals at the center

of the scale; and all such intervals are based on the Standard score scale, as indicated across the top of the chart.* The individual's placement on each of the variables is represented in Fig. 8:7 by a cross, rather than by a rectangle projected laterally from the mean point of reference as in Fig. 8:6.

The profile chart shown in Fig. 8:7, with descriptive terms for the decile and quintile intervals, is probably more widely used than the straight Standard score chart shown in Fig. 8:6. It should be emphasized, however, that the Standard score scale is used for both. In Fig. 8:7 this device is combined with the centile scale.

Both profile charts represent the results obtained on a battery of tests by two women applicants for admission to a school of nursing. The scores of each candidate are compared with the results obtained on these 17 tests by 10,000 applicants for admission to schools of nursing. In other words, these 10,000 cases constitute the normative group from whose test results on all 17 variables the Standard score scale was developed for each chart. Thus the basic requirement of a profile chart is met, namely, that the Standard scores must be based on means and standard deviations of tests all of which have been administered to the *same* group or kind of population.

It will be observed that less than 17 different tests were administered these two candidates, despite the fact that their results on 17 different variables are compared. The first test listed, the revised Army Alpha, was scored three ways, as indicated, to yield three variables. The Bernreuter Personality Inventory was also scored several ways to yield the four variables indicated. The Potts-Bennett Tests comprise a series of tests within a battery, each of which is scored separately rather than in terms of an over-all score.

The original scores made by each candidate on each of the first 13 variables are given in the first column at the right of the test names. The corresponding Standard scores are given in the next column at the right, and the results on the Bernreuter Personality Inventory are also indicated in this column. The general *order* in which the groups of variables is presented is somewhat arbitrary; however, two or more variables derived from the same test or inventory are of course listed together. It is because the general order is arbitrary that separate rectangles (as in Fig. 8:6) or crosses (as in Fig. 8:7) are used to denote the subject's position on each variable, instead of the successive crosses being connected with straight lines down the page. However, it was the latter type of chart that gave rise to the term "profile."

It is apparent that the patterns of each applicant's psychographs differ considerably. Thus Subject A (Fig. 8:6) does consistently above average on all the variables, and scores "Good" or "Very Good" on all the test variables except the Columbia Vocabulary. Her results on the Bernreuter are "Above Average" or "Average." Subject B, on the other hand, gives a considerably

---

* Although the theoretical limits of the first and tenth decile intervals are infinity, they are necessarily limited on the profile chart. These limits are S scores of 3.375 and 6.625.

more varied psychograph (Fig. 8:7), with a range from "Poor" on the Science Information Test to "Very High" on the Bernreuter Personality Inventory. Because of her excellent personality ratings, Subject B may prove to be the better nurse. The question in her case is whether she has sufficient aptitude to take the training required by the particular school of nursing to which she applied. If there were a shortage of candidates, she might well be considered; but if there were so many candidates that only a small proportion could be admitted, the final decision would have to be based on a consideration of the results for many other candidates as well as on Subject B's personality, history, and test performance. As for the test performance itself, it is to be emphasized that tests do not all have the same degree of reliability or validity (cf. Chapter 17, Section A), and their differences in this respect must be taken into account. Subject B, for example, has at least a "Low Average" rating in most of the more reliable variables.

The centile equivalents of the Standard score divisions of the psychographs in Figs. 8:6 and 8:7 are obtained by reference to Table 8:1. In order to determine the Standard score value that will be at C_{90} on the scale, the score value is located at a point that divides the distribution into two parts, with 10% beyond C_{90}. This is the point that marks the limit of 40% of the area above the mean. We locate in Table 8:1 the nearest value to 40, and we find it in the row for $x/\sigma = 1.2$ (or a Standard score of 6.2, since $S = z + 5.0 = 1.2 + 5.0 = 6.2$), and the next to the last column at the right, headed .08. S is therefore 6.28, or 6.3 when rounded to one decimal place.

The necessary values for the divisions used in Fig. 8:7 are summarized as follows:

C_{90} or better : S = 6.28 or better ; tenth decile —"Very Good"
C_{80} to C_{90} : S = 5.84 to 6.28⁻ ; ninth decile —"Good"
C_{60} to C_{80} : S = 5.25 to 5.84⁻ ; fourth quintile—"High Average"
C_{40} to C_{60} : S = 4.75 to 5.25⁻ ; third quintile —"Average"
C_{20} to C_{40} : S = 4.16 to 4.75⁻ ; second quintile—"Low Average"
C_{10} to C_{20} : S = 3.72 to 4.16⁻ ; second decile —"Only Fair"
Less than C_{10}: S = Less than 3.72; first decile —"Poor"

EXERCISES

1. Summarize the essential properties and implications of the normal bell-shaped distribution.
2. What properties do a rectangular distribution and the normal bell-shaped distribution have in common?
3. Why are the "practical limits" of the normal bell-shaped distribution taken as equal to plus and minus three standard deviation units from the mean?
4. What fundamental purpose in measuring people's abilities is served by z scores?
5. Under what circumstances can z scores be unambiguously interpreted in terms of centile intervals?

6. Compute the z score equivalents of the following original scores for a distribution whose mean is 85.3 and whose standard deviation is 14.5:
   a. an original score of 92.0
   b. an original score of 46.7
   c. an original score of 112.0
   d. an original score of 85.4
   e. an original score of 58.0
7. On the assumption that the distribution in the preceding exercise is of the normal, bell-shaped type, indicate the centile intervals in which each of the computed z scores lies.
8. Why is the value of the standard deviation of a distribution always larger than the value of the average deviation?
9. Convert the z scores obtained in Exercise 6 to Standard scores.
10. On the assumption that a distribution is of the normal bell-shaped type, and that its mean is 125.0 and its standard deviation 30.0, determine original score values of the following Standard scores:
    a. 5.2
    b. 2.5
    c. 7.6
    d. 4.1
    e. 6.3
11. On the assumption that the distribution of the Bennett Mechanical Comprehension test scores in Fig. 8:5 is normally distributed, determine the original score equivalents of the following:
    a. a Standard score of 6.2
    b. a z score of -1.1
12. What fundamental research purpose is served by an individual psychograph or profile chart?
13. On what assumptions is the use of individual profile charts based?
14. Devise a graphic method for comparing on the same psychograph the individual results of the two persons whose psychographs are presented in Figs. 8:6 and 8:7.

CHAPTER 9

The Product-Moment Method * for the Correlation of Variates

A. THE LINEAR CORRELATION OF BI-VARIATES

The lengths of the radii and circumferences of circles may vary; however, the association between the length of the radius and the circumference of circles is such that the relation between them is perfect. That is, any known variation in the length of the radius of a circle is accompanied by a definite amount of change in the length of the circumference of a circle. This relationship is expressed by

$$C = 2\pi r, \text{ or radius} = \frac{C}{2\pi}$$

Similarly, the height and width of the sides of a square are perfectly related. If the height of a square equals y and its width equals x, then x is always equal to y. Knowing the length of either side, we can compute, without error, the length of the other. Thus,

$$y = x, \text{ or } x = y$$

These are examples of relations which are perfect; that is, the relations are such that there is no variation in the length of the circumferences of circles with a given size of radius and no variation in the length of the side of squares with a base of a specified length.

In contrast, the relations characteristic of biological and social phenomena are not perfect. Co-relations between attributes or aspects of such phenomena are expressions of *some degree* of co-association or co-variability. Such co-relations may range from no correlation at all to values approaching perfect correlation. If the phenomena of the biological and social sciences were as strictly related as the mathematical properties of geometric figures, there would never have been any occasion for the development of statistical methods. It was because observation and measurement showed such relationships to be variable that the special techniques of applied mathematics, i.e., statistical methods, were developed. Historically, of course, the discovery was not so much finding that the co-relations of natural and social phenomena are variable as it was finding that these apparently chaotic and very complex

* This method of correlation is called the *product-moment method* because it is based on the method of moments (mean and standard deviation) described in chap. 7.

phenomena are less chaotic than originally imagined. As was indicated in Chapter 1, Quetelet and Galton were especially instrumental in making systematic observations and developing methods that served to describe the "law and order" characteristic of many aspects of these kinds of events.

The study of the possible relationships between two or more attributes or characteristics has been integral to the development of both the biological and the social sciences because it is through such studies that problems of law and causation in these fields have been opened to investigation. The analysis of laws and causal relations among variable phenomena is basically dependent upon the method of correlation. The statistical problem is one of employing a mathematical technique that will yield a satisfactory index of the *degree* of co-variation. Casual observation may suggest that short people weigh less than tall people and that tall people weigh more than short people, or, conversely, that people who weigh less tend to be shorter than people who weigh more. However, casual observation also indicates that the relationship between the height and weight of individuals is not perfect. The problem therefore is one of determining the degree of co-variation in height and weight. As shown in Chapter 4, the degree of co-variation between two co-related variables is expressed mathematically by the correlation coefficient. Such a coefficient, whose value depends upon the degree of co-variation manifest in the relation of two variables, may range from 0 (no correlation at all) to a value approaching $+1.0$ or -1.0 (in other words, approaching perfect correlation).

Casual observation also indicates that there is some correlation between the height and age of persons during the period of growth, but that after maturity there is likely to be very little correlation between these two variables. Psychological research has revealed that there is some degree of correlation between people's performances on psychological tests and their actual behavior in educational or working situations. The use of the Army classification tests is based on the well-tested observation that there is a relation between achievement on these tests and performance in many types of Army training and occupational situations. Similarly, research has clearly indicated that there is some degree of correlation between personality appraisals of individuals and their capacity to succeed in various types of training activities, such as dive bombing.

Correlation studies of the relationships between two attributes or factors are usually made for the purposes and problems of sampling and analytical statistics rather than for descriptive statistics alone. However, the degree of possible correlation between the empirical data of two variables is a *descriptive* problem. It is for this reason that methods for estimating and computing the degree of correlation between two variables are presented in the first part of this book. Later, in Chapters 16–18, we shall present implications of and procedures for correlation methods employed in studying populations through the analysis of sample data.

Pearson's Product-Moment *r*

The product-moment correlation coefficient, developed by Karl Pearson and symbolized by *r*, is often referred to as the Pearson *r*. It is equal to the ratio of the mean of the products of the paired deviations for the two variables correlated to the product of their respective standard deviations. Thus:

$$r = \frac{\frac{\Sigma(xy)}{N}}{\sigma_x \sigma_y} = \frac{\Sigma(xy)}{N\sigma_x\sigma_y} \qquad [9:1]$$

Pearson's product-moment correlation coefficient (*r*)

Before describing methods for the computation of *r* (Section C of this chapter) we shall give the background of its development and some of the considerations on which it is based, and present a graphic method by which it can be obtained. The latter will in particular serve to illustrate what *r* means, regardless of the method used to determine it.

The Cross-Tabulation of Bi-Variate Data

We saw in Chapter 4 that cross-tabulation of the data of two attributes is essential to studying the possible correlation between them. This is generally the case, whether the co-relationships being investigated are for the data of non-variable attributes or for the data of variates. The cross-tabulation of the data of bi-variates involves basically the same procedure as the cross-tabulation of the data of dichotomized attributes into a fourfold table. The chief difference is the fact that the cross-tabulation of bi-variate data is made in reference to a correlation matrix that is set up for *continuously distributed* variables. The simplest method of determining *by inspection* whether there is any noticeable correlation between two variables is to cross-tabulate the data into a graph known as a scattergram.

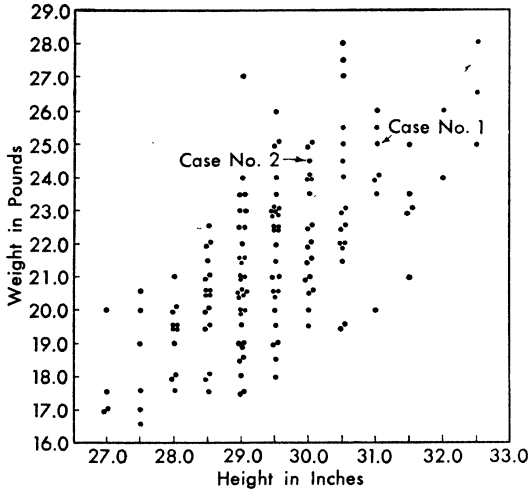
The Scattergram of Bi-Variate Data

A scattergram represents a cross-tabulation of bi-variate data that is usually made directly from the original data of each variable. That is, the data of the two variables correlated are not grouped into a limited number of class intervals but are plotted directly from the original measurements.

A scattergram that describes a fairly high degree of positive correlation is presented in Fig. 9:1. The matrix used for plotting a scattergram is the coordinate axes of a geometric field. The observed values of one variable are scaled on the *x*-axis, or abscissa. The observed values of the other variable are scaled on the *y*-axis, or ordinate. When the two variables being correlated are known or presumed to be causally related in a way such that variations in one variable are in some way and to some degree dependent upon variations in the other variable, the latter is called the *independent* variable, and is scaled on the *x*-axis. The former is called the dependent variable and is scaled on the *y*-axis.

In the height-weight data in Fig. 9:1, the height measurements have been scaled on the x-axis and the weight measurements on the y-axis. However, this should not be interpreted as necessarily implying that height is an independent variable such that weight is dependent upon it. Height and weight are both attributes or qualities of organisms, and hence any relationship between them is a function of the fact of organic unity. But the height or length of organisms is often thought of as being a more fundamental aspect or quality of the organism than is weight. For one thing, the weight of organisms is affected by environmental circumstances of growth and of living more than is their height. Nevertheless, such a difference does not in itself indicate that height is an independent variable and weight dependent on it. It would not be inaccurate, therefore, to scale the height measurements on the y-axis and the weight measurements on the x-axis.

Fig. 9:1. Scattergram of the Heights and Weights of One-Year Old Girl Infants *



* Data from J. G. Peatman and R. A. Higgons, "Growth Norms from Birth to the Age of Five Years: A Study of Children Reared with Optimal Pediatric and Home Care," American Journal of Diseases of Children, 44:1233–1247, 1938.

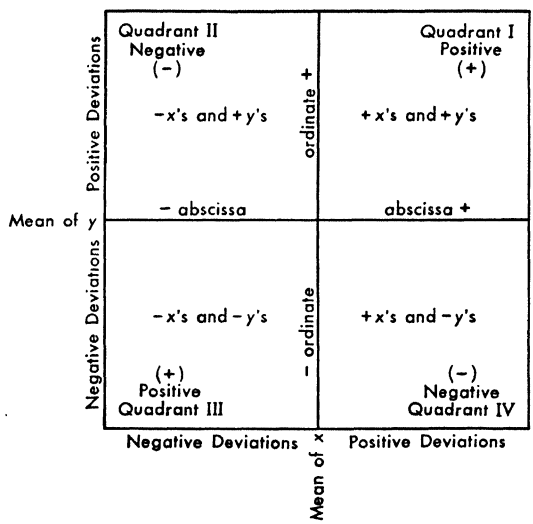
Whichever way any two variables are scaled, the result should not carry the implicit assumption that variations in the quality or factor scaled on the y-axis are dependent upon variations in the factor or quality scaled on the x-axis. The choice of axes for the two variables being correlated is usually purely arbitrary, and any evaluation of the result with respect to the problem of causality must be based upon information about the nature of the variables themselves, rather than upon the way in which they happen to be scaled on the coordinate axes.

Laying off the Scales of a Scattergram

A standard procedure is usually followed in scaling the measurements of each variable for a scattergram. Although the procedure can be varied, following the standard practice gives a descriptive picture that is unambiguous in

its implications. The standard procedure has already been described in Chapter 4, in connection with the correlation of the dichotomized data of bi-variates. As we saw there, a cross-tabulation of correlational frequencies is less likely to be interpreted ambiguously if the scales of each variable are laid off so that the result corresponds to the implications of a geometric field. A geometric field is a matrix which is divided into four quadrants by the co-ordinate axes. These quadrants are established for correlation by the intersection of lines drawn perpendicularly from the respective mean values of each scale. Such a relationship is illustrated in Fig. 9:2.

Fig. 9:2. The Four Quadrants of a Geometric Field



The upper right-hand quadrant of a geometric field is *positive*, because high values of one variable associated with high values of the other variable are located in this part of the total field. This is usually designated as Quadrant I. Similarly, low values of one variable associated with low values of the other are located in the lower left-hand quadrant, designated as Quadrant III. This quadrant is likewise positive because the variations, or deviations, of measures from the mean of each variable are negative, and the product of paired negative deviations is positive. The remaining two quadrants, II and IV, are both negative. Any paired observations located in Quadrant II, in the upper left-hand corner of the figure, represent instances in which the measures of the x variable are less than the mean of x, and the measures of the y variable are greater than the mean of y. Conversely, any paired observations located in Quadrant IV represent instances in which the X values are greater than the mean of x, and the Y values are less than the mean of y.

In order, therefore, for a scattergram to yield a result whose implications correspond to the positive character of Quadrants I and III and the negative character of Quadrants II and IV, the measures of each variable must be scaled as follows:

1. The measures of the x variable are scaled on the abscissa, beginning with the smallest values at the left side of the scale and ending with the largest

values at the right side of the scale. (In the height measures in Fig. 9:1, the least height was 27 inches and the maximum height was $32\frac{1}{2}$ inches.)

2. The measures of the y variable are scaled on the ordinate, beginning with the smallest values at the bottom of the scale and ending with the largest values at the top of the scale. (In the weight data in Fig. 9:1, the least weight was $16\frac{1}{2}$ pounds and the maximum weight was 28 pounds.)

When the measures of two variables being correlated are scaled in this manner, the scattergram gives a result that associates paired high values in positive Quadrant I, paired low values in positive Quadrant III, and low values of one variable with high values of the other variable in the negative Quadrants II and IV. Actually, of course, correlation is not simply a question of whether values are large or small; rather, the first step in the problem is the location of paired *deviations*, taken from the means or their respective variables. In other words, positive deviations of both variables are associated in Quadrant I; negative deviations of both variables are associated in Quadrant III; and negative deviations of one variable are associated with positive deviations of the other variable in Quadrants II and IV.

In psychological measurement there is a logical exception to scaling magnitudes in the order just described for the x- and y-axes. This exception holds when either one or both of the two variables being correlated yield measures such that the higher values mean *less* psychological ability or capacity, and the lower values mean *greater* psychological ability or capacity. Such a result typically appears in any measure of ability based upon the time required to achieve or complete a given number of tasks or items. Under these circumstances, the person with the smallest test score (in terms of time required) manifests psychologically greater ability than the person with the highest score (in terms of time required). In such cases the order of the scores is usually reversed so that the direct psychological implications of the scattergram will correspond to the usual meaning of a geometric field.

In practice, the means of each variable are often not drawn on a scattergram, because a scattergram, as ordinarily used, is a preliminary device to depict such correlation as may be present between two variables and is often made before the means are computed. If, however, the means of each variable are available, it is best to draw the intersection of the axes as projections from these means, as was done for the data in Fig. 9:3. It is to be observed that the great majority of the paired measures in this scattergram are in the negative quadrants (II and IV). Only 13 cases are located in positive Quadrant I, and only 9 in positive Quadrant III.*

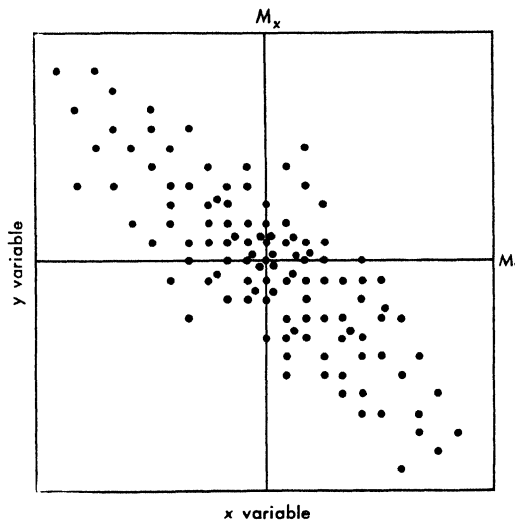
* Many years ago Sheppard, the English statistician, developed a geometric method for estimating a correlation coefficient from the ratio of frequencies in the four quadrants of the geometric field. The methods of tetrachoric correlation, described later in the next chapter, and phi correlation (chap. 4) are based upon ratios of positive and negative correlational frequencies.

As already indicated, the principal use of a scattergram is to give a graphic picture of the correlation between two variables, with respect to both degree and nature of the relationship. In the height-weight measures in Fig. 9:1, the degree of relationship is fairly well marked; the nature of the relationship is *positive* and appears to be *linear*. An inspection of the dots on the scattergram will bear out this statement.

The Assumption of Linear Correlation

The co-relation between any two variables can be described as either linear or non-linear. The product-moment method of correlation is based upon the assumption that such relationship as exists between two variables can be adequately described by a linear rather than a curvilinear function. In other words, the method is based upon a straight-line relationship. This means that for each successive change or difference in the measures of one variable, there is a proportionate change of a constant amount in the other variable. The direction of the change may be negative or positive. The term rectilinear correlation is synonymous with linear correlation. By contrast, curvilinear correlation occurs when the changes in the relationships are not of a constant, proportionate amount, and consequently cannot be described adequately by a straight line.

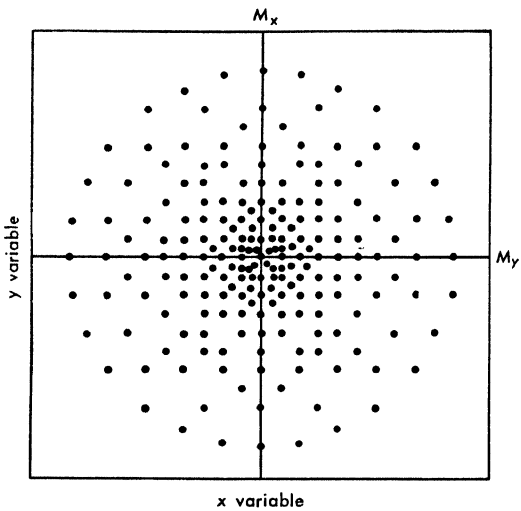
Fig. 9:3. A Scattergram That Illustrates Negative Correlation *



* Note that the correlation is linear and that as the degree of correlation increases, whether negative or positive, the scatter tends to form the shape of an ellipse that becomes increasingly narrow. The scatter of zero correlation tends to form a circle; cf. Fig. 9:4.

Linearity of a co-relation is obviously manifest when the scatter of the paired measurements of bi-variates clusters about an imaginary straight line. That the correlation shown in Fig. 9:1 is linear is apparent. A straight line can also readily be applied to the scatter in Fig. 9:3, illustrating negative correlation. However, in Fig. 9:4, illustrating zero correlation, it may not be so apparent that a line of any kind can be fitted to the scatter of the data. There is obviously no correlation between the two variables. That is, low values of

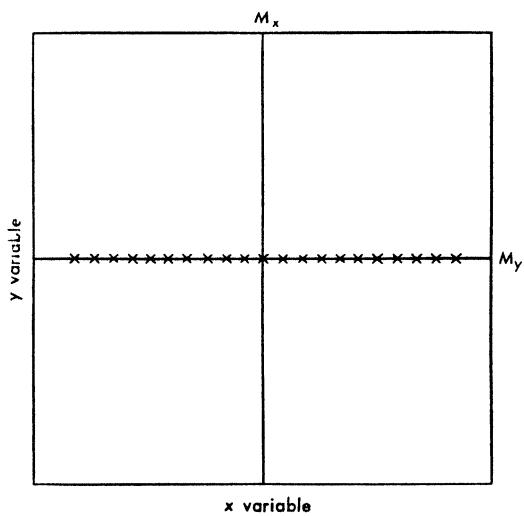
Fig. 9:4. A Scattergram That Illustrates Zero Correlation



the x variable are associated with as many high as low values of the y variable, and, similarly, high values of the x variable are associated with as many low as high values of the y variable.

In order to fit a line that will describe the trend of non-correlation, the *average variation* in the values of one variable must be plotted with respect to successive values of the other variable. This has been done in Fig. 9:5 (with the data from Fig. 9:4) for the variations of the y variable associated with successive class-interval values of the x variable. It is now apparent that a horizontal straight line can be fitted to the data. The *slope* of this line is zero.

Fig. 9:5. The Average Variation of Measures of y with Respect to x *

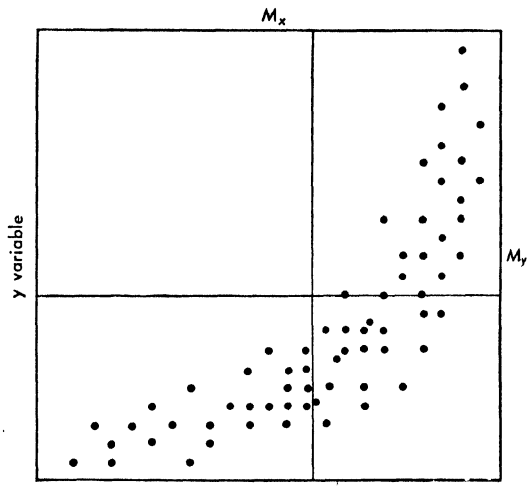


* This illustrates the Fact that the Zero Correlation of Fig. 9:4 is linear.

The method of product-moment correlation which yields the coefficient r (see Formula 9:1) is based upon the assumption that the co-variation observed in a set of observations can be adequately described by a straight-line function. At this point, we wish to emphasize that the making of a scattergram (or correlation chart, see page 207) is important in any study of the correlation between two variables, because the graph itself provides a picture that shows whether such relationship as may exist

is in fact satisfied by a linear function. If it is, the product-moment method of correlation is appropriate. But if the co-relationship shown in the scattergram differs considerably from a linear type of relation, non-linear methods of correlation must be employed.

Fig. 9:6. A Scattergram That Illustrates Non-Linear Correlation



In practice, because of a research worker's familiarity with many types of data, it is frequently safe to assume that the co-relationship is linear. In such instances, machine methods for computing the correlation coefficient may be warranted, despite the fact that they provide no cross-tabulated picture of the actual result. But whenever the variables being correlated are unfamiliar, or there is any question as to the nature of the possible co-relationship, a scattergram should be made to determine whether the product-moment method is appropriate for computing the correlation coefficient. The data in Fig. 9:6 represent a case in point. Here the paired observations do not scatter along a straight line nearly so well as along a curved line. If the product-moment method were used to express the degree of correlation between these two variables, the result would be very misleading. The correlation coefficient, r, would be considerably lower in value than a coefficient derived by a method that takes into account the curvilinear character of the association between two variates.

Plotting the Bi-Variate Data of a Scattergram

The data upon which the scattergram in Fig. 9:1 are based are presented in Table 9:1. The height-weight measurements of the first infant, No. 1, are 31.0 inches and 25 pounds respectively. This particular correlational frequency is designated on the scattergram in Fig. 9:1. In plotting each correlational frequency, the height measure is first located on the height scale at the bottom of the scattergram, and the weight measure is then located on the vertical scale at the left. Imaginary lines perpendicular to each scale are then projected from each of these two scale values into the geometric field, and a dot is made at their point of intersection. Each dot on a scattergram, therefore, represents a *correlational frequency* whose y and x values can always be

Table 9:1. Height and Weight Measurements of One-Year-Old Girl Infants

Case Number	Height (In in.)	Weight (In lbs.)	Case Number	Height (In in.)	Weight (In lbs.)	Case Number	Height (In in.)	Weight (In lbs.)
1	31.0	25.0	51	28.0	20.0	101	29.0	20.0
2	30.0	24.5	52	29.5	20.0	102	31.0	26.0
3	29.5	21.0	53	28.5	20.0	103	31.0	24.0
4	29.5	26.0	54	27.5	19.0	104	32.5	26.5
5	27.5	20.0	55	29.0	27.0	105	28.5	20.5
6	30.0	22.0	56	30.5	24.0	106	29.5	18.5
7	30.0	21.0	57	31.0	23.5	107	30.0	21.5
8	30.0	23.5	58	29.0	21.0	108	28.0	19.5
9	29.0	18.5	59	28.0	20.0	109	30.0	24.0
10	29.0	22.0	60	30.0	22.5	110	29.5	22.5
11	29.0	23.5	61	29.0	23.0	111	31.5	23.0
12	31.5	25.0	62	29.5	23.0	112	29.5	19.0
13	29.0	21.5	63	28.0	18.0	113	30.5	19.5
14	29.5	20.5	64	29.0	20.5	114	29.5	22.5
15	28.5	21.0	65	29.0	24.0	115	28.0	17.5
16	29.5	23.0	66	30.5	22.0	116	29.5	19.0
17	32.0	24.0	67	30.0	24.0	117	32.0	26.0
18	30.0	19.5	68	29.0	18.5	118	30.5	22.5
19	28.0	19.0	69	29.5	21.5	119	27.0	17.5
20	30.0	25.0	70	30.0	20.5	120	28.5	18.0
21	28.5	20.0	71	28.5	19.5	121	30.5	24.5
22	29.0	21.5	72	30.5	23.0	122	29.5	20.5
23	29.0	21.0	73	29.0	20.5	123	28.0	19.5
24	30.0	25.0	74	29.0	20.5	124	28.0	18.0
25	30.5	22.0	75	29.0	17.5	125	27.5	17.0
26	29.5	23.0	76	29.5	25.0	126	30.5	19.5
27	30.0	22.0	77	30.5	21.5	127	29.5	18.0
28	29.5	23.0	78	27.0	20.0	128	31.5	23.0
29	28.5	22.5	79	29.0	19.0	129	28.5	19.5
30	29.0	20.5	80	29.5	24.0	130	27.5	17.5
31	29.0	22.5	81	29.0	20.0	131	29.5	19.5
32	30.5	25.5	82	29.0	18.0	132	30.0	20.0
33	29.0	21.0	83	29.5	25.0	133	29.5	22.5
34	28.5	22.0	84	28.0	19.5	134	30.5	23.0
35	29.0	19.0	85	30.0	20.5	135	29.5	20.5
36	30.0	21.5	86	31.5	23.5	136	27.5	20.5
37	29.0	20.0	87	30.0	22.5	137	32.5	28.0
38	29.5	21.0	88	30.0	24.0	138	30.5	27.0
39	29.0	21.5	89	31.5	21.0	139	30.5	27.5
40	28.0	21.0	90	30.5	28.0	140	28.5	20.5
41	29.0	23.5	91	30.0	21.0	141	28.5	21.5
42	31.0	20.0	92	29.0	17.5	142	29.5	22.5
43	28.5	22.0	93	31.0	25.5	143	30.5	25.0
44	27.0	17.0	94	27.0	17.0	144	32.5	25.0
45	29.0	23.0	95	29.0	19.5	145	28.5	17.5
46	29.0	22.5	96	30.5	22.0	146	28.5	20.5
47	29.0	19.0	97	29.5	23.0	147	27.5	16.5
48	29.5	23.0	98	29.5	22.0	148	30.5	26.0
49	31.0	24.0	99	29.0	20.5	149	28.5	18.0
50	28.5	21.0	100	29.5	23.5	150	28.0	19.5
						151	28.5	20.5

obtained by referring to the scales of the coordinates. Thus, the dot for Infant No. 1 represents (reading up) a height measure of 31.0 inches and (reading across to the left) a weight measure of 25 pounds.

The second infant's height and weight measures, as given in Table 9:1, are represented on the scattergram by a dot at the intersection of imaginary lines projected from a measure on the height scale equal to 30.0 inches and a measure on the weight scale equal to 24.5 pounds. The paired data of the remaining 149 correlational frequencies are plotted in turn, to give the scattergram shown in Fig. 9:1. Sometimes, in making a scattergram, two or more correlational frequencies will have the same x-variable value and the same y-variable value. In this case, additional dots to represent the location of such correlational frequencies are plotted in close proximity to the first dot, rather than on top of it.

The Correlational Frequency: Paired Associates

Consideration of the kinds of statistical situations which lend themselves to the method of correlation should make clear the fundamental nature of bi-variate data. Essentially, the data of correlations are derived from the measurements of two attributes or traits that are logically associated by means of a group of *paired observations*. Each pair constitutes a correlational frequency, and unless observations of the variates of two attributes or traits have a factual basis for association by pairs, the method of correlation is not applicable or relevant.

The costs of advertising two products can be *compared* but they cannot be correlated unless there is a group of paired costs that can be cross-tabulated. The basis for associating such a group by pairs in the field of market research is often *calendar time*. For example, the costs of advertising two products can be cross-tabulated for a series of paired costs obtained for successive annual or semi-annual periods. Similarly, *time* is the basis for pairing observations of two variables that may not otherwise appear to have any relationship with each other. In fact, *time* is made the basis for pairing observations often used to illustrate *spurious* * *correlation*, as for example, a correlation between the annual precipitation in New Zealand and the birth rate in Wisconsin. Precipitation and birth rate for the same years can be paired; consequently, such data over a period of twenty-five or fifty years would constitute a group of 25 or 50 correlational frequencies paired on the basis of time. In this case, time would furnish the only basis for the associations by pairs; there is no other logical or reasonable basis for associating two such variables.

In psychology, biology, and the social sciences, the basis for correlational frequencies is usually the *individual organism*. The paired associates for the correlation between two variables such as height and weight, intelligence and

* Spurious correlation means that correlation obtained between two variables is in whole or in part due to factors other than those to which it is ascribed.

educational achievement, aptitude and personality ratings, attitude for A and attitude for B, etc., are obtained in each instance from the measurements or observations of the same persons. A *group* of individuals yields a number of paired associates and hence provides relevant data for cross-tabulation and correlation.

Blood relationship or *social relationships* of various kinds also yield paired associates, and hence a group of data that can be correlated. Galton's original correlational studies of inheritance were made from anthropometric data that were paired for parent and offspring. E. L. Thorndike and others have studied the relationship between the intelligence of siblings—brother-brother, brother-sister, or sister-sister pairs. These are examples of bi-variates derived from *genetically related pairs.* On the other hand, the possible relationship between intelligence, scholastic achievements, personalities, or interests, of best friends, husbands and wives, etc., has been studied by the method of correlation. Intelligence test scores of individuals may be paired with the intelligence test scores of their best friends; attitude scores of husbands may be paired with the attitude scores of their wives. A group of such paired associates is thus an example of bi-variates derived from the data of *socially related pairs.* In correlational problems such as these, a single attribute or trait is usually under consideration. That is, the data of each correlational frequency are measurements or observations of the same quality (as for example, measurements of intelligence, all of which are derived from the same test), but they are paired for genetically or socially related persons.

Some experimental situations in psychology and related fields, especially biology, also give rise to paired associates. This is particularly true of the experimental method of *equated groups,* in which the subjects of both the experimental and the control groups are matched, pair by pair. As we shall see later (Chapter 14, Section F), the correlation between the results for experimental and control groups, individually matched by pairs, is relevant for testing the significance of the mean difference between two such groups.

Two groups of data can often be compared with respect to their central tendencies, deviational tendencies, the form of their respective distributions, etc. But two variables cannot be *correlated* unless there is a logical or reasonable basis for cross-tabulating the data of each variable. Thus, the scholastic achievements of seniors and juniors in a college can be compared but not correlated, unless there is a meaningful basis for associating a particular junior student with a particular senior student, whereby a series of correlational frequencies can be established.

The Correlation Chart

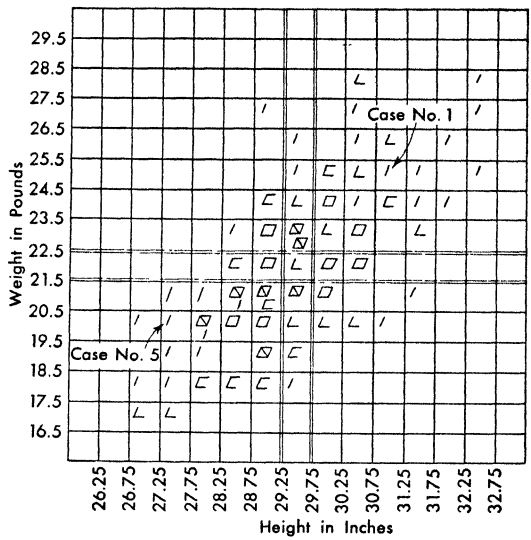
A picture of the distribution of correlational frequencies can also be obtained from a correlation chart. Each of the correlational frequencies in such a chart is represented not by a dot, but by a tally or by the total number of such frequencies in each cell of the matrix. The result is sufficiently similar to

a scattergram to be almost as satisfactory as the latter in depicting the bivariate relationship. Furthermore, the correlation chart has the advantage over the scattergram in that it is set up in such a way that the correlation coefficient itself can be directly computed from the cross-tabulated data.

Figs. 9:7 and 9:8 portray the height-weight data in Fig. 9:1 cross-tabulated into correlation charts. Fig. 9:7 shows a *tally* of the correlational frequencies. Fig. 9:8 represents the same result, but the correlational frequencies of each cell are summed, and the totals for each cell are indicated. The basic purpose of a scattergram is well served by either the correlation *tally chart* in Fig. 9:7 or the correlation *frequency chart* in Fig. 9:8. That is, both figures show (1) that there is a fair degree of correlation; (2) that the correlation is positive; and (3) that the bi-variate relationship is linear one, that is, it can be adequately described by a straight-line function.

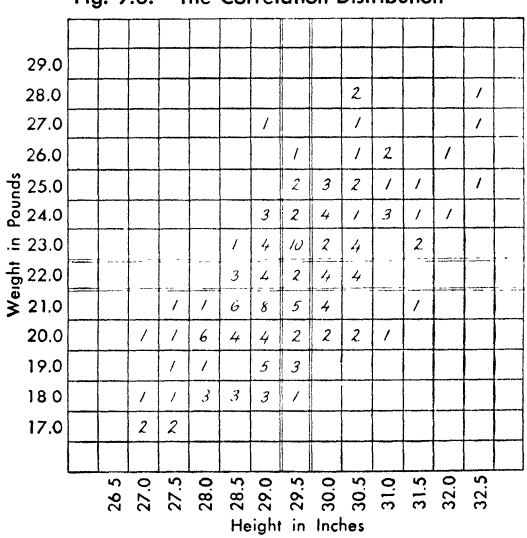
A correlation chart is constructed by procedures similar to those used in setting up the matrix for a scattergram. The chief difference is the fact that the data of each variable

Fig. 9:7. The Correlation Tally *



* Height-Weight Data of Table 9:1.

Fig. 9:8. The Correlation Distribution *



* From Tally of Fig. 9:7.

are *grouped into class intervals* characteristic of frequency distributions. Thus in Figs. 9:7 and 9:8, the height measurements have been distributed into 12 class intervals, each equal to half an inch. The weight data have likewise been distributed into 12 class intervals, each equal to one pound. The use of 12 class intervals for each of the two variables thus produces a correlation chart or matrix with 144 cells. Instead of only a 2 by 2 or 2 by 4 type of correlation table described in Chapter 4, we have, in the correlation of continuously distributed variables, 12 by 12-fold tables, or some other combination that produces a large number of cells. In practice, as was indicated in the development of frequency distributions in Chapter 5, no more than 20 class intervals for a variable need to be employed. And if fewer than 12 class intervals are used, it may be advisable to apply Sheppard's correction * to the standard deviation of each variable before r is computed.

The more marked the correlation between the data of two variables, the more cells there will be with no correlational frequencies. In positive correlations there will be more cells in the negative quadrants (II and IV) with no frequencies, and in negative correlations there will be more cells in the positive quadrants (I and III) with no frequencies.

The correlation tally in Fig. 9:7 was made directly from the original data in Table 9:1. The height-weight measures for Infant No. 1 are represented by the tally in the proper cell in Quadrant I. Similarly, the height-weight measures of the fifth infant are represented by the tally in the proper cell in the third quadrant. The distribution of the correlational frequencies in Fig. 9:8 is obtained from the tally of cross-tabulations in Fig. 9:7. In order to facilitate locating the correlational frequencies in Fig. 9:8, the mid-points of the class-intervals of each variable are used. Mid-point values, rather than the class limit values, are more relevant for further computational work with the data of a correlation chart (cf. Figs. 9:9 and 9:12), because all computations are based on the principles described in Chapter 7 for the mean and the standard deviation. That is, the mid-point of each class interval is assumed to be a representative value for all frequencies within the limits of the interval.

B. ESTIMATION OF PRODUCT-MOMENT r

Before presenting methods for computing the product-moment correlation coefficient, we shall illustrate the mathematical implications of a computed r by *estimating* the coefficient from regression lines fitted to a group of bivariate data.

Fitting Linear Regression Lines to Bi-Variate Distributions

The cross-tabulated relationship (whether zero, positive, or negative) between two variables, as illustrated by the correlation charts in Figs. 9:7

* For Sheppard's correction, see p. 167.

and 9:8, represents a bi-variate distribution. It has already been observed that the bi-variate data in Fig. 9:8 are distributed in such a manner that a *straight line* can be used to describe the nature of the relationship.

Mathematically, two problems arise in treating bi-variate distributions in which two attributes or qualities, x and y, are not invariantly related. If there were perfect correlation, there would be no particular mathematical problem, because x could be expressed as an invariant function of y, and y could be expressed as an invariant function of x. But, as we saw earlier, the statistical problem of correlation arose because bi-variate relations for phenomena of the biological and social sciences are not perfect or invariant. Hence, the first aspect of the mathematical problem is to find the best algebraic formulation that will express the relationship between two variables, x and y. The second aspect consists in developing a method that will denote the *degree* of correlation between two such variables.

One of the simplest types of algebraic relationships is the straight-line function. This is a linear equation that may be expressed as follows:

$$y = mx + k \qquad [9:2]$$

Linear equation (y on x)

where m is a constant denoting the slope of the straight line, and k defines the point at which the line intercepts or cuts across the y-axis of the coordinates. The equation may also be expressed as follows for the same data:

$$x = my + k \qquad [9:3]$$

Linear equation (x on y)

The algebraic formula for a linear relationship between two variables can thus be stated in two ways: either y as a function of x, or x as a function of y. These two equations are called *regression equations*, and straight lines fitted to bi-variate data are called *regression lines*.*

We shall now illustrate a procedure for fitting straight lines to the bi-variate data in Fig. 9:8, and thus describe a method for *estimating* the degree of correlation between the heights and weights of the 151 infants.

The Variation of Weight (y) *with Respect to Height* (x) · · · (y *on* x)

Two straight-line equations may be used to express the correlation between two variables whose relationship is linear. For the height-weight data in Fig. 9:8 these equations are for (1) the variation of weight with respect to

* Linear equations to express the algebraic relationship between bi-variates were employed by Galton in his development of a method of correlation for studying the relationship between characteristics of parents and their offspring. He observed, for example, that tall parents had, on the average, offspring shorter than themselves and that short parents had, on the average, offspring taller than themselves. He described this phenomenon as the law of filial regression (the heights of offspring *regress* toward the parental mean), and hence the equations and lines describing the relationship came to be known as *regression* equations and *regression* lines.

height, and (2) the variation of height with respect to weight. We shall first consider variation of height with respect to weight.

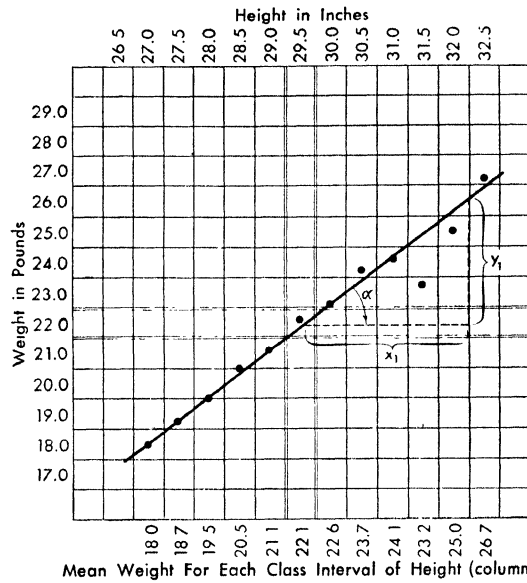
It is apparent from Fig. 9:8 that not all the infants with a height of 27 inches have the same weight; rather, their weight *varies* from 17 to 20 pounds (these are the mid-point values of each class interval for which there are correlational frequencies). Similarly, not all the infants with a height of $32\frac{1}{2}$ inches have the same weight; their weight varies from 25 to 28 pounds. These represent the variation in weight of the infants at the extremes of the height scale. But such variation in weight is characteristic of the infants of a given height at any point on the height scale. Thus, the infants with a height of $29\frac{1}{2}$ inches vary in weight from 18 to 26 pounds. Furthermore, this variation in weight follows an important pattern. Although the shortest infants may be as variable in weight as the tallest infants, the actual range in weight characteristic of their variation is different. Thus, the heaviest of the shortest infants (those with a height of 27 inches) weighed 20 pounds, whereas the lightest of the tallest infants (those with a height of $32\frac{1}{2}$ inches) weighed 25 pounds. There is therefore no overlapping in the weight of the shortest and tallest infants. On the other hand, as we go along the height scale from one class interval to the next, we see that there is a considerable degree of overlapping in the weight of each successive height group.

If the overlapping of the range in actual weights from one height class interval to another were at a *maximum*, the correlation would be zero, because maximum overlapping in the variability of weights would signify not only that the shortest infants were just as variable in weight as the tallest infants, but that the actual range in their respective weights was the same. In other words, if the correlation were zero, the shortest infants would vary in weight, say, from 17 to 28 pounds, and the tallest infants would vary in weight from 17 to 28 pounds. The extent of the variation of y with respect to the successive class-interval values of x is thus basic to the meaning and interpretation of a measure of correlation. The greater the variation in y, the less the correlation between the two variables, or, conversely, the less the variation in values of y for given values of x, the more marked the correlation will be.

If the correlation between x and y were perfect (1.00), all the infants of a given height would have exactly the same weight; there would be no variation in weight for a given height. This is what is meant by an *invariant* relationship. Just as the standard deviation is used to summarize the degree of variation characteristic of a single variable, so it is also used to summarize the degree of scatter characteristic of a bi-variate distribution. When the scatter about the regression line is zero, the correlation is perfect. The use of σ to measure the degree of scatter will be described later (cf. the standard error of estimate, Chapter 16, Section B). At this point it should be emphasized that a measure of the scatter of correlational frequencies about a straight-line function provides an index of the correlation between the two variables. The greater the scatter, the less the correlation; the less the scatter, the greater the correla-

tion. Similarly, the greater the scatter, the less accurately values of one variable can be predicted from given values of the other; and, conversely, the less the scatter, the greater the accuracy with which values of one variable can be predicted from given values of the other.

Fig. 9:9. Means of the Variations in Weight for Successive Class-Interval Measurements of Height *



* From the data of Fig. 9:8.

The variation of weights with respect to the heights shown in Fig. 9:8 is *summarized* in Fig. 9:9. The scatter in Figs. 9:1, 9:7, and 9:8 has already revealed that the variation of weight with respect to height can in all likelihood be described by a straight-line equation. That this is in fact the case is shown in Fig. 9:9. The values plotted represent the *means* of the variations in weight for the midpoint measures of height of successive class intervals. As indicated at the bottom of the figure, the mean of the variation in weight for the infants with a height of 27 inches is 18.0 pounds. This mean value was obtained from the four correlational frequencies in the first column of Fig. 9:8. An inspection of Fig. 9:8 indicates that of the four infants whose height was 27 inches:

2 infants weighed 17 pounds
1 infant " 18 "
1 " " 20 "

The sum of the weights of these four infants is 72 pounds, and their mean weight is therefore 18.0 pounds. Similarly, the mean of the variation in weight of the infants with a height of $27\frac{1}{2}$ inches is 18.7 pounds; the mean weight of those with a height of 28 inches is 19.5 pounds, etc. Thus, a change in height is accompanied by an *average* change in weight.

This is the *meaning* of statistical correlation, or co-variation. To say that there is some degree of co-variation between two variables is to assert that a change in one will be accompanied by some degree of change, *on the average*, in the other. However, it should be emphasized that the implications of

correlation also are basically dependent on the *variability* characteristic of the relationship. That is, a low degree of correlation is indicative of a great deal of *scatter* or variability of measures about the regression line, whereas a high degree of correlation signifies relatively little scatter. The straight line fitted to the scatter indicates the trend in the *average* change in the values of one variable for given values of the other variable. The product-moment correlation coefficient, r, is used as an index that represents all these aspects about linear co-variation.

Inspection of Fig. 9:9, in which all the mean weights are plotted with respect to the successive height measures of each class interval, reveals that, with one exception, all the mean values cluster along a line that can be drawn most simply as a straight line. The exception is the weight of those infants with a height of $31\frac{1}{2}$ inches (twelfth column in the figure). The five infants with a height of $31\frac{1}{2}$ inches had an average weight of 23.2 pounds. However, since only a few cases are involved, and the divergence of this mean from the straight line is not very marked, this has no serious effect on the result.

The straight line drawn to the data in Fig. 9:9 has been fitted by *inspection* rather than by a mathematical method. Although a graph that shows the best-fitting regression lines is not ordinarily constructed in connection with the *computation* of r (see Section C), it is our purpose here to fit a straight line to a group of bi-variate data, *estimate* r from the result, and thereby illustrate what is basically involved in computing r.

For the relation of y on x, the equation of a straight line has already been given as

$$y = mx + k$$

In the case of product-moment correlation, y represents measurements of the ordinate variable expressed as deviations from the mean ($y = Y - M_y$); m is a constant denoting the slope of the straight line fitted to the means of the variation in y taken with respect to successive class interval values of x; and k represents the point of intercept on the ordinate axis of the straight line. For straight-line functions for bi-variate data, this point of intercept is at the intersection of the *means* of the y and x variables. Since the intersection of the means is taken as the origin of the coordinates, k is equal to zero. In other words, a straight line fitted to bi-variate data should pass through the origin if a linear function satisfactorily describes the relationship.

The equation of the straight line fitted to bi-variate data thus becomes $y = mx$. In order to use this equation in estimating values of y (weights) from given values of x (heights), it is necessary to obtain a value for m that is equal to the *slope* of the fitted straight line. The slope of the straight line fitted to the data in Fig. 9:9 is equal to the tangent of the angle made by this straight line with the x-axis. This angle is designated as α, and its tangent is equal to the ratio of y_1 to x_1. For the data in Fig. 9:9, m is equal to a ratio of approximately $\frac{4.0}{5.5}$, or .73.

This value, .73, thus represents the slope of the straight line fitted to the data in Fig. 9:9. However, the actual value of the tangent of α depends upon how the two variables have been scaled in the correlation chart. Were it not for m's dependence upon the method of scaling, the preceding ratio could be used directly as a satisfactory estimate of the degree of correlation between the two variables. But this ratio is unsatisfactory in its present form because a value of correlation obtained by this method is capricious, since it depends upon the way in which the y variable and x variable are scaled on the two coordinate axes. What is needed, therefore, is a method of scaling each of the two variables that will have the same implications for all such problems. Fortunately, this problem can be solved by converting the measures of both variables to z scores (cf. p. 177). That is, if the original measures for each variable are taken as deviations from their respective means and are expressed in terms of their respective standard deviations, the bi-variate data will be comparable in all problems for which linear correlation is appropriate. Furthermore, the slope of the best-fitting straight line for such transformed bi-variate data will have unambiguous implications; the value of the slope of the best-fitting straight line is thus the product-moment correlation coefficient, r.

The z Score Correlation Chart

In order to estimate r from the actual data in a correlation chart, the original values of each variable must be converted into z scores. We have seen that a z score is a deviation (x) taken in terms of the standard deviation of its distribution. Thus:

$$z_x = \frac{X - M_x}{\sigma_x}$$

And similarly,

$$z_y = \frac{Y - M_y}{\sigma_y}$$

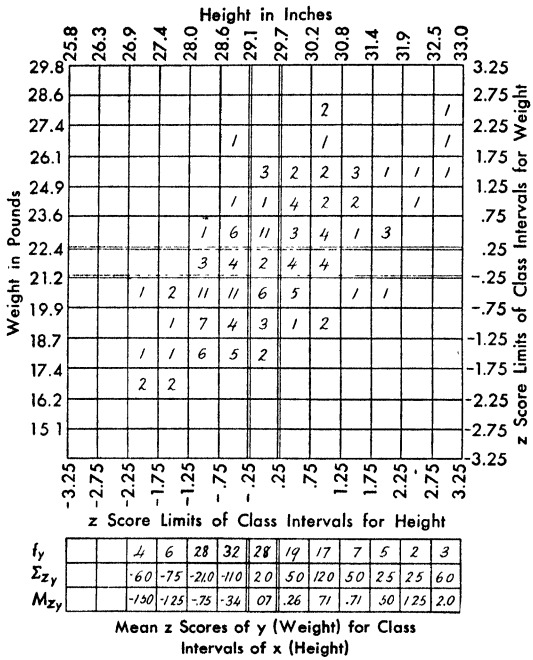
There are two ways in which the height-weight data in Fig. 9:8 can be converted into z scores. Either each weight and height measure in Table 9:1 can be converted into z_x and z_y, or a correlation chart can be constructed in which the limits of the intervals on the height and weight scales are set up both in terms of z and in terms of corresponding values of the original measures. When many cases are to be cross-tabulated, the first method is cumbersome and time-consuming. Consequently, the latter method will be used here; it is illustrated by the correlation chart shown in Fig. 9:10.

Conversion of Original Score Limits to z Score Limits

In this correlation chart, the scales of the x and y variables have been made comparable in terms of their respective standard deviations. In other words, the scales of both variables have been converted into z scores. For illustrative purposes, the range of each class interval has been taken as equal to a z score

value of .50, or half a standard deviation unit. The z score limits of each class interval for the x variable (height) are scaled at the bottom of the chart. The corresponding original height score limits (in inches) are scaled at the top of the chart. Similarly, the weight variable is scaled in z score units at the right of the chart and the original weights (in pounds) are scaled at the left. By this method, the original height and weight measures in Table 9:1 can be cross-tabulated in the correlation chart and at the same time converted to z score values without any additional computations. The original data are thus fed into an original score correlation chart, and they come out as z scores.

Fig. 9:10. A z Score Correlation Chart for the Estimation of Product-moment r *



* Height-Weight Data of Table 9:1.

The computations required in this method of transforming the correlational frequencies of original scores into z scores consist in (1) computing the means and standard deviations of each variable, and (2) determining original score limits that correspond to the desired z score limits. On computation, the means and standard deviations of the height-weight variables are as follows:

$$\begin{aligned} \text{mean weight} &= 21.8 \\ \text{mean height} &= 29.4 \end{aligned}$$

$$\begin{aligned} \sigma \text{ weight} &= 2.48 \\ \sigma \text{ height} &= 1.14 \end{aligned}$$

It will be recalled that the z score value of the mean of a distribution is equal to zero, since a mean value does not deviate from itself. Thus, if $M_x = 29.4$,

$$z_{M_x} = \frac{29.4 - 29.4}{1.14} = 0$$

And since the z score value of any height score is equal to

$$z_x = \frac{X - M_x}{\sigma_x} = \frac{X - 29.4}{1.14}$$

any value of X in terms of z will be equal to

$$X = z_x\sigma_x + M_x \qquad [9:4]$$

To determine the value of an original score from a z score

Thus, $$X = z_x(1.14) + 29.4$$

And the X value of $z = 0$ (the mean z score) will be:

$$X = 0(1.14) + 29.4 = 29.4 \qquad (X, \text{ when } z = \text{zero})$$

The value of X can thus be obtained for any z score value by means of Formula 9:4, the conversion formula. For Fig. 9:10, the successive X values of the limits of the successive class intervals are needed. If these are obtained by starting at the mean of the distribution, the height measurement corresponding with a z score value of .25 will be equal to

$$X = .25(1.14) + 29.4 = 29.68 \qquad (X, \text{ when } z = .25)$$

This is the limit indicated at the top of the correlation chart in this figure for a point in the scale of height scores corresponding to a z score of .25 (at the bottom of the chart). Similarly, the height measurements corresponding to z score values of .75 and 1.25 are:

$$X = .75(1.14) + 29.4 = 30.26 \qquad (X, \text{ when } z = .75)$$
$$X = 1.25(1.14) + 29.4 = 30.82 \qquad (X, \text{ when } z = 1.25)$$

In this fashion the height measurements corresponding to the z score limits of each class interval in the height scale can be determined. However, an alternative and simpler procedure is as follows:

1. Determine the height measurements for $z_x = .25$ and $z_x = -.25$.
2. Determine the range in height units (inches) of a class interval equal to a range of .50 z score units (one-half a standard deviation).
3. Add the range value in inches obtained in the preceding step to the X value of $z = .25$, to find the value of $z = .75$ in inches. Continue adding in this fashion to determine the value in inches of each successive class-interval limit above the mean height.
4. Subtract the range value in inches of a class interval in height (obtained in the second step) from the X value of $z = -.25$, to find the value of $z = -.75$ in inches. Similarly, subtract the same range value in inches from the height value of $z = -.75$, to obtain the value of $z = -1.25$ in inches. Continue subtracting in this fashion to determine the value in inches of each successive class-interval limit below the mean height.

This procedure can be summarized as follows:

Where the z interval equals .50, the range in inches of any z score interval for the height variable is equal to the difference between the height values

of $z = .25$ and $z = -.25$. The value of X when $z = .25$ has already been found to be 29.68. The value of X when $z = -.25$ is as follows:

$$X = -.25(1.14) + 29.4 = 29.12$$

Therefore, the difference between 29.68 and 29.12 gives the range in inches of any class interval whose range is equal to one-half a z score unit. This difference, .56 inch, can now be used as a constant amount to be added to and subtracted from the height measures corresponding to z scores of .25 and $-.25$. Thus, the values in inches of the successive class-interval limits *above* the mean are as follows:

1. Range value, in inches, of class intervals equal to .50 z score units is .56.
2. When $z = .25$, $X = 29.68$ inches
3. When $z = .75$, $X = 29.68 + .56 = 30.24$ inches
4. When $z = 1.25$, $X = 30.24 + .56 = 30.80$ "
5. When $z = 1.75$, $X = 30.80 + .56 = 31.36$ "
6. When $z = 2.25$, $X = 31.36 + .56 = 31.92$ "
7. When $z = 2.75$, $X = 31.92 + .56 = 32.48$ "
8. When $z = 3.25$, $X = 32.48 + .56 = 33.04$ "

Ordinarily, it is unnecessary to carry the limits beyond $z = +3.25$ because in most distributions values beyond this point do not occur. In the height data in Fig. 9:8 the maximum height is 32.50 inches and hence is included in the class interval whose upper limit is equal to a z score of 3.25.

The height values of the limits of the class intervals *below* the mean of the distribution of height scores are found by the same method, except that the constant value of .56 inch is *subtracted* from the height value of the successive limits. Thus,

9. When $z = -.25$, $X = 29.12$ inches
10. When $z = -.75$, $X = 29.12 - .56 = 28.56$ inches
11. When $z = -1.25$, $X = 28.56 - .56 = 28.00$ "
12. When $z = -1.75$, $X = 28.00 - .56 = 27.44$ "
13. When $z = -2.25$, $X = 27.44 - .56 = 26.88$ "
14. When $z = -2.75$, $X = 26.88 - .56 = 26.32$ "
15. When $z = -3.25$, $X = 26.32 - .56 = 25.76$ "

A similar procedure is used to convert the ordinate scale of *weights* into intervals whose limits will correspond to the intervals on the z score scale at the right of the chart in Fig. 9:10. Any value of Y in terms of z is equal to:

$$Y = z_y\sigma_y + M_y \qquad [9:4a]$$

Since $M_y = 21.8$ lbs., and $\sigma_y = 2.48$ lbs.,

$$Y = z_y(2.48) + 21.8$$

The original score values of the class-interval limits are indicated at the left of the chart.

With both scales of original measures converted into z score intervals, the final step in making a z score correlation chart consists in cross-tabulating the

original data (Table 9:1) into the chart shown in Fig. 9:10. The scales at the left and top of this chart are the reference scales for the tally. This figure shows the final result for the infants' heights and weights in terms of the number of correlational frequencies per cell, rather than the tally itself. The z score interval for any case can now be readily determined by referring to the scales at the right and bottom of the chart.

The Regression Line for z_y on z_x

Once the correlation frequencies of original data are reorganized into a z score correlation chart like that in Fig. 9:10, a regression line can be fitted to the data and from this line the product-moment correlation coefficient can be directly estimated. The procedure for estimating r consists in computing the average values of the z scores of one variable that are associated with the successive class-interval values of the other variable.* Such computations for the original measures were made in Fig. 9:9 for the variation of the original weight scores of the y variable with respect to the original height scores of the x variable. Therefore, the same relationship of y with respect to x will now be developed in terms of z scores.

The z score limits of each of the 13 class intervals of variables x and y have been selected in Fig. 9:10 so as to yield mid-point values that are convenient for arithmetical manipulation. Thus, the mid-point values of each class interval, beginning with the lower end of each scale, are as follows:

$$-3.0,\ -2.5,\ -2.0,\ -1.5,\ -1.0,\ -.50,\ 0,\ .50,\ 1.0,\ 1.5,\ 2.0,\ 2.5,\ 3.0$$

These, then, are the z score values to be used in working with the correlation frequencies of the z score correlation chart. Fig. 9:10 shows that there are four infants whose *height* measurements are within the interval whose mid-point value is -2.0, in other words, whose mid-point value is 2.0 standard deviation units below the mean of the height distribution. The *weight* of these four infants varies, when converted to z scores, from -0.5 to -2.0. The average of the z score variation of their weight is therefore computed as follows:

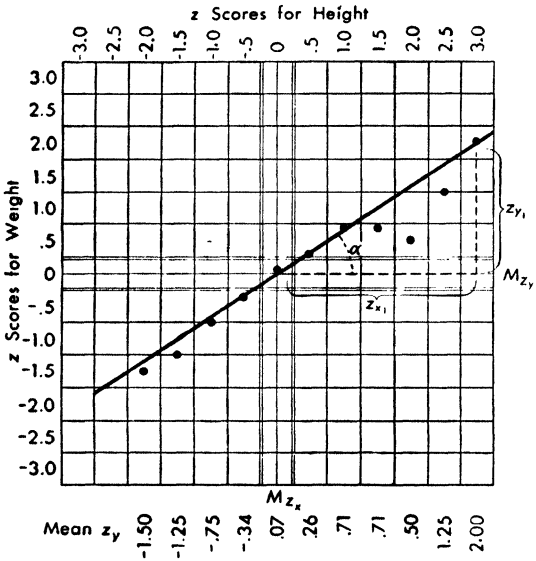
2 infants with weight equal to z scores of -2.0	$=-4.0$
1 infant with weight equal to z score of -1.5	$=-1.5$
1 infant with weight equal to z score of -0.5	$=-0.5$
Σ	$=-6.0$
Mean	$=-1.50$

The sum of the z score weights of these four infants is thus -6.0 and the average of these four variations is a z score of -1.50. Similarly, the mean of the z score variation in the weight of the infants with z scores of -1.5 for

* That these are average values is indicated by the use of the bar above the symbol for the dependent variable.

the height variable is found to be −1.25; the mean z score weight of those with a z score height of −1.0 is −.75. These mean values, as well as those for the remaining intervals of height, are indicated at the bottom of the correlation chart in Fig. 9:11.

Fig. 9:11. Estimation of Product-Moment r from a Straight-Line Function for $\bar{z}_y$ on z_x *



* From Data of Fig. 9:10.

Estimating r

The matrix in Fig. 9:11 is the same as the z score correlation chart in Fig. 9:10. The means of the variation in z score weights for the successive class intervals of height have been plotted and a straight line has been fitted, by inspection, to these means. The slope of this straight-line function is again (as in Fig. 9:9) equal to the tangent of the angle α made by this best-fitting straight line with the abscissa. Since the scales of the variables x and y have been made comparable in terms of z scores, we shall now symbolize the slope of the straight line by r. We find that r is equal to approximately the following ratio:

$$r_{yx} = \frac{z_{y_1}}{z_{x_1}} = \frac{2.0}{3.0} = .67$$

This value, .67, is the slope of the straight line fitted to the bi-variate data in Fig. 9:11. *It is the estimated value of* r, *the product-moment correlation coefficient.*

The Regression Equation of $\bar{z}_y$ on z_x

The product-moment correlation coefficient, r, is thus seen to be the slope of a straight-line function fitted to bi-variate data that have been converted into comparable deviate scales, each of which is taken in terms of the standard deviations of each variable. The regression equation by which the z score values of the y variable can be estimated from the z score values of the x variable, is as follows:

$$\bar{z}_y = r_{yx} z_x \qquad [9:5]$$

Regression equation of $\bar{z}_y$ on z_x

For the height-weight relationship in Fig. 9:11, this linear equation is equal to:

$$\bar{z}_y = .67z_x$$

In order to estimate actual weight scores from given height scores, it is necessary to express the preceding equation in terms of the original height and weight score values. If X symbolizes any original height measure and Y any original weight measure, we have

$$\frac{Y - M_y}{\sigma_y} = r_{yx}\frac{X - M_x}{\sigma_x} \qquad [9:5a]$$

Rearranging the terms of this linear equation in order to solve for Y from any given value of X, we have

$$\bar{Y} = \frac{r_{yx}\sigma_y(X - M_x)}{\sigma_x} + M_y \qquad [9:5b]$$

or, as usually expressed for convenience in computing,

$$\bar{Y} = r_{yx}\frac{\sigma_y}{\sigma_x}(X - M_x) + M_y \qquad [9:6]$$

Regression equation of $\bar{Y}$ on X (original score form)

Substituting the values already obtained for the means and standard deviations of each variable, and using the value of r estimated from the regression line in Fig. 9:11, we have:

$$\begin{aligned}\bar{Y} &= .67\frac{2.48}{1.14}(X - 29.4) + 21.8\\ &= 1.46(X - 29.4) + 21.8\\ &= 1.46X - 21.12\end{aligned}$$

The regression equation for the data in Fig. 9:11, from which the infants' weights can be estimated from any given height value, is thus

$$\bar{Y} = 1.46X - 21.12$$

where X is the height in inches and $\bar{Y}$ is the *average* estimate of weight in pounds. Infants with a height of 29 inches (the mid-point height value for the interval ranging from 28.75 to 29.25 inches) would, on the average, have the following weight in pounds:

$$\bar{Y} = 1.46(29.0) - 21.12 = 21.2 \text{ pounds}$$

Thus, on the basis of the estimate of correlation obtained from the straight line fitted to the height-weight data in Fig. 9:10, we would expect infants whose height was 29 inches to have an *average* weight of 21.2 pounds.

The Regression Equation in Descriptive Statistics *

From the point of view of descriptive statistics the preceding predictive estimate is unnecessary because the average weight of infants whose height was between 28.75 and 29.25 inches can be determined directly from the

* For its use in sampling and analytical statistics, see chap. 16.

original data. The basic value of the method of correlation for descriptive statistics is not the regression equation per se, from which values of one variable can be estimated from those of another, but rather the correlation coefficient as an index to *summarize* the degree of correlation between bi-variates.

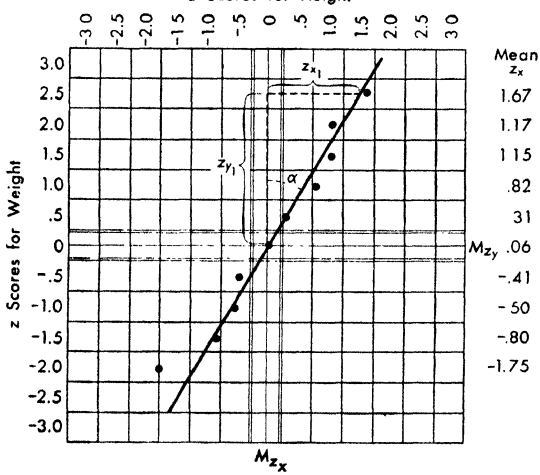
The purpose of introducing at this time the regression equation for product-moment r has been to emphasize the following points:

1. The fact that product-moment correlation is based on the assumption that a straight-line function will adequately describe the relationship between two variables.
2. The fact that r, the product-moment correlation coefficient, can be estimated by means of a graphic method.
3. The fact that r is in reality the slope of a straight line fitted to a bi-variate distribution set up in terms of z scores.
4. The fact that there are two regression equations for each bi-variate relationship.

The Regression of $\bar{z}_x$ on z_y

We shall now describe the basis for the second straight-line function and the corresponding regression equation for the height-weight data in Fig. 9:8. Having considered the development of the relation of z_y to z_x (the way in which weights vary with respect to successive class-interval height measurements), we shall discuss the relationship of z_x to z_y—in other words, how the height of the infants varies with respect to their weight measurements. The necessary correlational data for this relationship are already cross-tabulated in the z score correlation chart (Fig. 9:10). We shall use this z score matrix again to determine the means of the variations in height with respect to different weights. The results are presented in Fig. 9:12.

Fig. 9:12. Estimation of Product-Moment r from a Straight-Line Function for $\bar{z}_x$ on z_y *



* From Data of Fig. 9:10.

The correlational frequencies in Fig. 9:10 indicate that there were four

infants whose weight was in the z score interval ranging from -2.25 to -1.75. Two of these infants had z scores for height that also ranged from -2.25 to -1.75. The other two, however, had z scores for height ranging from -1.75 to -1.25. The mean z score for the height of these four infants is therefore -1.75. This mean is noted opposite the bottom row at the right of Fig. 9:12. Similarly, the next class interval of the weight z scores (-1.75 to -1.25) in Fig. 9:10 shows that there were 15 infants in this weight group and their z scores for height varied from -2.00 (mid-point value) to zero. The mean of their z scores, $-.80$, is likewise noted at the right of Fig. 9:12. The mean z scores for height for the remaining weight groups are also indicated at the right of this chart. Each mean is then plotted in the matrix in the chart, and a straight line is drawn by inspection to these means, giving the regression line of $\bar{z}_x$ on z_y. From the slope of this line, a second estimate of the correlation between z and y can be made. The coefficient, r, is equal to the tangent of the angle made by the regression line and the ordinate axis (α in Fig. 9:12). The tangent is therefore equal to the following ratio:

$$r_{xy} = \frac{z_{x_1}}{z_{y_1}} = \frac{1.625}{2.5} = .65$$

The estimated value of the product-moment correlation coefficient from the relation of the x variable to the y variable is thus .65.

The Regression Equation for $\bar{z}_x$ on z_y

The straight-line equation for the relationship of z_x to z_y is as follows:

$$\bar{z}_x = r_{xy}z_y = .65z_y \qquad [9:7]$$

Regression equation of $\bar{z}_x$ on z_y

This is the regression equation of x on y in terms of z scores. In order to convert it into a form from which values of X (height in inches) can be estimated from given values of Y (weight in pounds), we proceed as before; namely, we express the z score values in original scores as follows:

$$\frac{X - M_x}{\sigma_x} = r_{xy}\frac{Y - M_y}{\sigma_y} \qquad [9:7a]$$

From this we convert the expression for the solution of X for any value of Y:

$$\bar{X} = \frac{r_{xy}\sigma_x(Y - M_y)}{\sigma_y} + M_x \qquad [9:7b]$$

or, as usually expressed for convenience in computing,

$$\bar{X} = r_{xy}\frac{\sigma_x}{\sigma_y}(Y - M_y) + M_x \qquad [9:8]$$

Regression equation of $\bar{X}$ on Y (original score form)

Substituting in this regression equation the values obtained for the means and standard deviations of the height and weight variables and the value of r estimated from the straight line in Fig. 9:12, we find the equation of X on Y to be equal to:

$$\overline{X} = .65\frac{1.14}{2.48}(Y - 21.8) + 29.4$$

$$\overline{X} = 0.30(Y - 21.8) + 29.4$$

$$\overline{X} = .30Y + 22.86$$

This is therefore the regression equation for variations in the infants' height as associated with different values of weight. We would expect infants with a weight of, say, 26 pounds to have an *average* height of 30.7 inches:

$$\overline{X} = .30(26.0) + 22.86 = 30.7 \text{ inches}$$

The Regression Coefficients

From the preceding, we have seen that there are two regression equations that describe the co-relationship of two variables. One equation describes the variations of the y variable with respect to the x variable, and the other describes the variations of the x variable with respect to the y variable. Expressed in z score form, these equations were found to be as follows:

[9:5]

$$\bar{z}_y = r_{yx}z_x$$

Regression equation of $\bar{z}_y$ on z_x

[9:7]

$$\bar{z}_x = r_{xy}z_y$$

Regression equation of $\bar{z}_x$ on z_y

In these equations, each measurement of the two variables is expressed in terms of its distance from the mean of its respective distribution, taken in units of the standard deviation of the distribution. The correlation coefficients, r_{yx} and r_{xy}, of these two equations are identical with the *regression coefficient* and with each other.

Regression equations are, however, written more usually in terms of x and y deviation measures, or in terms of original measures. We saw that the respective regression equations of Y on X and X on Y were obtained by expressing the z score equation in terms of original measures:

[9:6]

$$\overline{Y} = r_{yx}\frac{\sigma_y}{\sigma_x}(X - M_x) + M_y$$

Regression equation of Y on X

[9:8]

$$\overline{X} = r_{xy}\frac{\sigma_x}{\sigma_y}(Y - M_y) + M_x$$

Regression equation of $\overline{X}$ on Y

From these equations the average values of Y associated with X, or of X associated with Y, can be estimated. It should be observed, however, that

the regression coefficients in Formulas 9:6 and 9:8 are not equal in value to the correlation coefficient itself; rather, the regression coefficient of the first equation is $r_{yx}\frac{\sigma_y}{\sigma_x}$, and the regression coefficient of the second is $r_{xy}\frac{\sigma_x}{\sigma_y}$. These regression coefficients are usually symbolized by b, as follows:

$$b_{yx} = r_{yx}\frac{\sigma_y}{\sigma_x} \qquad [9:9]$$

Regression coefficient of $\bar{y}$ on x

$$b_{xy} = r_{xy}\frac{\sigma_x}{\sigma_y} \qquad [9:10]$$

Regression coefficient of $\bar{x}$ on y

We have seen why m alone is not a suitable estimate of the slope of the best-fitting straight line for bi-variate data which are not reduced to comparable deviation scales (in terms of their respective standard deviations). We now see that an appropriate mathematical adjustment can be made in determining the regression coefficients by means of the ratio of the standard deviations of the respective distributions. Only if this adjustment has already been made by converting original measures into z scores will r, the correlation coefficient, also be equal to the regression coefficient. When this conversion has been made, the regression coefficient is symbolized by the Greek letter beta:

$$\beta_{yx} = r_{yx} \qquad [9:11]$$

Regression coefficient of $\bar{z}_y$ on z_x

$$\beta_{xy} = r_{xy} \qquad [9:12]$$

Regression coefficient of $\bar{z}_x$ on z_y

Regression Equations Expressed in Terms of x and y

Regression equations are often expressed in terms of the deviations x and y, where $x = X - M_x$ and $y = Y - M_y$. Expressing the regression equations in Formulas 9:5 and 9:7 in terms of x and y gives the following, inasmuch as $z_y = \frac{y}{\sigma_y}$, and $z_x = \frac{x}{\sigma_x}$:

$$\frac{y}{\sigma_y} = r_{yx}\frac{x}{\sigma_x}$$

or, for the solution of $\bar{y}$:

$$\bar{y} = r_{yx}\sigma_y\frac{x}{\sigma_x} = r_{yx}\frac{\sigma_y}{\sigma_x}x \qquad [9:13]$$

Regression equation of $\bar{y}$ on x

This, then, is the regression equation of $\bar{y}$ on x, the regression coefficient being the same as in the equation in Formula 9:6.

And

$$\frac{x}{\sigma_x} = r_{xy}\frac{y}{\sigma_y}$$

or, for the solution of $\bar{x}$:

$$\bar{x} = r_{xy}\sigma_x\frac{y}{\sigma_y} = r_{xy}\frac{\sigma_x}{\sigma_y}y \qquad [9:14]$$

Regression equation of $\bar{x}$ on y

Again the regression coefficient in the above formula is the same as that in Formula 9:8.

Standard Formula for r

The last set of regression equations, Formulas 9:13 and 9:14, is expressed in terms of the deviations x and y. The standard formulation of the product-moment correlation coefficient is also expressed in terms of these same terms, as indicated earlier in this chapter.

$$r_{xy} = \frac{\frac{\Sigma(xy)}{N}}{\sigma_x\sigma_y} = \frac{\Sigma(xy)}{N\sigma_x\sigma_y} \qquad [9:1]$$

Product-moment r

where $\Sigma(xy)$ is the algebraic sum of the products of the deviation values, x and y, obtained for each pair of associated measures, N is the number of paired frequencies, and σ_x and σ_y are the standard deviations of the variables correlated. The correlation coefficient, r_{xy}, is thus formulated as the ratio of the mean of the product deviations, $\left[\frac{\Sigma(xy)}{N}\right]$, to the product of the standard deviations of the two variables.

In order for the preceding formula to be used in computing the product-moment correlation coefficient, each original pair of measures being correlated must be converted into x and y deviations, either at the beginning or toward the end of the computations. We shall see in the next section that in the short method for computing r this conversion is made toward the end. That is, instead of each original measure being converted into its respective x and y value, the initial computations are made from the original measures themselves and the value of the numerator, $\Sigma(xy)$, is obtained as one of the final steps in the computation (see Fig. 9:13).

The correlation coefficient can also be computed from the z score measures of a bi-variate distribution. In this case, r is the mean of the algebraic sum of the products of each associated pair of z scores:

$$r_{xy} = \frac{\Sigma(z_x z_y)}{N} \qquad [9:15]$$

r from z scores

These two formulas for r symbolize what is basically involved in computing the product-moment correlation coefficient. They represent two processes that can be used to obtain the slope of the straight line which best fits the bi-variate data. Any method for computing r is essentially a mathematical

procedure for obtaining the best-fitting straight line by the method of *least squares*, a method which yields a mathematical result such that the errors of fit are at a minimum. Consequently, from the slope of such a straight line a precise value for r_{xy} can be obtained.

In *estimating* the value of r from regression lines fitted to bi-variate data, we have seen that a determination can be made from the regression either of $\bar{y}$ on x, or of $\bar{x}$ on y. Inasmuch as these lines are fitted by inspection, they cannot be expected to yield exactly the same values for r. However, if they are fitted carefully, the difference between the two values of r should not exceed .05. On the other hand, in *computing* the value of r, only one coefficient is obtained. It can be regarded as the correlation either between x and y, or between y and x. When the errors of fitting regression lines to bi-variate data are reduced to a minimum, as is the case when r is *computed*, the slopes of both lines are identical, and therefore $r_{yx} = r_{xy}$. In practice, the subscripts for r are usually written as xy, the notation yx being discarded.

C. COMPUTATION OF PRODUCT-MOMENT r

Summary of Mathematical Implications of r

In the preceding section we described a method for *estimating* Pearson's product-moment correlation coefficient from the cross-tabulated data of two variables. We saw that the estimating process is not particularly difficult, once its implications are clear. Before proceeding with methods for *computing* r, it will be relevant to review the meaning of the correlation coefficient from a mathematical point of view.

When the data of each variable have been converted to comparable z score scales, r is the slope of two regression lines (linear functions) fitted to the variations of y with respect to x and to the variations of x with respect to y. When the correlation coefficient is *estimated* from the cross-tabulated data of a z score correlation chart, two values for the correlation coefficient are obtained, one from the slope of the regression of $\bar{z}_y$ on z_x and the other from the slope of the regression of $\bar{z}_x$ on z_y. These two values need not be identical, because the fitting of regression lines by inspection cannot yield a precise result. On the other hand, when the product-moment correlation coefficient is *computed*, only one coefficient is obtained. It is a mathematical average of the respective slopes of $\bar{z}_y$ on z_x and $\bar{z}_x$ on z_y. Furthermore, in effect, it is computed by the mathematical method of least squares. This means that the straight-line functions are fitted (algebraically, not graphically) to the bi-variate data in such a way as to reduce the errors of fit to a minimum.* This

* For the mathematical nature of the method of least squares, see M. Phillip, *The Principles of Finance and Statistical Mathematics*, Prentice-Hall, New York, rev. ed., 1941, chap. 14; and C. C. Peters and W. R. Van Voorhis, *Statistical Procedures and Their Mathematical Basis*, McGraw-Hill, New York, 1940, p. 299.

is why the computed value of r provides a more accurate index of correlation than does the estimate from a straight-line function fitted by inspection.

Various Methods for the Computation of r

The computation of r involves only two steps in addition to those described in Chapter 7 for computing the mean and standard deviation of a variable. The first of these two steps has already been described in the present chapter—the now familiar cross-tabulation, in a correlation chart, of the data of the two variables to be correlated. However, this step is often carried out only algebraically and not by means of a correlation chart. (Cf. Table 9:2.) The second step consists in computing the *products of the deviations* for each pair of associated measures in the bivariate distribution, and then summing and averaging these products.

We have seen (Formula 9:1) that the product-moment correlation coefficient is algebraically equal to the ratio of the mean of the product deviations to the product of the standard deviations of the two variables being correlated. Many methods of computation have been devised for obtaining this ratio. In this chapter we shall describe the following methods, which are among the most commonly used procedures:

I. Product-moment r computed from ungrouped data (long method).
II. Product-moment r computed from grouped data (short method).
III. Product-moment r computed from ungrouped data (machine method).

The relative advantages and disadvantages of these three methods will be apparent as we proceed.

D. METHOD I: PRODUCT-MOMENT r FROM UNGROUPED DATA (LONG METHOD)

In the long method for ungrouped data, the mean and standard deviation of each variable are computed and the original measures are each expressed as deviations (x and y) from their respective means. The products of the deviation values of each pair of measures are then obtained, and all the products for the bi-variate distribution are summed to obtain $\Sigma(xy)$. With these computations made, the correlation coefficient is readily obtained, being equal to the ratio of the mean of the product deviations to the product of the standard deviations of the two distributions. Thus,

$$r_{xy} = \frac{\frac{\Sigma(xy)}{N}}{\sigma_x \sigma_y} = \frac{\Sigma(xy)}{N\sigma_x\sigma_y} \qquad [9:1]$$

Pearson's product-moment correlation coefficient

If the σ's in this formula are expressed in deviation terms, the computations can be simplified by canceling out the N's:

$$r_{xy} = \frac{\Sigma(xy)}{N\sqrt{\frac{\Sigma x^2}{N}}\sqrt{\frac{\Sigma y^2}{N}}} = \frac{\Sigma(xy)}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}} \qquad [9:16]$$

Pearson r (alternate form for computation)

where Σx^2 represents the sum of the squared deviations of the x variable, and Σy^2 represents the sum of the squared deviations of the y variable.

Order of Operations for Method I

Method I is illustrated in Table 9:2. For purposes of simplification, the data of only 20 cases have been used. The paired associates in this table represent the scores made by 20 persons on two separate administrations of a digit-span test.* The scores made on the first administration are given as the x variable and represent the average result of four trials. The average scores received by each person on the second administration of the test are given as the y variable, and likewise represent the average of four trials.

The order of computation in Table 9:2 is as follows. (The similarity to the procedure already developed for computing the mean and standard deviation from ungrouped data will be apparent.)

1. The data of each of the variables are arranged by associated pairs in two adjacent columns (columns 2 and 3 of the table). In the case of these data, the basis for each associated pair is the subject taking both administrations of the test.
2. The mean of each variable is next obtained by summing the scores of each variable and dividing by N, the number of measures.
3. The deviations of the measures of each variable from their respective means are obtained and entered in adjacent columns (columns 4 and 5), to yield the deviations x and y, where $x = X - M_x$, and $y = Y - M_y$. Care is necessary in differentiating the positive and negative deviations.
4. The deviations of each variable are squared (columns 6 and 7) to give x^2 and y^2 for computing the standard deviations.
5. The products of each associated pair of x and y deviations are obtained to give the product deviations (column 8). The signs of the deviations must be carefully noted in computing these products so that the correct sign will be entered in the column.
6. The product deviations of the last column are summed algebraically and then averaged (divided by N) to give the mean of the product deviations. This is the necessary value for the numerator of r in Formula 9:12. The ratio of this value to the product of the standard deviations of each variable gives the correlation coefficient, r.

* These data are from J. G. Peatman and N. M. Locke, "Studies in the Methodology of the Digit-Span Test," *Archives of Psychology*, Monograph No. 167, 1934. The correlation coefficient obtained here between the digit-span scores of two separate administrations of the test constitutes an index of the reliability of the test, by the method of test-retest. (Cf. chap. 17, Section B.)

Table 9:2. Computation of the Product-Moment Correlation Coefficient (r)—Long Method with Ungrouped Data

(Correlation of Digit-Span Test Scores for 20 College Students: Test and Retest) *

(1) Subjects	(2) Digit-Span Scores 1st Test X	(3) Digit-Span Scores Retest Y	(4) Deviations from Mean of x x	(5) Deviations from Mean of y y	(6) Deviations Squared x^2	(7) Deviations Squared y^2	(8) Product Deviations xy
A	8.5	8.25	1.05	.60	1.1025	.3600	.6300
B	6.75	8.25	−.70	.60	.4900	.3600	−.4200
C	6.25	7.75	−1.20	.10	1.4400	.0100	−.1200
D	6.0	5.5	−1.45	−2.15	2.1025	4.6225	3.1175
E	7.0	7.25	−.45	−.40	.2025	.1600	.1800
F	9.25	9.25	1.80	1.60	3.2400	2.5600	2.8800
G	5.5	5.75	−1.95	−1.90	3.8025	3.6100	3.7050
H	9.25	8.25	1.80	.60	3.2400	.3600	1.0800
I	7.75	6.75	.30	−.90	.0900	.8100	−.2700
J	5.0	5.25	−2.45	−2.40	6.0025	5.7600	5.8800
K	6.5	7.25	−.95	−.40	.9025	.1600	.3800
L	7.75	8.5	.30	.85	.0900	.7225	.2550
M	9.0	8.0	1.55	.35	2.4025	.1225	.5425
N	7.5	8.25	.05	.60	.0025	.3600	.0300
O	7.0	7.75	−.45	.10	.2025	.0100	−.0450
P	7.75	8.25	.30	.60	.0900	.3600	.1800
Q	5.75	5.75	−1.70	−1.90	2.8900	3.6100	3.2300
R	8.75	9.5	1.30	1.85	1.6900	3.4225	2.4050
S	8.75	8.50	1.30	.85	1.6900	.7225	1.1050
T	9.0	9.00	1.55	1.35	2.4025	1.8225	2.0925
N = 20	149.0 (ΣX)	153.0 (ΣY)	0 (Check sum)	0 (Check sum)	34.0750 (Σx^2)	29.9250 (Σy^2)	27.6925 −.8550 26.8375 (Σxy)

$$\text{Mean}_x = \frac{149.0}{20} = 7.45 \qquad \text{Mean}_y = \frac{153.0}{20} = 7.65$$

$$\sigma_x = \sqrt{\frac{34.0750}{20}} = 1.305 \qquad \sigma_y = \sqrt{\frac{29.9250}{20}} = 1.223$$

$$\text{Mean of Product Deviations} = \frac{26.8375}{20} = 1.3419$$

$$\text{Correlation Coefficient } r = \frac{1.3419}{(1.305)(1.223)} = \frac{1.3419}{1.5960} = .84$$

* These 20 cases were drawn randomly from the original group of 142 subjects. The coefficient for the total group was .86, as compared with .84 for these 20 cases.

In the example in Table 9:2, the mean of the product deviations is 1.342; the product of the standard deviations of the two variables is 1.596; the ratio of 1.342 to 1.596 is .84; and r is therefore .84. There is consequently a marked positive relationship in digit-span performance on the two administrations of the test for these 20 subjects.

Shortcomings of Method I

This method of computing r has two shortcomings. (1) The amount of arithmetical work is unnecessarily laborious, as was the case in computing the standard deviation by the long method. (2) Of real importance in the case of unfamiliar data, the form of the bi-variate distribution cannot readily be seen unless a scattergram or correlation chart of the data is made. This is likely to prove a serious handicap at times, inasmuch as the basic assumption in the method of product-moment correlation is that the relationship between two variables is linear. Unless the data are cross-tabulated, the investigator cannot be sure that a straight-line function will be appropriate for the bi-variate data. Hence, the long method for ungrouped data should be avoided except in the case of familiar variables when one can be confident that a linear function will satisfactorily describe the relationship.

E. METHOD II: PRODUCT–MOMENT r FROM GROUPED DATA (SHORT METHOD)

Although a long method can be used for computing r from grouped data cross-tabulated in a correlation chart, it is so unnecessarily laborious that we shall omit it here and describe only the short method. The difference between the two is analogous to the difference in the two methods for computing the standard deviation described in Tables 7:8 and 7:10. Most of the steps in the short method have already been developed in computing the mean and the standard deviation (Table 7:10); in fact, aside from making an original cross-tabulation of the bi-variate data, only two additional steps are involved. They are described in Fig. 9:13, which illustrates Method II, and they involve the computation of $\Sigma x'$, $\Sigma y'$, and $\Sigma x'y'$. The correlation chart in Fig. 9:13 will be recognized as the same as that in Fig. 9:8, the cross-tabulation of the weights and heights of 151 girl infants.

The formula for the computation of r by Method II is as follows:

$$r_{xy} = \frac{\frac{\Sigma(fx'y')}{N} - \left(\frac{\Sigma fx'}{N}\right)\left(\frac{\Sigma fy'}{N}\right)}{\sqrt{\frac{\Sigma(fx'^2)}{N} - \left(\frac{\Sigma fx'}{N}\right)^2}\sqrt{\frac{\Sigma(fy'^2)}{N} - \left(\frac{\Sigma fy'}{N}\right)^2}} \qquad [9:17]$$

Pearson's r by short method with deviations from guessed means

This formula may also be written with the symbol c representing the operations already done in computing the mean and standard deviation by the short method for grouped data:

$$r_{xy} = \frac{\frac{\Sigma(fx'y')}{N} - c_x c_y}{\sqrt{\frac{\Sigma(fx'^2)}{N} - c_x^2}\sqrt{\frac{\Sigma(fy'^2)}{N} - c_y^2}} \qquad [9:17a]$$

In this form the numerator gives the mean of the product deviations from the actual means of both variables, and the denominator gives the standard deviations of each in *unit* interval terms. To obtain the standard deviations in original score terms, each root value in the denominator must be multiplied by i, the size of the respective class intervals.

The Frequency Distributions of Each Variable from the Correlation Chart

The frequencies of the class intervals of each variable can be readily obtained from the cross-tabulated data. The summation of all the frequencies in each row of the chart in Fig. 9:13 gives the frequencies for each class interval of the y variable. Similarly, the summation of all the frequencies in each column gives the frequencies for each class interval of the x variable. These frequency summations are entered at the right and the bottom of the chart; thus the frequency distribution for the weight (y) variable is in the first column at the right, and the frequency distribution for the height (x) variable is in the first row at the bottom. The summation of each of these frequency distributions gives N, the total number of associated pairs or frequencies. The sum of the frequencies of each variable should, of course, be made independently so as to provide a check for the value of N. The value of N is entered in the square at the lower right-hand corner, just outside the correlation matrix.

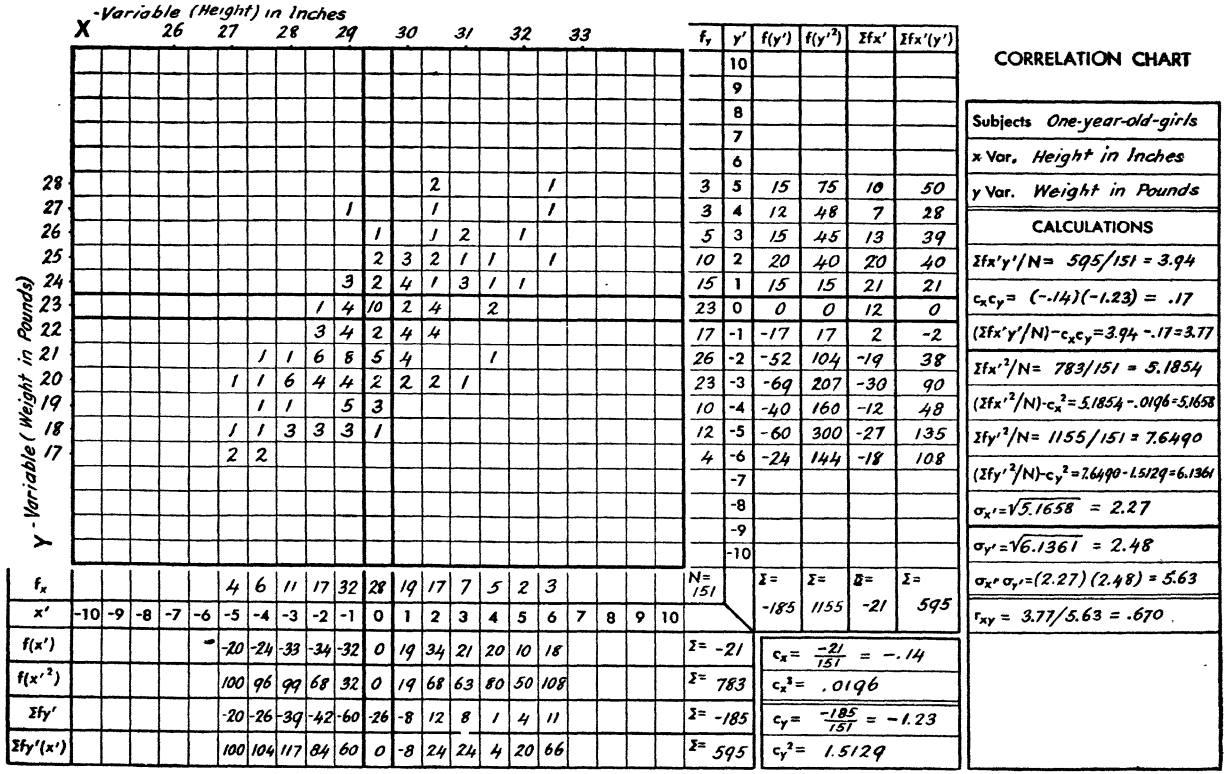
The Standard Deviations of Each Variable from the Correlation Chart

The frequency distributions of each variable having been obtained, the next step is to compute their respective standard deviations. The procedure is the same as that used in determining standard deviations by the short method from grouped data (Table 7:10). Appropriate columns and rows are provided at the right of and below the matrix of the correlation chart for entering the relevant computations. The y variable will be computed first, and then the x variable.

The unit interval deviations (y') from the guessed mean are entered in the column headed y' at the right of the correlation matrix. The products of the frequencies and the unit interval deviations from the guessed mean are next computed and entered in the column headed $f(y')$. The sum of these products is necessary for the computation of c_y, the correction factor for the deviations from the guessed mean. As we saw earlier,

$$c_y = \Sigma(fy')/N$$

The products of the frequencies and the squared deviations from the guessed mean are entered in the column headed $f(y'^2)$. The sum of the products of this column gives the total of the squared deviations from the guessed mean. All



CORRELATION CHART

Subjects	*One-year-old-girls*
x Var.	*Height in Inches*
y Var.	*Weight in Pounds*
CALCULATIONS	
$\Sigma fx'y'/N$ =	$595/151 = 3.94$
$c_x c_y$ =	$(-.14)(-1.23) = .17$
$(\Sigma fx'y'/N) - c_x c_y$ =	$3.94 - .17 = 3.77$
$\Sigma fx'^2/N$ =	$783/151 = 5.1854$
$(\Sigma fx'^2/N) - c_x^2$ =	$5.1854 - .0196 = 5.1658$
$\Sigma fy'^2/N$ =	$1155/151 = 7.6490$
$(\Sigma fy'^2/N) - c_y^2$ =	$7.6490 - 1.5129 = 6.1361$
$\sigma_{x'}$ =	$\sqrt{5.1658} = 2.27$
$\sigma_{y'}$ =	$\sqrt{6.1361} = 2.48$
$\sigma_{x'}\sigma_{y'}$ =	$(2.27)(2.48) = 5.63$
r_{xy} =	$3.77/5.63 = .670$

Fig. 9:13. Computation of the Product-Moment Correlation Coefficient (*r*), Short Method—Grouped Data Cross-Tabulated in Correlation Chart. Height (*x*) and Weight (*y*) Measures of 151 One-Year-Old Girls, from Data of Table 9:1.

the data necessary for computing the standard deviation of the y variable are now available.*

The sum of the frequencies times the unit interval deviations from the guessed mean ($\Sigma fy'$) is equal to -185, and N is 151. The correction, c_y, is therefore -1.23:

$$c_y = \frac{-185}{151} = -1.23$$

The sum of the squared deviations [$\Sigma f(y'^2)$] from the guessed mean is 1155, and the standard deviation for weight is 2.48 pounds:

$$\sigma_y = i_y \sqrt{\frac{\Sigma f(y'^2)}{N} - c_y^2}$$

$$\sigma_y = 1.0\sqrt{\tfrac{1155}{151} - (-1.23)^2} = 2.48$$

The size of the class interval, i, is equal to 1.0.

The same procedure is used for computing the standard deviation of the x variable (height), and is indicated at the bottom of the chart. The unit deviations (x') of each class interval from the guessed mean are entered in the second row below the correlation matrix. The products of the frequencies for each class interval and their respective unit deviations, $f(x')$, are entered in the third row. The sum of these products divided by the number of cases gives the correction, c_x:

$$c_x = \frac{-21}{151} = -.14$$

The products of the frequencies and the squared deviations, $f(x'^2)$, are entered in the fourth row. The standard deviation for height is found to be 1.14 inches:

$$\sigma_x = i_x \sqrt{\frac{\Sigma f(x'^2)}{N} - c_x^2}$$

$$\sigma_x = 0.5\sqrt{\tfrac{783}{151} - (-.14)^2} = 0.5(2.27) = 1.135$$

where 0.5 is equal to i, the size of the class interval for this particular variable.

As already indicated, r, the product-moment correlation coefficient, is the ratio of the average of the product deviations to the product of the standard deviations. We now have the computations necessary for determining the denominator of this ratio, namely, the product of the standard deviations:

$$\sigma_x \sigma_y = (1.135)(2.48) = 2.81$$

The computation of the ratio necessary to express the value of r can be simplified by omitting i_x and i_y (the size of the class intervals) from the final computations. They can be omitted because they will cancel out algebraically from both the numerator and the denominator of the ratio for r. The product of the standard deviations that will be used is as follows:

$$\sigma_{x'} \sigma_{y'} = (2.27)(2.48) = 5.63$$

* An additional column may be added at the right of the chart in order to apply Charlier's check for the sums needed in computing σ (cf. Table 7:10).

The prime signs are used with the x and y subscripts to indicate that the product of the standard deviations has been obtained for unit intervals of deviation, rather than for actual intervals of the original variables.

The Product Deviations

The next step is to make the necessary computations for the numerator of the ratio for r. In other words, we need to obtain the product of the deviations for each pair of associated measures, sum these products, and calculate their mean:

$$\text{Mean of product deviations} = \frac{\Sigma xy}{N}$$

There are several methods of obtaining the product deviations from the cross-tabulations of a correlation chart. The method shown in Fig. 9:13 is in general use and provides not only an independent check of the sum of the product deviations, but at the same time a check for each of the correction factors, c_x and c_y. These values are obtained from the sum of the last two columns at the right of the matrix and of the last two rows at the bottom of the matrix. The computations required for the product deviations developed from the columns at the right will be described first.

The next to the last column, headed x', gives for all the cases in each class interval (row) of the y (weight) variable the sum of the deviations of the other variable (x) from the guessed mean of x. Thus, according to Fig. 9:13, in the highest class interval of the weight variable, y, there are three cases with an average weight of 28 pounds. Two of them are two unit intervals above (*to the right of*, on the chart) the guessed mean of the height variable, x, and the third case is six unit intervals above. Hence, the sum of the deviations ($\Sigma x'$) of these three cases from the guessed mean of the x variable is equal to

$$2(2) + 1(6) = 10$$

This figure is entered in the next to the last column at the right of the chart.

The next class interval of the y variable, with a mid-point of 27 pounds, has three cases. The first is one unit interval below (*to the left of*, on the chart) the guessed mean of x and therefore has an x' value of -1. The second is two unit intervals, and the third is six unit intervals, above the guessed mean of the x variable. The algebraic sum of the deviations of these three cases is equal to:

$$-1 + 2 + 6 = 7$$

The cases in each of the remaining class intervals of the y variable (rows) are in turn *summed* with respect to their unit interval deviations from the guessed mean of the other variable, x, and these sums are entered in the next to the last column at the right. The sum of the entries in this column gives the sum of all the deviations in the distribution from the guessed mean of x, and

should be equal to the sum already obtained in the usual manner (third row below the matrix in Fig. 9:13). In other words,

$$\Sigma fx' = \Sigma f(x')$$

The computation of the product deviations is now simple since we have the sum of the deviations from the guessed mean of x for each of the class intervals of y. The products, $x'y'$, for each class interval of y are readily obtained by multiplying y' and $\Sigma fx'$, because, algebraicallly,

$$y'(\Sigma fx') = \Sigma f(x'y')$$

for all frequencies in each row. These products are entered in the last column at the right of the chart. The sum of all these product deviations gives the total of the product deviations for the two variables $\Sigma f(x'y')$. For the height-weight data of the 151 infants, this is equal to 595. The average of this, 595/151, is 3.94, and is the mean of the product deviations from the guessed means of each variable.

It is now necessary to correct for the fact that these product deviations were obtained from the guessed means rather than from the actual means This correction is the product of the correction factors, c_x and c_y; in other words,

$$\left(\frac{\Sigma fx'}{N}\right)\left(\frac{\Sigma fy'}{N}\right)$$

It has already been pointed out that the final computation of r can be simplified by omitting the original size of the class intervals of each variable, viz., i_x and i_y, from the ratio. Hence, the average of the product deviations from the actual means of each variable, the class intervals being expressed as unit deviations from their means, is:

$$\frac{\Sigma f(x'y')}{N} - c_x c_y = \frac{595}{151} - (-.14)(-1.23)$$
$$= 3.94 - .17 = 3.77$$

Ratio for r

The correlation coefficient, r, is now readily computed, since it is the ratio of the mean of the product-deviations to the product of the standard deviations of each variable. For the data in Fig. 9:13, r is equal to .670:

$$r_{xy} = \frac{\text{Mean of product deviations}}{\text{Product of standard deviations}}$$
$$= \frac{3.77}{5.63} = .670, \text{ or } .67$$

Checking $\Sigma(x'y')$

Ordinarily, before computing the ratio for the correlation coefficient, it is wise to check the calculations of the product deviations. Such a check can be made independently by computing the sum of the product deviations

with respect to the class intervals of the x variable as well as of the y variable. These computations are given in the last two rows below the correlation matrix in Fig. 9:13. This time the sum of the deviations from the guessed mean of the y variable is obtained for each class interval of the x variable. Thus, the first column of data in the correlation chart shows four cases for the class interval of height that has a mid-point of 27 inches. All four cases are below the guessed mean of the other variable, y: the first, three unit intervals below; the second, five intervals below; and the remaining two, six intervals below. The sum of these four deviations is

$$-3 + (-5) + 2(-6) = -20$$

This figure for the column is entered in the fifth row below the chart. The deviations from the guessed mean of the y variable for the cases in each of the other eleven class interval columns of the x variable are obtained similarly and entered in their respective columns in the fifth row below the chart. The grand total of all these sums is -185, a value that checks with that already obtained in the $f(y')$ column at the right of the chart.

The sum of the product deviations is now checked by multiplying the $\Sigma y'$ values of each column by x'. These products are entered in the last row at the bottom of the chart, and their grand total is obtained to give $\Sigma(x'y')$. This total, 595, confirms the correctness of the sum of the product deviations already obtained in the last column at the right of the chart.

Means and Standard Deviations from the Correlation Chart

It has already been observed that the method just used for calculating the product-moment correlation coefficient does not directly employ the actual means and standard deviations of the original distributions of the two variables being correlated. These two measures, which are ordinarily needed for descriptive and analytical purposes, are obtained by the short method for grouped data already described in Chapter 7. Thus, for the data in Fig. 9:13,

$$\text{Mean}_x = G.M._x + i_x\left(\frac{\Sigma fx'}{N}\right)$$

$$\text{Mean height} = 29.5 + (0.5)(-.14)$$
$$= 29.4 \text{ inches } (x \text{ variable})$$

$$\text{Mean}_y = G.M._y + i_y\left(\frac{\Sigma fy'}{N}\right)$$

$$\text{Mean weight} = 23.0 + 1.0(-1.23)$$
$$= 21.8 \text{ pounds } (y \text{ variable})$$

$$\text{Standard deviation of height} = i_x(\sigma_{x'})$$
$$\sigma_x = 0.5(2.27) = 1.14 \text{ inches}$$

$$\text{Standard deviation of weight} = i_y(\sigma_{y'})$$
$$\sigma_y = 1.0(2.48) = 2.48 \text{ pounds}$$

F. METHOD III: PRODUCT-MOMENT r FROM UNGROUPED DATA (MACHINE METHOD)

Machine Computation

The method for computing the product-moment correlation coefficient to be described in this section has an advantage over both Methods I and II in that the *original measures* can be directly employed. No conversion to deviation values, x and y, is necessary. The underlying principle is similar to that of Method II, for all computations are made from guessed means and then *corrected*. However, as implied by the title of this method, the original data are not cross-tabulated into a correlation chart; hence Method III may at times present the same disadvantage as Method I, viz., the *form* which a bi-variate distribution takes cannot be seen unless a scattergram of the data is constructed.

Method III is particularly valuable when machines are available for computations and all the inter-correlations of a number of variables are required. In the machine procedure, the original data are usually punched on cards (cf. the I.B.M. cards, Chapter 2), one card being used for each subject. After the data are punched, the cards are fed into a machine that is set to make the necessary multiplications and additions. The totals needed for the computation of r by Formula 9:18 are obtained direct from the machine totals and the remaining calculations are facilitated by an electric calculator.

Method III often proves valuable even though punch cards and machines are not available. This is the case provided the total number of paired frequencies, N, is not greatly in excess of 100, and provided an adding or calculating machine and tables of squares and square roots (see Table I, Appendix C) are available.

This method also has the important advantage of embodying a systematic set of *checks* for all computations.* These checks are particularly necessary to determine the accuracy of the method because the multiplication and addition of many numbers are involved.

The Guessed Means Taken as Equal to Zero

In Method III a short cut is employed, in that deviations are taken from guessed means. However, in contrast to Method II, the guessed means are taken as equal to zero, as was done for the computation of M and σ in Method III in Chapter 7. Each original measure thus becomes a deviation value from a guessed mean. Each X and Y become x' and y' respectively, although the actual values of the original scores are not changed. All com-

* The system of checks used in connection with Method III has been adapted from Clark Hull, *Aptitude Testing*, World Book Co., Yonkers, 1928, pp. 427–439.

putations are made from the original measures. Since this method utilizes the short-cut device of guessed means taken as equal to zero, the final computations for r are analogous to those with Formula 9:17 in Method II.

If any of the original measures of a bi-variate distribution are negative numbers, the original data must be modified before the guessed means are taken as zero. This is done by adding to each original measure a constant amount of sufficient size to convert all the original measures into positive numbers. After this conversion is made, the method proceeds in the manner to be described. The only additional correction necessary, because of the use of converted values, arises in computing the final value of the mean of the distribution. The standard deviation will not be affected since the addition of a constant amount to each measure of a distribution leaves the value of σ unchanged.

The Formula for r (Method III)

The formula for computing r by Method III may be expressed in terms similar to those used in Formula 9:17:

$$r_{xy} = \frac{\frac{\Sigma(x'y')}{N} - \left(\frac{\Sigma x'}{N}\right)\left(\frac{\Sigma y'}{N}\right)}{\sqrt{\frac{\Sigma(x'^2)}{N} - \left(\frac{\Sigma x'}{N}\right)^2}\sqrt{\frac{\Sigma(y'^2)}{N} - \left(\frac{\Sigma y'}{N}\right)^2}} \qquad [9:17b]$$

The above formula is identical with Formula 9:17, except for the omission of the f's, which denote the multiplication of the deviation values of each class interval by the number of frequencies. Since the data are not grouped into class intervals in Method III, this multiplication is unnecessary and hence the f's do not appear in the formula.

When the guessed means of distributions of positive numbers are taken as equal to zero,

$$x' = X, \quad \text{and} \quad y' = Y$$

In other words, each original measure, its numerical value remaining unchanged, becomes a deviation from the guessed mean:

$$x' = X - G.M._x \quad \text{and} \quad y' = Y - G.M._y$$

If X and Y are substituted for x' and y' in Formula 9:17b, we have:

$$r_{xy} = \frac{\frac{\Sigma(XY)}{N} - \left(\frac{\Sigma X}{N}\right)\left(\frac{\Sigma Y}{N}\right)}{\sqrt{\frac{\Sigma(X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2}\sqrt{\frac{\Sigma(Y^2)}{N} - \left(\frac{\Sigma Y}{N}\right)^2}} \qquad [9:18]$$

Pearson's r: alternate form of 9:17 for ungrouped data with guessed means equal to zero

Since $\frac{\Sigma X}{N}$ yields the mean of the x variable, and $\frac{\Sigma Y}{N}$ yields the mean of the y variable, Formula 9:18 may be rewritten as follows:

$$r_{xy} = \frac{\frac{\Sigma(XY)}{N} - M_x M_y}{\sqrt{\frac{\Sigma(X^2)}{N} - (M_x)^2}\sqrt{\frac{\Sigma(Y^2)}{N} - (M_y)^2}} \qquad [9:18a]$$

But inasmuch as any sum divided by N, the number of measures yielding the sum, is a mean measure, all the terms of the above formula can be expressed as means:

$$r_{xy} = \frac{M_{XY} - M_x M_y}{\sqrt{M_{X^2} - (M_x)^2}\sqrt{M_{Y^2} - (M_y)^2}} \qquad [9:18b]$$

where M_{XY} is the mean of the products of the paired original measures of x and y, M_{X^2} is the mean of the squares of the original measures of the x variable, and M_{Y^2} is the mean of the squares of the original measures of the y variable. M_x and M_y are, as usual, the means of the measures of each distribution.

Inter-Correlation Coefficients

It has already been pointed out that Method III is labor-saving whenever inter-correlations between three or more variables are required. This is true because the sums of the original measures of each variable must be obtained only once, and the sums of the squares of the original measures of each variable must likewise be computed only once. The means are obtained from the sums of the original measures; the standard deviations of each variable are computed from the sums of the squares. In the following pages we shall illustrate Method III for three variables, x, y, and z. However, in order to simplify the illustration, the data of only ten cases are used.

The *inter-correlation* of three or more variables signifies that each variable is correlated with every other one. If only three variables are to be inter-correlated, only three coefficients will be needed: r_{xy}, r_{xz}, and r_{yz}. However, the number of inter-correlation coefficients increases rapidly with an increase in the number of variables to be inter-correlated. For n variables, the number of inter-correlation coefficients is:

$$\text{Number of inter-}r\text{'s} = \frac{n(n-1)}{2} \qquad [9:19]$$

Number of inter-correlation coefficients between n variables

Thus if there are ten variables to be inter-correlated,

$$10(9)/2 = 45$$

and 45 r's must be computed.

In the example developed in Tables 9:3 to 9:6, three variables have been inter-correlated to illustrate the labor-saving characteristic of Method III

Table 9:3. Method III for *r*: Work Table and Checks; Original Data, Means, Squares, and Cross-Products

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
				Check Columns							
Subjects	Original			Sums	Sums Squared	Squares for σ's			Cross-Products		
	X	Y	Z	$(X+Y+Z)$	$(X+Y+Z)^2$	X^2	Y^2	Z^2	XY	XZ	YZ
1	15	30	9	54	2916	225	900	81	450	135	270
2	13	24	10	47	2209	169	576	100	312	130	240
3	11	24	5	40	1600	121	576	25	264	55	120
4	9	20	7	36	1296	81	400	49	180	63	140
5	17	26	9	52	2704	289	676	81	442	153	234
6	7	15	6	28	784	49	225	36	105	42	90
7	15	34	11	60	3600	225	1156	121	510	165	374
8	14	27	8	49	2401	196	729	64	378	112	216
9	16	36	9	61	3721	256	1296	81	576	144	324
10	11	22	4	37	1369	121	484	16	242	44	88
$N = 10$	$\Sigma = 128$	258	78	464	22600	1732	7018	654	3459	1043	2096
	ΣX	ΣY	ΣZ	$\Sigma(X+Y+Z)$	$\Sigma(X+Y+Z)^2$	$\Sigma(X^2)$	$\Sigma(Y^2)$	$\Sigma(Z^2)$	$\Sigma(XY)$	$\Sigma(XZ)$	$\Sigma(YZ)$
Means	12.8	25.8	7.8	46.4	2260.0	173.2	701.8	65.4	345.9	104.3	209.6

The mean of the x variable: $M_X = 12.8$
The mean of the y variable: $M_Y = 25.8$
The mean of the z variable: $M_Z = 7.8$

Check I: $\Sigma X + \Sigma Y + \Sigma Z = \Sigma(X+Y+Z)$
$128 + 258 + 78 = 464$

Check II: $M_X + M_Y + M_Z = M_{(X+Y+Z)}$
$12.8 + 25.8 + 7.8 = 46.4$

Check III: $\Sigma(X^2) + \Sigma(Y^2) + \Sigma(Z^2) + 2[\Sigma(XY) + \Sigma(XZ) + \Sigma(YZ)] = \Sigma[(X+Y+Z)^2]$
$1732 + 7018 + 654 + 2(3459 + 1043 + 2096) = 22600$

Check IV: $M_{X^2} + M_{Y^2} + M_{Z^2} + 2(M_{XY} + M_{XZ} + M_{YZ}) = 2260$
$173.2 + 701.8 + 65.4 + 2(345.9 + 104.3 + 209.6) = 2260$

when the inter-correlations of variables are to be obtained. As shown in Table 9:3, the measures of each variable are summed only once. Similarly, the squares of the measures of each variable are computed and summed only once. When the inter-correlations of ten variables are required, the only additional labor consists in computing the products of each pair of measures for each inter-correlation coefficient. Since ten variables yield 45 inter-correlation coefficients, 45 columns will be required for their cross-products. On the other hand, inter-correlating 45 variables with Method II requires 45 separate correlation charts, for each of the 45 sets of bi-variate data must be cross-tabulated into separate charts for the computations described in Table 9:2. Thus it should be apparent that Method III really saves labor when many inter-correlation coefficients are needed, provided, of course, that a calculating or adding machine is available.

The product of the reciprocal of N and any other number is equal to the quotient of the latter number divided by N. Thus, where ΣX represents any other number:

$$\frac{\Sigma X}{N} = \Sigma X\left(\frac{1}{N}\right) = \Sigma X \text{ (reciprocal of } N)$$

If $\Sigma X = 342$ and N is 50:

$$\frac{\Sigma X}{N} = \frac{342}{50} = 342\left(\frac{1}{50}\right) = 342(.02) = 6.84$$

For the data in Table 9:3, the use of reciprocals to obtain the means of the sums of each column is obviously unnecessary, because the total number of cases is ten. The division of each sum by 10 requires only a shifting of the decimal point one place to the left.

The mean of the x variable is seen to be 12.8; that of the y variable, 25.8; and that of the z variable, 7.8.

Checks for the Computations in Table 9:3

Before any of the checks now to be described for Table 9:3 are used, and before columns 5 through 12 are computed, the original entries in columns 2 through 4 should be checked. If any errors are made in entering the original measures in these three columns, the ensuing computations cannot be correct and the checks will not reveal such errors.

The following four checks are required to insure the correctness of the computations in Table 9:3:

Check I:

$$\begin{aligned} \Sigma X + \Sigma Y + \Sigma Z &= \Sigma(X + Y + Z) \\ 128 + 258 + 78 &= 464 \\ 464 &= 464 \end{aligned}$$

This check establishes the correctness of the additions in columns 2 through 4, which are used to obtain the means of the distribution. This check is based on the fact that the total of the sums of the *columns* in any table of numbers should equal the total of the sums of the *rows* in the table.

Check II:

$$\begin{aligned} M_X + M_Y + M_Z &= M_{(X+Y+Z)} \\ 12.8 + 25.8 + 7.8 &= 46.4 \\ 46.4 &= 46.4 \end{aligned}$$

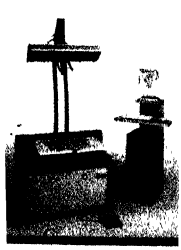
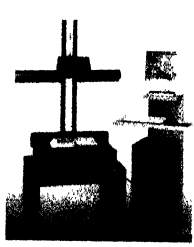
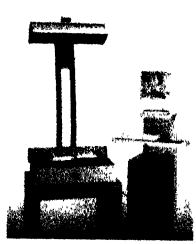
Check II is based on the same principle as the preceding, and is made for computing the means of columns 2 through 4.

Check III:

$$\begin{aligned} \Sigma(X^2) + \Sigma(Y^2) + \Sigma(Z^2) + 2[\Sigma(XY) + \Sigma(XZ) + \Sigma(YZ)] &= \Sigma[(X + Y + Z)^2] \\ 1732 + 7018 + 654 + 2(3459 + 1043 + 2096) &= 22600 \\ 22600 &= 22600 \end{aligned}$$

If you are looking for a perfect solution for the digitization and preservation of valuable cultural works and documents, talk to us. We have been in the market since 1961 and offer you complete digital and analog systems for scanning and microfilm technology at the highest level

Zeutschel, the Future of the Past.

Check III, which establishes the correctness of the remaining computations in the table, is particularly valuable because it enables detection of any possible error in squaring the original measures or in calculating their cross-products. This check is developed on the fact that the sum of the square of the sums of a series of numbers is equal to the sums of the squares of each number plus twice the sum of all their cross-products (taken two at a time). The check sum, which is obtained from column 6, is 22,600 for the data in Table 9:3.

Check IV:

$$M_X^2 + M_Y^2 + M_Z^2 + 2(M_{XY} + M_{XZ} + M_{YZ}) = 2260$$
$$173.2 + 701.8 + 65.4 + 2(345.9 + 104.3 + 209.6) = 2260$$
$$2260 = 2260$$

Check IV establishes the correctness of the means of columns 7 through 12, and is based upon the same principle as Check III. The check for the means is obtained from the mean of column 6.

Before proceeding with the computation of the standard deviations in Table 9:4, we shall summarize the order of operations in Table 9:3.

Summary of Operations: Table 9:3

1. Number consecutively each subject in the group (column 1).
2. Record each subject's original scores for each of the three variables in the X, Y, and Z columns (columns 2, 3, and 4). Check all these entries for accuracy.
3. Sum each subject's scores and record each sum in column 5, headed $(X + Y + Z)$. Square each of these sums and enter the results in column 6, headed $(X + Y + Z)^2$.
4. Square each original score of X in column 2 and enter it in column 7, headed X^2. Similarly, square each original score of Y and Z in columns 3 and 4 respectively, and enter the squared values in columns 8 and 9, headed Y^2 and Z^2.
5. Compute the cross-products for each subject's score pairs for all variables taken two at a time. Enter these cross-products in columns 10, 11, and 12, headed XY, XZ, and YZ.
6. *Sum* all the columns of the table.
7. Obtain the means of each column by dividing each sum by N, the total number of cases (in this case $N = 10$).
8. Apply checks I, II, III, and IV as indicated. The computations in Table 9:3, made from the original scores, and the means of all the columns are correct if the results satisfy the four checks. (Note: If decimals are dropped from the mean values of each column, checks II and IV may not be *exact.*)

Computation of Standard Deviations of All Variables (Table 9:4)

The computation of the standard deviation of an ungrouped distribution by the short method, when the original scores are taken as deviations from a guessed mean equal to zero, may be formulated as follows:

[9 : 20]

$$\sigma_v = \sqrt{M_{V^2} - (M_v)^2}$$

Standard deviation with guessed mean of variable taken equal to zero

where the subscript v is used to represent *any* of the variables being inter-correlated, M_{V^2} is the mean of the deviations squared, and M_v is the mean of the distribution. This formulation is for use with the means obtained from Table 9:3 and is the same as the following:

$$\sigma_v = \sqrt{\frac{\Sigma(V^2)}{N} - \left(\frac{\Sigma V}{N}\right)^2} \qquad [9 : 20a]$$

This is identical with Formula 7:6a.

The variables being inter-correlated, and for which the standard deviations are therefore necessary, are listed in column 13 of Table 9:4. The means of the distributions of each variable are listed in column 14, their values having been obtained from Table 9:3. Each of these means is squared in column 15. In column 16 are entered the values of the means of the squared deviations from columns 7 through 9 of Table 9:3.

Table 9:4. Standard Deviations for All Variables in Table 9:3

$$\sigma_v = \sqrt{M_{V^2} - (M_v)^2}$$

(Where v Stands for Any Variable)

(13) Variable v	(14) Means (from Table 9:3) M_v	(15) Means Squared $(M_v)^2$	(16) Means of Squares (from Table 9:3) M_{V^2}	(17) Variance σ^2 $M_{V^2} - (M_v)^2$	(18) Standard Deviation σ $\sqrt{M_{V^2} - (M_v)^2}$
x	12.8	163.84	173.2	9.36	3.06
y	25.8	665.64	701.8	36.16	6.01
z	7.8	60.84	65.4	4.56	2.14
Check sums	46.4	890.32	940.4	50.08	11.21

The standard deviation of the x variable: $\sigma_x = 3.06$
The standard deviation of the y variable: $\sigma_y = 6.01$
The standard deviation of the z variable: $\sigma_z = 2.14$

Check V: $\Sigma M_{V^2} - \Sigma(M_v)^2 = \Sigma\sigma^2$
$940.4 - 890.32 = 50.08$

The *variance*, σ^2, of each distribution (the square of the standard deviation), is entered in column 17 and is equal to the difference between the mean of the squared deviations (original scores) and the square of the mean of the distribution. The latter value, obtained from column 15, is the correction factor, c, developed earlier for computing the standard deviation by the short method. When the guessed mean is taken as equal to zero, c_x is equal to $\Sigma X/N$, which gives the mean of the distribution.

The values of the standard deviations of each distribution are next computed and entered in column 18. They are equal to the square root of the variances (column 17) obtained for each variable.

The only check necessary for most of the computations in Table 9:4 is Check V, which appears at the bottom of the table. In order to apply it, columns 15 through 17 must be summed. The principle of this check is similar to that of Check I for the data in Table 9:3; that is, the sum of a series of differences obtained row by row from a table should be equal to the differences between the sums of the respective columns. Thus, the sum of the variances obtained in column 17 should equal the difference between the sum of the values in column 16 and the sum of the values in column 15.

Check V does not verify the accuracy of any entries in Table 9:4 taken from Table 9:3; these have to be checked independently. Nor does this check test the accuracy of the square roots of the variances, i.e., the standard deviation values in column 18. However, these latter values will be tested in Table 9:6 by Check VIII.

Computation of the Mean of the Product Deviations of Each Bi-Variate Distribution (Table 9:5)

The computation of the mean of the product deviations of a bi-variate distribution by the short method may be formulated as follows:

$$M_{uv} = M_{UV} - M_u M_v \qquad [9:21]$$

Mean of product deviations of correlated variables with guessed means at zero

where the subscripts u and v stand for any two variables being correlated. M_{UV} is the mean of the product deviations when the deviations are taken from a guessed mean of zero and are consequently equal to the values of the original measures of the distributions being correlated. M_u and M_v are, as usual, the means of the respective distributions of the variables being correlated; they are the correction factors for the computation of r by this method. The data in columns 20, 21, and 22 in Table 9:5 are from Table 9:3.

Check VI establishes the correctness of the products of the means in column 23 and is based on the sums of columns 14 and 15 of Table 9:4. Check VII verifies the sum of the means of the product deviations of column 24.

Table 9:5. Mean of Product Deviations for Each Pair of Variables in Table 9:1

$$M_{uv} = M_{UV} - M_u M_v$$

(where u and v stand for any two variables being correlated)

(19) Variables Correlated U and V	(20) Means of Cross-Products (from Table 9:3) M_{UV}	(21) Means of Distributions Correlated (from Table 9:3) M_u	(22) M_v	(23) Products of Means of Distributions $M_u M_v$	(24) Means of Product Deviations $M_{UV} - M_u M_v$
x with y	345.9	12.8	25.8	330.24	15.66
x with z	104.3	12.8	7.8	99.84	4.46
y with z	209.6	25.8	7.8	201.24	8.36
Check sums	659.8			631.32	28.48

The mean of the product deviations of x and y: $M_{xy} = 15.66$
The mean of the product deviations of x and z: $M_{xz} = 4.46$
The mean of the product deviations of y and z: $M_{yz} = 8.36$

Check VI:

$$(M_X + M_Y + M_Z)^2 - [(M_X)^2 + (M_Y)^2 + (M_Z)^2] = 2[\Sigma(M_x M_y + M_x M_z + M_y M_z)]$$
$$(46.4)^2 - 890.32 = 2(631.32)$$
$$2152.96 - 890.32 = 1262.64$$

Check VII:

$$\Sigma(M_{XY} + M_{XZ} + M_{YZ}) - \Sigma(M_x M_y + M_x M_z + M_y M_z) = \Sigma(M_{xy} + M_{xz} + M_{yz})$$
$$659.8 - 631.32 = 28.48$$

The preceding formulation of the mean of the product deviations is for use with the means obtained from Table 9:3. Formula 9:21 may be stated in original score form as follows:

$$\text{Mean of product deviations} = \frac{\Sigma(UV)}{N} - \left(\frac{\Sigma U}{N}\right)\left(\frac{\Sigma V}{N}\right) \qquad [9:21a]$$

Computation of the Correlation Coefficients (Table 9:6)

The final step in computing the inter-correlation coefficients is shown in Table 9:6, and is based upon the formulation for r in Formula 9:18b. That is, the correlation coefficient in each case is the ratio of the product deviations obtained in Table 9:5 to the products of the standard deviations of the two variables being correlated. The standard deviations were obtained in Table 9:4. In column 26 of Table 9:6 are entered the means of the product deviations from Table 9:5. In columns 27 and 28 are entered the standard deviations of the variables being correlated from Table 9:4. Column 29 shows the products of the standard deviations. In column 30 are entered the ratios of the means of the product deviations to the products of the standard deviations to give the correlation coefficients. The coefficients for the data used to illustrate Method III range from .65 to .85.

Table 9:6. The Product-Moment Correlation Coefficients

$$r_{uv} = \frac{M_{uv}}{\sigma_u \sigma_v}$$

(where u and v stand for any two variables being correlated)

(25)	(26)	(27)	(28)	(29)	(30)
Variables Correlated	Means of Product Deviations (from Table 9:5, column 24)	Standard Deviations (from Table 9:4, column 18)		Products of Standard Deviations	Product-Moment Correlation Coefficients
u and v	M_{uv}	σ_u	σ_v	$\sigma_u \sigma_v$	$\frac{M_{uv}}{\sigma_u \sigma_v}$
x with y	15.66	3.06	6.01	18.39	$\frac{15.66}{18.39} = .85 = r_{xy}$
x with z	4.46	3.06	2.14	6.54	$\frac{4.46}{6.54} = .68 = r_{xz}$
y with z	8.36	6.01	2.14	12.86	$\frac{8.36}{12.86} = .65 = r_{yz}$
Check sum				37.79	

The correlation coefficient for variable x with y: $r_{xy} = .85$
The correlation coefficient for variable x with z: $r_{xz} = .68$
The correlation coefficient for variable y with z: $r_{yz} = .65$

Check VIII: $(\Sigma\sigma)^2 - \Sigma(\sigma^2) = 2\Sigma$ (product of standard deviations)
$(11.21)^2 - 50.08 = 2(37.79)$
$75.58 = 75.58$

The final check, Check VIII, to be applied to the computations of Table 9:6 is as follows:

Check VIII:

$$(\Sigma\sigma)^2 - \Sigma(\sigma^2) = 2\Sigma \text{ (product of standard deviations)}$$
$$(11.21)^2 - 50.08 = 2(37.79)$$
$$75.58 = 75.58$$

This check tests the accuracy of the products of the standard deviations computed in this table. The first term of the check, $(\Sigma\sigma)^2$, is obtained from the sum of column 18 in Table 9:4 and the second term is obtained from the sum of column 17 in the same table. Check VIII thus tests the accuracy not only of the standard deviations obtained in Table 9:4, but also of the products of the standard deviations in column 29 in Table 9:6.

The means of the product deviations entered in column 26 must be independently checked with the values for these means obtained in Table 9:5. There remain to be checked only the ratios for the correlation coefficients themselves, in column 30. A method for checking these ratios independently is to repeat the operation in reverse, either (1) by multiplying the correla-

tion coefficient by the product of its standard deviations to give the value of the mean of the product deviations, or (2) by dividing the mean of the product deviations by its corresponding correlation coefficient to give the value of the corresponding product of the standard deviations. If the computation of r is correct, either of these results should check, except for dropped decimals.

G. OTHER METHODS FOR THE COMPUTATION OF r

We have illustrated three of the commonly used methods for computing product-moment correlations. Methods I and III, for ungrouped data, ordinarily have an advantage over Method II, since they give a more precise estimate of r than is the case when the data are grouped into class intervals. Method II, however, has the advantage inherent in the portrayal of a bivariate distribution by a correlation chart; the investigator can usually see whether a straight-line function is in reality the appropriate one to use in the correlation of the two variables. Furthermore, the sacrifice of mathematical precision when the data of one or both variables are grouped into class intervals of several units is negligible when there are at least 12 class intervals and 40 or more correlational frequencies.

Short Methods II and III of course have the computational advantage over Method I that is implied by the characterization "short." Method III, however, is satisfactory from a computational point of view only if a machine calculator is available, particularly for adding columns of numbers, squares, and cross-products.

In the next chapter we shall discuss methods for the correlation of bivariate data that provide estimates of r in cases in which the use of the product-moment method is not feasible or convenient. However, before these different techniques are considered, two additional methods for computing r should be discussed briefly: one by the *sums* of paired deviations; the other, by the *differences* of paired deviations.* Both give a result which is algebraically the same as the *product*-moment method.

The Method of Sums for r

If s represents the sum of two paired deviations, x and y, then

$$r = \frac{\sigma_s^2 - \sigma_x^2 - \sigma_y^2}{2\sigma_x\sigma_y} \qquad [9:22]$$

Pearson r by the method of sums

where σ_s^2 is equal to the variance of the sums of all paired deviations; σ_x^2 and σ_y^2 are the variance of each variable correlated; and σ_x and σ_y are their standard deviations.

* Cf. Peters and Van Voorhis, *op. cit.*, pp. 101–103, for the derivation of these formulas from product-moment r.

This procedure for computing r is based on the *sums* of associated pairs rather than on their products. The variance of the paired sums is as follows:

$$\sigma_s^2 = \frac{\Sigma(x+y)^2}{N} \qquad [9:23]$$

Variance of paired sums

This method for computing r is used only infrequently; however, the method of *differences* has a very practical application that we shall illustrate with the digit-span test scores in Table 9:2.

The Method of Differences for r

If d represents the difference between the deviations, x and y, of two paired measures, the variance of all the paired differences of a bi-variate distribution equals:

$$\sigma_d^2 = \frac{\Sigma(x-y)^2}{N} \qquad [9:24]$$

Variance of paired differences

and the correlation coefficient by the method of differences is:

$$r = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_d^2}{2\sigma_x\sigma_y} \qquad [9:25]$$

Pearson r by method of differences

This formula can, under certain circumstances, be simplified to yield a convenient method for computing a reliability coefficient of a test (cf. Chapter 17, Section B). Thus, if the variance and the means of two forms of a test, or two halves of a test, can be assumed to be equal, Formula 9:25 becomes:

$$r = 1 - \frac{\Sigma(D^2)}{2N\sigma_x^2} \qquad [9:26]$$

Special case of 9:25

where $\Sigma(D^2)$ is the sum of the squared differences between the original values, X and Y, of each pair of associated measures; N is the number of correlational frequencies, and σ_x^2 is the variance of the test.

The correlation coefficient for the digit-span data in Table 9:2 is computed in Table 9:7 by this latter method. We saw in Table 9:2 that the mean of the first digit-span test (x) was 7.45, and that of the second test (retest with another series of numbers) was 7.65. The standard deviation of the first was 1.30, and of the second, 1.22. The means and standard deviations of each test were thus similar, although not identical.

The difference, D, between each subject's pair of scores is obtained in column 4 and these differences are squared in column 5 to give D^2. The sum of the squared differences is 11.1250. If the square of the standard deviation of the first administration of the test, which was found in Table 9:2 to be 1.3,

Table 9:7. Computation of r by the Method of Differences (Digit-Span Test Data and Variance of x from Table 9:2) *

(1) Subjects	(2) Digit-Span Scores 1st Test X	(3) Retest Y	(4) Differences Between X and Y D	(5) D^2
A	8.5	8.25	.25	.0625
B	6.75	8.25	−1.50	2.2500
C	6.25	7.75	−1.50	2.2500
D	6.0	5.5	.50	.2500
E	7.0	7.25	−.25	.0625
F	9.25	9.25	0	0
G	5.5	5.75	−.25	.0625
H	9.25	8.25	1.00	1.0000
I	7.75	6.75	1.00	1.0000
J	5.0	5.25	−.25	.0625
K	6.5	7.25	−.75	.5625
L	7.75	8.5	−.75	.5625
M	9.0	8.0	1.00	1.0000
N	7.5	8.25	−.75	.5625
O	7.0	7.75	−.75	.5625
P	7.75	8.25	−.50	.2500
Q	5.75	5.75	0	0
R	8.75	9.5	−.75	.5625
S	8.75	8.50	.25	.0625
T	9.0	9.00	0	0
$N = 20$	149.0	153.0		$\Sigma = 11.1250$

$$r = 1 - \frac{\Sigma D^2}{2N\sigma_x^2} = 1 - \frac{11.1250}{2(20)(1.3)^2} = .84$$

* The value of σ_x for these data was 1.3.

is used for the variance of the test, the coefficient r by the method of differences is as follows:

$$r = 1 - \frac{11.1250}{2(20)(1.3)^2} = 1 - \frac{11.1250}{67.60}$$
$$= 1 - .165 = .835, \text{ or } .84$$

which is identical with the value for r obtained in Table 9:2 by the product-moment method.

When this abbreviated method of differences is employed alone, two additional columns are necessary in Table 9:7 in order to obtain the variance of the test. These are columns (4) and (6), for x and x^2 in Table 9:2. If the abbreviated method of Formula 9:26 cannot be employed, the variance and standard deviation of y must also be computed, as indicated in Formula 9:25.

EXERCISES

1. Cite two examples of perfect correlation other than those mentioned in this chapter.
2. What is the difference between a scattergram and a correlation chart?
3. Under what circumstances is it advisable to make a scattergram or correlation chart of a bi-variate distribution?
4. What can be inferred about the correlation between two variables from a scattergram?
5. What are the basic assumptions for the use of Pearson's product-moment method of correlation?
6. What is the difference between a correlational frequency and a statistical frequency?
7. Why must the data of a bi-variate distribution be associated by individual pairs in order for a correlation coefficient to be calculated?
8. What essential properties do some of the fourfold charts in Chapter 4 and the correlation chart in Fig. 9:8 have in common?

Use the data in Table 5:14 for the following six problems:

9. Set up a correlation chart in z score form and estimate the degree of correlation between the average grades and intelligence test scores of the college freshmen from the regression line of grades on intelligence test scores.
10. Compute the correlation between the intelligence test scores of college freshmen and their best friends by means of the correlation chart method used in Fig. 9:13.
11. Using the data for only the last 25 cases, compute by the long method of Table 9:2 the product-moment correlation between grades of the college freshmen and of their best friends.
12. Use Method III (Tables 9:3–9:7) to compute the inter-correlations between the grades, intelligence test scores, and ages of the college freshman group.
13. Set up a regression equation in original score form and predict for the data in Exercise 10 the average intelligence test score received by the best friends of college freshmen whose intelligence test scores were 90.
14. Set up a regression equation in original score form and predict for the data in Exercise 11 the average grade received by the best friends of college freshmen whose average grade was 73.
15. Use the method of differences (Formula 9:25) to compute the correlation of the height-weight data in Table 9:1.

Use the data of Table 9:8 for the following problems:

16. Make a scattergram and then determine the correlation between 1943–44 total annual expenditures and expenditures for instruction in the 68 cities.
17. What are the mean total expenditure and the mean expenditure for instruction?
18. Is the variability in total expenditures among the 68 cities *relatively* different from the variability in expenditures for instruction?
19. What is the average amount spent for instruction of cities whose total expenditures were
    a. over $200.00
    b. less than $100.00
    c. between $130.00 and $140.00
    d. between $170.00 and $180.00

Table 9:8. Total 1943–44 Annual Expenditure per Pupil in Average Daily Attendance, and Expenditure per Pupil for Instruction, in the School Systems of 68 Cities of from 30,000 to 100,000 Population *

City	Total Annual Expenditure	Annual Expenditure for Instruction
Fort Smith, Ark.	$ 68.97	$ 54.13
Little Rock, Ark.	70.66	55.14
Alhambra, Calif.	156.62	118.13
Glendale, Calif.	166.22	130.27
Santa Barbara, Calif.	175.59	131.69
Stamford, Conn.	178.86	140.74
Waterbury, Conn.	167.22	131.73
West Hartford, Conn.	136.83	106.33
Aurora (East Side), Ill.	126.67	96.79
Danville, Ill.	98.53	68.91
Decatur, Ill.	93.98	73.56
Elgin, Ill.	119.39	93.10
Moline, Ill.	117.39	79.53
Quincy, Ill.	114.52	87.26
Rock Island, Ill.	94.41	71.55
Elkhart, Ind.	113.83	86.54
Evansville, Ind.	126.59	97.65
Davenport, Iowa	126.71	92.15
Dubuque, Iowa	164.45	120.65
Ottumwa, Iowa	101.25	78.11
Covington, Ky.	128.35	102.85
Lexington, Ky.	102.40	83.44
Brookline, Mass.	195.44	149.63
Chicopee, Mass.	146.48	109.06
Holyoke, Mass.	175.01	131.75
Lynn, Mass.	152.82	114.39
Medford, Mass.	123.78	100.99
Salem, Mass.	147.45	113.67
Battle Creek, Mich.	131.02	95.34
Dearborn, Mich.	170.70	124.74
Jackson, Mich.	129.87	94.94
Kalamazoo, Mich.	142.05	105.87
Lansing, Mich.	130.12	99.53
Jackson, Miss.	72.70	59.46
Joplin, Mo.	81.72	60.73
Nashua, N. H.	128.52	90.47
Atlantic City, N. J.	205.31	156.14
East Orange, N. J.	203.96	164.91
Hoboken, N. J.	207.95	149.51
Irvington, N. J.	181.69	137.99
Montclair, N. J.	242.66	192.95
New Brunswick, N. J.	178.22	143.25
Plainfield, N. J.	179.04	140.75
Albuquerque, N. Mex.	88.78	71.25
Elmira, N. Y.	142.92	111.39

* From *School Life*, December, 1945, p. 23.

Table 9:8 — (*Continued*)

City	Total Annual Expenditure	Annual Expenditure for Instruction
Jamestown, N. Y.	$161.01	$116.56
Troy (Union District), N. Y.	166.50	117.20
White Plains, N. Y.	257.54	193.73
Cleveland Heights, Ohio	196.40	137.70
Lakewood, Ohio	190.43	143.98
Marion, Ohio	85.88	63.73
Steubenville, Ohio	126.92	96.72
Harrisburg, Pa.	161.30	120.88
New Castle, Pa.	132.50	96.18
Wilkes-Barre, Pa.	151.29	114.41
Cranston, R. I.	120.66	97.27
Spartanburg, S. C.	73.10	62.30
El Paso, Tex.	78.54	62.35
Port Arthur, Tex.	81.13	63.78
Waco, Texas	79.17	65.79
Petersburg, Va.	90.65	76.17
Portsmouth, Va.	84.89	71.75
Everett, Wash.	110.18	82.89
Madison, Wis.	155.07	122.36
Oshkosh, Wis.	136.25	105.12
Racine, Wis.	127.19	97.13
Sheboygan, Wis.	117.80	87.80
West Allis, Wis.	145.43	112.75

CHAPTER 10

Special Methods for the Linear Correlation of Variables

Several methods, other than those described in the preceding chapter, are available for the linear correlation of two variables. These other methods yield coefficients analogous in their general implications to r, and are used (1) when the product-moment method described in Chapter 9 cannot be directly employed because of the nature of the data to be correlated, or (2) when one of these other methods is more desirable as a short-cut computational procedure. The following methods will be described:*

A. Correlation of Ranks.
B. Serial Correlation.
C. Tetrachoric r.

A. CORRELATION OF RANKS

Purpose of the Method

A method for the correlation of ranks was developed several decades ago by the English psychologist and statistician, the late Charles Spearman, in order to provide an estimate of linear correlation when either or both of the variables being correlated could at least be ranked if not quantitatively differentiated. This and the several other methods that were eventually developed are ordinarily used for variables that cannot be differentiated in a satisfactory quantitative way, but for which ratings can be differentiated and then ranked. However, product-moment r is a general method for the correlation of linear relationships, whether or not the data of one or both variables are quantitatively differentiated or are in the form of centiles or ranks. Hence, the method now to be described is essentially a short-cut procedure for bi-variate data.

Variables that can be differentiated by ranks often occur in the ratings of aesthetic qualities, of personalities, of achievement or success, of quality of performance on a job, etc. Ratings are commonplace in psychological measurement.

* The ϕ coefficient and the Coefficient of Mean Square Contingency are also used to give estimates of r, i.e., linear correlation. Since they were presented in chap. 4, they will not be discussed further in the present chapter.

Methods for correlating ranks are also sometimes used as short-cut devices for the correlation of variables that are actually quantitatively differentiated, as when only small groups or samples of data are available for correlation. Rank methods are often employed under these circumstances rather than the product-moment method because they are easier to apply, and, under certain circumstances, they yield as satisfactory a result.

Spearman's Rank-Difference Method

Three variations of methods for the correlation of ranks have been developed.* They are as follows:

1. The Rank-Difference Method.
2. The Rank-Product Method.
3. The Rank-Sum Method.

The first of these has long been used and is known as the Spearman rank-difference method. The computational procedure is based upon the *differences* between the ranks of each associated pair of ratings or scores for the variables being correlated. The second method is based upon the *products* of these ranks, and the third is based upon the *sum* of these ranks. The tables developed by Du Bois have made the computations necessary for any one of these methods somewhat simpler.† Inasmuch as all these methods yield the same result, only the one most commonly used, the *rank-difference* method, will be described.

When a different rank can be assigned to each member or case of the variables being correlated, the coefficient obtained is equivalent to the product-moment r. In other words, if the ratings obtained for a variable are all different, they can be ranked as follows: 1, 2, 3, 4, $\cdots n$. Often this ranking is impossible because two or more cases may receive the same rating or score. Several procedures have been developed for treating such duplications.‡ However, if ratings or scores are duplicated in only a small proportion of the cases, the resulting coefficient should not be particularly distorted and consequently will still yield a satisfactory estimate of linear correlation.

In order to distinguish a correlation coefficient obtained by a rank method from one obtained by the product-moment method, ρ (rho), the symbol for the Greek letter r, is used.

The development of Spearman's rank-difference method is illustrated in Table 10:1. The correlation coefficient itself is computed by the following formula:

* These three variations are analogous to those for product-moment r described at the end of the last chapter, viz., the methods of *sums* and of *differences*, and of *product* deviations. Spearman's *rank-difference* method is a special case of the method of differences for r. (Cf. the end of the present section.)

† Philip H. Du Bois, "Formulas and Tables for Rank Correlation," *Psychological Record*, 3:46–56, 1939.

‡ *Ibid.*

$$\rho = 1 - \frac{6\Sigma(D^2)}{N(N^2 - 1)} \qquad [10:1]$$

Spearman's rank-difference coefficient for linear correlation

where $\Sigma(D^2)$ represents the sum of the squares of the differences between the ranks of each associated pair, 6 is a constant, and N as usual is the number of associated pairs in the group or sample of the variables correlated.

Table 10:1. Spearman's Method of Rank-Difference Correlation: Correlation Between Aptitude Test Scores and Achievement Ratings

(1) Subject	(2) Achievement Ratings by Ranks	(3) Scores on Aptitude Test	(4) Rank or Mid-Rank for Ties: Aptitude Test	(5) Rank Differences (D)	(6) D^2
A	1	72	2	1.0	1.0
B	2	60	5.5	3.5	12.25
C	3	65	4	1.0	1.0
D	4	60	5.5	1.5	2.25
E	5	76	1	4.0	16.0
F	6	68	3	3.0	9.0
G	7	52	12.5	5.5	30.25
H	8	56	8.0	0	0
I	9	54	10.5	1.5	2.25
J	10	56	8.0	2.0	4.0
K	11	52	12.5	1.5	2.25
L	12	54	10.5	1.5	2.25
M	13	48	14	1.0	1.0
N	14	56	8.0	6.0	36.0
O	15	36	19	4.0	16.0
P	16	46	15	1.0	1.0
Q	17	40	17.5	.5	.25
R	18	30	20	2.0	4.0
S	19	40	17.5	1.5	2.25
T	20	44	16	4.0	16.0
					$\Sigma = 159.00$

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} = 1 - \frac{6(159)}{7980} = 1 - .1196 = .88$$

The variables correlated in Table 10:1 (achievement ratings and aptitude scores) are suggestive of the type of situation for which the rank method is useful. Results for only 20 subjects are presented in the table. The subjects are designated by the letters A to T in column 1. Each subject's achievement rating is listed in rank order in column 2. No ties in ratings are indicated. Subject A received the highest rating and consequently has a rank of 1; Subject B received the next highest rating and has a rank of 2, etc. The aptitude test scores of the 20 subjects are listed in column 3. They range in

size from 30 to 76, but they are not listed in order of size because each subject's aptitude score must be paired with his achievement rating. Only if the correlation between scores and ratings were perfect (1.00) would the scores in column (3) prove to be listed in order of size.

Ranking the Test Scores

Since the achievement ratings are already ranked in column 2, it is unnecessary to adjust them further in preparing to compute the correlation coefficient. However, the aptitude test scores in column 3 must be ranked before the computations can be made. These ranks, which are presented in column 4, are obtained as follows:

The subject receiving the highest score on the aptitude test (Subject E) is given a rank of 1. The one with the next highest score (Subject A) is given a rank of 2. The third highest score was received by Subject F; the fourth highest, by Subject C. Subjects B and D, however, both received the same score, 60. These two cases represent a duplication and require an adjustment in the ranking procedure, since one subject cannot very well be given a rank of 5 and the other a rank of 6. Several methods have been developed for making this adjustment,* but the one generally found most satisfactory and the simplest to use is *averaging the ranks* involved in any duplication.

Averaging Ranks

Since the scores of Subjects B and D are the fifth and sixth cases in order of size, their average rank is 5.5 and therefore this rank is assigned to both of them in column 4. The seventh highest score in column 3 is 56, but three subjects—H, J, and N—have this score. These three cases are the seventh, eighth, and ninth in order of size, and hence the average of their ranks is 8.0. Although the ranks in column 4 can be obtained by inspection, the surest way to avoid error is to *list* the scores to be ranked in order of size and to number the order of the scores thus listed, as follows:

76....1		52....12	= 12.5
72....2		52....13	
68....3		48....14	
66....4		46....15	
60....5	= 5.5	44....16	
60....6		40....17	= 17.5
56....7		40....18	
56....8	= 8.0	36....19	
56....9		30....20	
54....10	= 10.5		
54....11			

Such a procedure makes it obvious that there are two cases with a value of 60, they were fifth and sixth in order of size, and therefore their average rank

* Philip H. Du Bois, "Formulas and Tables for Rank Correlation," *Psychological Record*, 3:46–56, 1939.

is 5.5. Similarly, there are three cases with the value of 56; they were the seventh, eighth, and ninth in order of size; and their average rank is 8.0. There are two scores of 54: here the average rank is 10.5. There are also two scores of 52 with an average rank of 12.5, and there are two scores of 40 whose average rank is 17.5.

The rank of the lowest score should of course be equal to the total number of cases when there are no duplications of this lowest score. In other words, the rank of the lowest score should always be equal to N, the number of cases in the distribution, unless there are duplications of the lowest score.

Once the ranks are obtained by the above procedure, they are entered in column 4 for each subject.

The Computation of Rho

With the data of both variables now ranked, we can proceed to the computation of the correlation coefficient. This involves two steps. (1) The *difference* between the ranks of each associated pair must be obtained and entered in column 5, and (2) these differences must be squared and entered in column 6. Since only the sum of the squared differences is needed to compute ρ, it is unnecessary to take into account the direction of the difference between each rank pair and to use a plus or minus sign in column 5.

The sum of the squared differences is indicated at the bottom of column 6. This value, 159.0, is needed for the computation of rho at the bottom of the table. The coefficient is equal to .88. Relatively, this is a high correlation. Despite the fact that there were very few associated pairs whose ranks on the achievement ratings and aptitude tests were the same, the correlation is substantial. This is so because the differences between the ranks of associated pairs were not very large. The greatest difference was for Subject N, whose ranks on the achievement rating and the aptitude test were 14 and 8.0 respectively. This rho coefficient of .88 is analogous to a product-moment r and is indicative of a degree of co-relationship between two variables that would be implied by r itself.

When the differences between the ranks of the associated pairs of two variables are at a maximum, the correlation coefficient will not be zero but will approach -1.00 as a limit. This is so because the subject with the highest rank on one variable will have the lowest rank on the other, etc., and a perfect inverse relation will obtain. A low correlation, that is, one close to zero, occurs when there is no relationship between the ranks of the two variables being correlated.

The Relation of r to Rho

We saw in the preceding chapter that linear correlation may under certain circumstances be obtained by a method of differences, for which the formula was

$$r = 1 - \frac{\Sigma D^2}{2N\sigma_x^2} \qquad [9:26]$$

This formula is based on the assumption that the means and standard deviations of both variables correlated are equal. This assumption holds for each series of ranks of two correlated variables. Thus, if two series of 5 ranks each are correlated, the mean rank of both will be 3, since

$$(1 + 2 + 3 + 4 + 5)/5 = 3$$

Generally, the mean of any series of ranks, 1 to n, is equal to

$$M_{\text{ranks}} = \frac{n + 1}{2} \qquad [10:2]$$

Means of a series of ranks, 1 to n.

And the standard deviation of a series of ranks is:

$$\sigma_{\text{ranks}} = \sqrt{(x_1^2 + x_2^2 + x_3^2 \ldots + x_n^2)/N} \qquad [10:3]$$

Standard deviation of a series of ranks, 1 to n

where x_1, x_2, $x_3 \cdots x_n$ are the successive ranks expressed as deviations from the mean rank. Thus, for 5 ranks:

$$\sigma_{\text{ranks}} = \sqrt{(2^2 + 1^2 + 0 + 1^2 + 2^2)/N} = \sqrt{10/5} = \sqrt{2} = 1.41$$

Under these circumstances of equal means and standard deviations, Formula 9:26 for the correlation of differences may be adapted to the special case of *differences between ranks*, as follows:

$$r = 1 - \frac{\Sigma D^2}{2N\sigma_x^2}$$

$$= 1 - \frac{\Sigma D^2}{2N\sigma^2_{\text{ranks}}}$$

The standard deviation of a series of n ranks can be obtained more readily by the following formula than by its equivalent, Formula 10:3.

$$\sigma_{\text{ranks}} = \sqrt{\frac{N^2 - 1}{12}} \qquad [10:3a]$$

Standard deviation of n ranks

where n, the number of ranks in a distribution, is equal to N, the number of cases to be ranked. Substituting this value of σ_{ranks} in the above formula and substituting rho for r, since we are now dealing with the special case of *rank* differences, we have:

$$\rho = 1 - \frac{\Sigma D^2}{\frac{2N(N^2 - 1)}{12}} = 1 - \frac{\Sigma D^2}{\frac{N(N^2 - 1)}{6}} = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)} \qquad [10:1]$$

B. SERIAL CORRELATION

The linear correlation of a continuously distributed variable with one which is segmented or divided into only a few or even two classes is known as serial correlation. Until recently, methods for serial correlation were available only for biserial r, in which the segmented variable is dichotomized.

We shall present first the methods of biserial r and point-biserial r and then describe methods of serial correlation for variables that are segmented into more than two broad classes.

Biserial Correlation

Purpose of the Method

The biserial method for linear correlation is useful for situations in which one of the bi-variates is dichotomized rather than continuously distributed. The method is based on the assumption that the variable which is dichotomized would, if quantitatively differentiated, yield the normal, bell-type of distribution. The continuously distributed variable that is correlated with the dichotomized one is not, however, assumed to be normally distributed. The correlation coefficient obtained by the method of biserial r is symbolized by r_{bi} because it is analogous in its implications about co-variability to product-moment r.

Biserial r has been extensively used in analyses of the value of single items of psychological tests, and of the relation between test results and a dichotomized criterion, such as ratings of success and failure, good and poor, etc. (cf. Chapter 17, Section C). The usefulness of biserial r in this latter type of situation is illustrated by Table 10:2, showing the relationship between a clerical proficiency test and ratings of 133 clerical employees in the relevant skill.

When the problem in psychometrics is to determine which *items* of a test yield the best and which the poorest results, a coefficient of correlation is useful as an index of the differentiating value of single items. A valuable test item is one that is answered in such a way that those doing well, either on the test as a whole or on an independent criterion of efficiency, generally respond in the same way to the item, whereas those doing poorly on the test as a whole or on the independent criterion of efficiency, generally give a contrasting response. An item of little or no value is one that produces no differentiation with respect to the criteria used. By means of an empirical analysis of the items on a preliminary or tentative test, the investigator can establish a final test consisting only of items whose usefulness has been functionally demonstrated.

Responses to test items are often dichotomized initially; that is, the answers are scored as right or wrong, correct or incorrect. Variables that are quantitatively distributed are also sometimes dichotomized for purposes of correlation analysis into "satisfactory" and "unsatisfactory," "success" and "failure," etc. Thus, the correlation between the aptitude scores and the achievement ratings of the 20 subjects in Table 10:1 might be recast and obtained by the method of biserial correlation. In fact, the latter would ordinarily be used for validation of aptitude test scores when the group included many more individuals and it would consequently be relatively difficult to assign different

achievement ratings to all the members. In such a situation, each member of the group could be given a criterion rating of his performance in terms of "satisfactory" or "unsatisfactory."

Computation of Biserial r

We shall illustrate the computation of biserial r in several different situations for which this method is characteristically valuable. First, we shall apply it in evaluating the usefulness of an aptitude test for predicting success or failure in a clerical situation. We shall then use the method to evaluate an item of a vocabulary test against the total test score. This constitutes an example of internal validation, inasmuch as the differentiating value of the item is analyzed with respect to the total test of which it is a part, rather than with respect to an external criterion of success or failure. The internal validation of test items is common in psychometrics, but unfortunately the procedure is not as sound as validation against an independent, external criterion of efficiency (cf. Chapter 17, Section D).

The basic formula for biserial r is as follows:

$$r_{bi} = \left(\frac{M_h - M_t}{\sigma_t}\right)\left(\frac{p_h q}{y}\right) \qquad [10:4]$$

Biserial coefficient for linear correlation

where

M_h is the mean of the distribution of test scores for the part of the total group receiving the higher criterion rating.

M_l is the mean score for the remainder of the total group (lower rating part).

σ_t is the standard deviation of the test distribution for the total group.

p_h is the proportion of the total group in the *higher* criterion group.

q is the proportion of the total group in the *lower* criterion group.

y is the value of the ordinate on a normal curve at the point that divides the total distribution into two parts, with the proportion of the area *above* the point equal to p_h. (The value of the ordinate at such a point can be obtained from Table 10:3, which gives these values for a normal distribution in which the total area is taken as unity.)

A more convenient modification of this formula makes unnecessary the computation of the mean test score result for the group with the lower criterion ratings. It is as follows: *

$$r_{bi} = \left(\frac{M_h - M_t}{\sigma_t}\right)\left(\frac{p_h}{y}\right) \qquad [10:4a]$$

Dunlap's formula for biserial correlation

The symbols have the same meaning as in Formula 10:4, except that M_t, the mean of the total distribution, is employed instead of M_l, and p_h is used instead of $p_h q$. Formula 10:4a is used in the following examples.

* J. W. Dunlap, "Note on Computation of Biserial Correlations in Item Evaluation," *Psychometrika*, 1:51–60, 1936. Dunlap's article includes a table of p/y values to four decimal places for p values of .000 to .999. In the same issue of *Psychometrika* Dunlap also presents a nomograph for the computation of biserial correlations.

The computations needed, therefore, in calculating biserial r by this latter method are those that will yield the mean and standard deviation of the test scores of the total group, and the mean of the test scores for the part of the total group that receives the higher ratings. The short method has been used in computing M and σ in Tables 10:2 and 10:5. Whatever method is used for computing these two measures, the distributions used must be set up in the same way as those obtained from cross-tabulating the continuous and dichotomous variables.

Biserial Correlation of Clerical Proficiency Test Results with Independent Ratings of Efficiency

A clerical proficiency test, consisting of 53 items of an information type and designed to measure proficiency (not aptitude, or potential skill), was developed for use in the classification and selection of clerical workers whose chief task would be the preparation of technical correspondence in proper form.* The test was administered to a group of 133 employed stenographers, typists, and clerks, each of whom was independently rated on the following 3-point scale for efficiency in clerical work:

1. NO SKILL: For employees with no training who were judged to have had no significant experience in the technical correspondence under consideration.
2. SOME SKILL: For employees who had had appreciable training or experience, or both, but were not judged competent in the technical correspondence at the time the test was administered.
3. SKILLED: For employees judged fully competent in the technical correspondence, a competent person being defined as one who may be charged with responsibility for the preparation of the correspondence in proper form, from a copy or notes or instructions which provide only the substance of such communications, with no review for form prior to the preparation of final copy.

All ratings were obtained by the psychologist in conference with the supervisors of the 133 subjects. The ratings were as follows: 33 of the total group were rated as NOT SKILLED, 37 as having SOME SKILL, and 63 as SKILLED.

It will be observed that the independent criterion ratings for this group were classified into three rather than two classes. A 3-point rating scale is often easier and more desirable than a 2-point scale for such purposes. A triserial coefficient for computing a linear correlation between a scale of trichotomized ratings and a distribution of test scores will be presented later in this section. Biserial r can, however, be employed with the clerical proficiency test data if the ratings of two adjacent classes are combined; but it cannot be employed with only the upper and lower parts of a distribution, as, for example, the SKILLED vs. the NOT SKILLED, the SOME SKILLED being omitted.

* Data through courtesy of E. E. Cureton, of Richardson, Bellows, Henry and Co., Inc., New York City.

This is so because biserial r is based on the assumption that the dichotomy is for a continuous normal distribution. Since both the NOT SKILLED and the SOME SKILLED groups may be defined as NOT SKILLED, they will be combined, and the result will be a dichotomy of the SKILLED and the NOT SKILLED. This dichotomy is used in Table 10:2, which lists the test scores made by the total group as well as by each dichotomized part, and illustrates the computation of biserial r for the results.

Table 10:2. Biserial Correlation of Clerical Proficiency Test Results with Criterion Ratings

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Criterion Ratings							
Test Scores	Some Skill and No Skill f_l	Skilled f_h	Total Group f_t	t'	ft'	ft'^2	h'	fh'
51–52	0	1	1	9	9	81	6	6
49–50	1	5	6	8	48	384	5	25
47–48	0	7	7	7	49	343	4	28
45–46	3	8	11	6	66	396	3	24
43–44	2	4	6	5	30	150	2	8
41–42	2	8	10	4	40	160	1	8
39–40	1	9	10	3	30	90	0	0
37–38	6	7	13	2	26	52	−1	−7
35–36	7	2	9	1	9	9	−2	−4
33–34	6	3	9	0	0	0	−3	−9
31–32	4	2	6	−1	−6	6	−4	−8
29–30	8	4	12	−2	−24	48	−5	−20
27–28	6	1	7	−3	−21	63	−6	−6
25–26	9	2	11	−4	−44	176	−7	−14
23–24	3		3	−5	−15	75		
21–22	1		1	−6	−6	36		
19–20	6		6	−7	−42	294		
17–18	4		4	−8	−32	256		
15–16	1		1	−9	−9	81		
	$N_l = 70$	$N_h = 63$	$N_t = 133$		307	2700		99
					−199			−68
					$\Sigma = 108$			31

$$M_t = 33.5 + 2(108/133) = 35.12$$

$$\sigma_t = 2\sqrt{2700/133 - (108/133)^2} = 8.86$$

$$M_h = 39.5 + 2(31/63) = 40.48$$

$$p_h = 63/133 = .47$$

$$y = .398 \text{ (from Table 10:3, } a = .03)$$

$$r_{bi} = \left(\frac{40.48 - 35.12}{8.86}\right)\left(\frac{.47}{.398}\right) = (.605)(1.182) = .72$$

The scores obtained from the 53-item clerical test are grouped in class intervals of two units in column 1 of this table. The distribution of frequencies

for the test results for all 133 cases, the total group, is given in column 4. The cross-tabulation of each dichotomized part with the test results is shown in columns 2 and 3. The test results for the NOT SKILLED (the SOME SKILLED combined with the NOT SKILLED) are listed in column 2 and include 70 cases. This part of the total group, the lower part, is symbolized by l. The test results for the 63 in the SKILLED group, the higher part, are given in column 3 and symbolized by h. It is apparent at once that the average test results for the SKILLED group are somewhat higher than the results for the NOT-SKILLED group, which of course indicates a positive correlation between these results and the criterion ratings. The problem now is to determine the *degree* of this correlation.

The initial figures for the mean and standard deviation of the test results for the total group are indicated in columns 5, 6, and 7, and the final computations are shown at the bottom of the table. M_t is equal to 35.12 and σ_t is equal to 8.86.

The initial figures for the mean test score of the part of the total group with the higher ratings are given in columns 8 and 9, and, as indicated at the bottom of the table, M_h is equal to 40.48.

There remains only to determine the values of p_h and y. These computations are shown at the bottom of the table, and the two values are .47 and .398 respectively.

The correlation coefficient is now computed and is found to be .72. This then is an index of the degree of relationship between the group's results on the test of clerical proficiency and the independent criterion ratings of their efficiency or skill for this type of clerical work. Such a coefficient for this type of correlation is called a *validity* coefficient of a test. It is the index of the validity of this test in differentiating the skilled and not-skilled clerical workers for technical correspondence of the type under study. The coefficient, .72, is satisfactory for an achievement or proficiency type of test. It shows that this test should be of considerable aid in the immediate differentiation of SKILLED and NOT-SKILLED applicants for such positions; the test should differentiate those who will need training from those who can go right to work. Only in the case of perfect correlation, however, would it be possible to make such a differentiation with no errors. Thus columns 2 and 3 in Table 10:2 show that a few of the NOT-SKILLED group made fairly high scores on the test; 9 out of the 70 had scores greater than 38. However, two-thirds of the SKILLED group scored above 38. None of this group had scores below 25, whereas more than 20% of the NOT-SKILLED group scored less than this. Between the test scores of 25 and 38 there is considerable overlapping; consequently, in using the test, those who received scores within this range are better classified as *doubtful* rather than as SKILLED or UNSKILLED.

Before proceeding with the discussion of biserial r, we wish to emphasize that if there is little or no difference between the means of the distributed variable of the dichotomized parts of the whole (or between the means of

either part and the whole), there is obviously no basis for correlation. Many problems for which a biserial r could be calculated are analyzed by comparing the means of the two groups. If the difference between the means is zero or insignificant (see Chapter 14, Sections E and F, for the implications of a *significant difference between means* in sampling theory), it follows that the correlation is zero and does not need to be calculated. On the other hand, when, as in the preceding example, the difference between the means is considerable, there is a basis for some degree of correlation, and biserial r is a convenient coefficient for indexing the degree of the relationship. More information can usually be obtained from biserial r than from only the means of the dichotomized parts of a distribution.

The Ordinate Values of Table 10:3

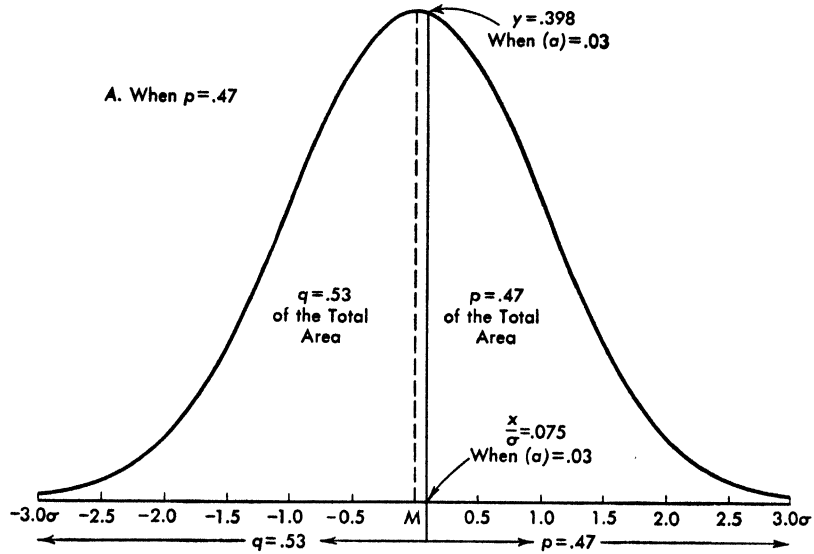
As indicated at the bottom of Table 10:2, the value for y in computing biserial r was obtained from the ordinate values in Table 10:3. This table

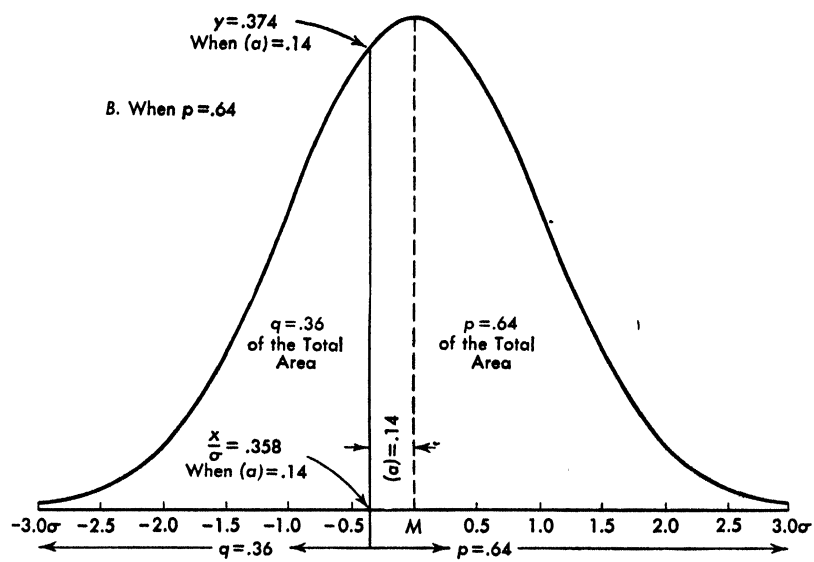
Table 10:3. Deviates (x/σ) in Terms of σ Units and Ordinates (y) for Given Areas Measured from the Mean of a Normal Distribution Whose Total Area = 1.00 *

Area from the Mean a	x/σ z	Ordinates y	Area from the Mean a	x/σ z	Ordinates y
.00	.000	.399			
.01	.025	.399	.26	.706	.311
.02	.050	.398	.27	.739	.304
.03	.075	.398	.28	.772	.296
.04	.100	.397	.29	.806	.288
.05	.126	.396	.30	.842	.280
.06	.151	.394	.31	.878	.271
.07	.176	.393	.32	.915	.262
.08	.202	.391	.33	.954	.253
.09	.228	.389	.34	.995	.243
.10	.253	.386	.35	1.036	.233
.11	.279	.384	.36	1.080	.223
.12	.305	.381	.37	1.126	.212
.13	.332	.378	.38	1.175	.200
.14	.358	.374	.39	1.227	.188
.15	.385	.370	.40	1.282	.176
.16	.412	.366	.41	1.341	.162
.17	.440	.362	.42	1.405	.149
.18	.468	.358	.43	1.476	.134
.19	.496	.353	.44	1.555	.119
.20	.524	.348	.45	1.645	.103
.21	.553	.342	.46	1.751	.086
.22	.583	.337	.47	1.881	.068
.23	.613	.331	.48	2.054	.048
.24	.643	.324	.49	2.326	.027
.25	.675	.318	.50	∞	.000

* See also Table I, Appendix B.

Fig. 10:1. Illustrating the Dichotomy on a Normal, Bell-Type Distribution for $p = .47$ and $p = .64$. The ordinate values of y and the abscissa values of x in terms of x/σ are taken from Table 10:3 for the given areas, (a), in terms of their proportions of the whole.





gives both the z score distance from the mean (second column) and the ordinate values, y (third column), at any point for areas, (a), of the normal distribution taken with respect to the mean (first column). The area is given in terms of *proportions* of the area from the mean, beginning at the mean itself, where (a) is equal to .00, and ranging to the theoretical limit of a proportion of .50, above or below the mean. This latter distance is infinite from the mean and the ordinate value is zero.

The total area of the normal distribution is taken as equal to 1.00. The ordinate value, y, at the mean of such a distribution is equal to .399. In Table 10:2, and also in A of Fig. 10:1, .47 of the total group was in the higher part; consequently, we need to locate a point above the mean that divides the total distribution into two parts, viz., .47 and .53. This will be an (a) value of .03 (see Table 10:3), since .50 − .47 = .03. Reading across the table for this value of (a) gives the value of y, the ordinate, as .398.

If, as in Table 10:5, the proportionate size of the higher group is greater than .50, the point on the curve that divides the total group into its two parts is *below* the mean of the distribution. When p is equal to .64, as in B in Fig. 10:1, the ordinate value, y, is at a point below the mean that includes .14 of the area, since .50 of the area lies above the mean. In Table 10:3, this will be an (a) value of .14. Reading across the table for this value, we see that y is equal to .374.

Values for Biserial r *from Table 10:4*

The computations required for the final term of Formulas 10:4 and 10:4a for biserial r can be facilitated by referring to Table 10:4. The proportionate values of the frequencies for each dichotomized part in biserial correlation are given in the first two columns headed p and q respectively. The product of p and q in column 3 is useful for a number of problems but is not needed for Formulas 10:4 and 10:4a, since the ratio, pq/y, needed for Formula 10:4 is given in column 5, and the ratio p/y, needed in Formula 10:4a—the one used for Table 10:2—is given in column 4.

The value for p, the proportionate number of frequencies in the higher group in Table 10:2, was .47. Locating this value near the bottom of the first column in Table 10:4 and reading across the row to column 4, we see that the ratio of p/y is equal to 1.1815. This checks with the value obtained in Table 10:2, where the ratio of p_h to y was computed. Columns 6 and 7 of Table 10:4 are useful in computing point-biserial r, described in the next section.

It will be observed that the values for Table 10:4 run only to $p = .50$. If the proportionate size of the higher group in a biserial correlation is greater than .50, the table may still be used by reversing the p and q values, or by employing Formula 10:4 for biserial r. If the p and q values are reversed,

Table 10:4. Values Employed in the Determination of Biserial and Point-Biserial Correlations *

(1) p	(2) q	(3) pq	(4) $\frac{p}{y}$	(5) $\frac{pq}{y}$	(6) $\sqrt{pq}$	(7) $\sqrt{\frac{p}{q}}$
.01	.99	.0099	.3745	.3700	.0994	.1005
.02	.98	.0196	.4132	.3935	.1380	.1428
.03	.97	.0291	.4412	.4264	.1703	.1758
.04	.96	.0384	.4640	.4452	.1959	.2042
.05	.95	.0475	.4850	.4605	.2179	.2293
.06	.94	.0564	.5038	.4736	.2375	.2526
.07	.93	.0651	.5212	.4844	.2551	.2744
.08	.92	.0736	.5380	.4950	.2713	.2950
.09	.91	.0819	.5542	.5044	.2862	.3145
.10	.90	.0900	.5698	.5129	.3000	.3333
.11	.89	.0979	.5851	.5207	.3129	.3416
.12	.88	.1056	.6000	.5278	.3249	.3693
.13	.87	.1131	.6147	.5347	.3363	.3865
.14	.86	.1204	.6289	.5410	.3470	.4035
.15	.85	.1275	.6432	.5469	.3571	.4201
.16	.84	.1344	.6576	.5523	.3666	.4365
.17	.83	.1411	.6717	.5574	.3756	.4525
.18	.82	.1476	.6860	.5627	.3842	.4685
.19	.81	.1539	.7001	.5670	.3923	.4844
.20	.80	.1600	.7143	.5714	.4000	.5000
.21	.79	.1659	.7287	.5758	.4073	.5156
.22	.78	.1716	.7430	.5793	.4142	.5311
.23	.77	.1771	.7576	.5832	.4208	.5465
.24	.76	.1824	.7720	.5868	.4271	.5620
.25	.75	.1875	.7867	.5900	.4330	.5773
.26	.74	.1924	.8015	.5929	.4386	.5928
.27	.73	.1971	.8167	.5960	.4439	.6082
.28	.72	.2016	.8318	.5989	.4490	.6236
.29	.71	.2059	.8472	.6016	.4538	.6391
.30	.70	.2100	.8628	.6037	.4582	.6547
.31	.69	.2139	.8787	.6062	.4625	.6703
.32	.68	.2176	.8949	.6086	.4665	.6860
.33	.67	.2211	.9114	.6107	.4702	.7018
.34	.66	.2244	.9279	.6125	.4737	.7178
.35	.65	.2275	.9449	.6143	.4770	.7338
.36	.64	.2304	.9623	.6159	.4800	.7500
.37	.63	.2331	.9799	.6173	.4828	.7664
.38	.62	.2356	.9979	.6187	.4854	.7829
.39	.61	.2379	1.0164	.6200	.4877	.7996
.40	.60	.2400	1.0355	.6214	.4899	.8165
.41	.59	.2419	1.0548	.6222	.4918	.8336
.42	.58	.2436	1.0744	.6230	.4935	.8509
.43	.57	.2451	1.0947	.6241	.4951	.8686
.44	.56	.2464	1.1156	.6247	.4964	.8864
.45	.55	.2475	1.1369	.6254	.4975	.9045
.46	.54	.2484	1.1590	.6258	.4984	.9230
.47	.53	.2491	1.1815	.6262	.4991	.9417
.48	.52	.2496	1.2048	.6265	.4996	.9508
.49	.51	.2499	1.2287	.6266	.4999	.9802
.50	.50	.2500	1.2534	.6266	.5000	1.0000

* This table was developed by E. K. Taylor of the Adjutant General's Office, and is reproduced by permission.

Formula 10:4a must be changed, and the mean of the lower part must be substituted for the mean of the higher part, as follows:

$$r_{bi} = \left(\frac{M_t - M_l}{\sigma_t}\right)\left(\frac{p_l}{y}\right) \qquad [10:4b]$$

Biserial correlation changed for use with Table 10:3 when $p > .50$

The use of this formula will be illustrated in the following section.

Biserial Correlation of a Vocabulary Test Item with the Total Test Score

A vocabulary test * consisting of 80 multiple-choice items and designed to measure general vocabulary knowledge at a high level of difficulty was administered to 181 college students. Each test item was scored as correct or incorrect. The "best" items will be those that most consistently differentiate students with good vocabulary ability from students with only poor vocabulary ability. However, no really independent ratings of this ability are usually available for evaluating the validity of vocabulary test items, and consequently they are often validated *internally*, that is, against the total test score taken as the criterion of the ability. If then the total score on the vocabulary test is assumed to measure vocabulary ability, its use as a criterion will permit an item analysis of the test from which the relative adequacy of each item can be determined. The items that correlate highest with the criterion will be "best," and those that correlate least with the criterion will be "poorest."

In a sense, this problem reverses the position of the criterion in the biserial correlation in Table 10:2, for this time the variable is the criterion. The dichotomy is the result obtained for *one* test item, scored as *correct* or *incorrect*, as follows:

Test Item:

__5__ *basal:* 1. mean 2. malcontent 3. sly 4. justifiable 5. essential

The data for the 181 students' results are cross-tabulated for biserial correlation in columns 2 and 3 of Table 10:5. There is considerable overlapping in the total test scores of those who responded correctly (column 3) and those who responded incorrectly (column 2) to the particular item. However, a somewhat greater proportion of those with higher total test scores did answer the item correctly. Thus, the 4 students with a total test score of 70 or better answered the item correctly; 7 out of 8 with a total test score of 67 to 69 answered it correctly, etc. Some degree of correlation is observable from an inspection of the results, and it is positive. The means of both the lower and the higher groups are computed in Table 10:5. The biserial coefficient is found to be .55 by Formulas 10:4, 10:4a, or 10:4b, as follows:

* A special vocabulary test administered by the author to students in general psychology classes at the College of the City of New York.

Table 10:5. Biserial Correlation of a Vocabulary Test Item with the Total Test Score

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Total Test Scores	Responses to Item: Incorrect f_l	Responses to Item: Correct f_h	Total Group f_t	h', l', t'	ft'	ft'^2	fl'	fh'
70–72		4	4	7	28	196		28
67–69	1	7	8	6	48	288	6	42
64–66	2	7	9	5	45	225	10	35
61–63	5	14	19	4	76	304	20	56
58–60	3	16	19	3	57	171	9	48
55–57	6	13	19	2	38	76	12	26
52–54	5	16	21	1	21	21	5	16
49–51	3	18	21	0	0	0	0	
46–48	7	8	15	−1	−15	15	−7	−8
43–45	7	6	13	−2	−26	52	−14	−12
40–42	8	2	10	−3	−30	90	−24	−6
37–39	5	3	8	−4	−32	128	−20	−12
34–36	8	2	10	−5	−50	250	−40	−10
31–33	1	0	1	−6	−6	36	−6	
28–30	4	0	4	−7	−28	196	−28	
	$N_l = 65$	$N_h = 116$	$N_t = 181$		313	2048	62	251
					−187		−139	−48
					$\Sigma = 126$		−77	203

$M_t = 50.0 + 3(126/181) = 52.09$

$\sigma_t = 3\sqrt{2048/181 - (126/181)^2} = 9.87$

$M_h = 50.0 + 3(203/116) = 55.25$

$M_l = 50.0 + 3(-77/65) = 46.45$

$p_h = 116/181 = .641$

$y = .374$ (From Table 10:3, $a = .14$)

$pq/y = .616$ (From Table 10:4, where p_h is taken as $q = .64$)

$p_l/y = .962$ (From Table 10:4, $p_l = .36$)

By Formula 10:4:

$$r_{bi} = \left(\frac{M_h - M_t}{\sigma_t}\right)\left(\frac{pq}{y}\right)$$

$$= \left(\frac{55.25 - 46.45}{9.87}\right)\left(\frac{(.641)(.359)}{.374}\right) = (.892)(.616) = .55$$

the value for the ratio pq/y, .616, being obtained from column 5 of Table 10:4.

By Formula 10:4a:

$$r_{bi} = \left(\frac{M_h - M_t}{\sigma_t}\right)\left(\frac{p_h}{y}\right)$$

$$= \left(\frac{55.25 - 52.09}{9.87}\right)\left(\frac{.641}{.374}\right) = (.320)(1.714) = .55$$

The value of the ratio p_h/y must be computed; the value of y, .374, is obtained from Table 10:3.

By Formula 10:4b:

$$r_{bi} = \left(\frac{M_t - M_l}{\sigma_t}\right)\left(\frac{p_l}{y}\right)$$

$$= \left(\frac{52.09 - 46.45}{9.87}\right)\left(\frac{.359}{.374}\right) = (.571)(.962) = .55$$

the value of the ratio p_l/y, .962, being obtained from column 4 of Table 10:4.

A test item correlation of .55 with a criterion, whether internal (as in this case) or external and independent, is fairly satisfactory in that such an item does result in some differentiation between those with more ability and those with less ability in the attribute under consideration. If the vocabulary test were to be shortened to the best 50 items, this is one that might be retained. It is unlikely that there would be 50 or more other items that would yield a higher correlation with the criterion. Thus, by means of biserial r, a validity coefficient can be obtained for each item of a test, and the most effective items can be selected for any abbreviation or revision of the test itself.

Point-Biserial Correlation

We saw that the preceding method for the biserial correlation of one continuously distributed variable with a dichotomized variable was based on the assumption not only that the relationship is linear but also that the dichotomized variable is in reality a quality or attribute that would yield a normal, bell-shaped distribution if it could be measured on a continuous scale. This latter assumption is the most reasonable one that can be made for the dichotomies in the examples in Tables 10:2 and 10:5, even though the ratings of skill in technical correspondence are admittedly not on a continuous scale, and it may be difficult to see how responses to a test item can ever yield a normal distribution, since such responses are usually scored as *correct* or *incorrect*. Nevertheless, normality is usually assumed for test items provided the test itself, as a whole, yields a normal distribution of test results. This assumption is based on the argument that the form of the whole is derived from the form of each of its parts. From this it follows that if the whole test yields a distribution of the normal type, each item in it would necessarily yield a normal distribution if the responses to it could be differentiated on a sufficiently fine scale.

Situations often arise, however, in which there is no logical justification for assuming the normality of the dichotomized attribute in biserial correlation. One of the attributes may be in the form of a true dichotomy, as was true of some of those described in Chapters 2–4—for example, the dichotomy male and female. Although sex differences in results for a continuously distributed variable, such as intelligence test scores, are usually analyzed by

comparing the means of each sex group, we have already indicated that if there is a mean difference, some degree of correlation exists between the dichotomy and the variable. The smaller this difference, the less the degree of correlation; and conversely, the greater the difference, the greater the degree of correlation.

Questionnaire results in market research investigations often yield true dichotomies. Thus, the following questions of fact will be answered by *Yes* or *No:* Do you own an automobile? Do you own a piano? Do you have any brothers or sisters now living? Have you ever traveled in an airplane? Do you reside in San Francisco? The correlation of answers to such questions with the age, income status, education, intelligence, etc., of the respondents often gives insights into the relationships between a character of a population and a product. The classification of adults into the MARRIED and NOT MARRIED, PARENTS and NOT PARENTS, also represents true dichotomies.

Fortunately, there is available a method of correlation that can be employed to correlate two attributes, one of which is continuously distributed, the other being a true dichotomy, or a dichotomized variable that is not normally distributed. The assumption is made that a linear function is adequate to describe the relationship. As in biserial r, the continuous variable is not assumed to be normally distributed. The coefficient obtained is called *point-biserial r* and is equal to the following: *

$$r_{pt\text{-}bi} = \left(\frac{M_P - M_Q}{\sigma_t}\right)\sqrt{pq} \qquad \begin{matrix}[10:5] \\ \text{Point-biserial } r\end{matrix}$$

The symbols here have the same meaning as those in Formula 10:4, except that for a true dichotomy "higher" and "lower" have no meaning, and hence one part of the dichotomy is symbolized as P and the other part as Q.

As Formula 10:4 was simplified to Formula 10:4a and 10:4b for biserial r, point-biserial r may also be obtained from measures derived from the total distribution and only one of the dichotomized parts, as follows:

$$r_{pt\text{-}bi} = \left(\frac{M_P - M_T}{\sigma_t}\right)\sqrt{\frac{p}{q}} \qquad \begin{matrix}[10:5a] \\ \text{Point-biserial } r\end{matrix}$$

The formula for point-biserial r is thus not very different from that for biserial r. The first ratio is identical in both Formulas 10:4 and 10:5, but the second ratio is different. Thus, in Formula 10:4, this ratio is based on the ordinate value y of the normal, bell-shaped curve at the point that divides the area of the distribution into two parts corresponding to the proportions of the total group in the respective divisions of the dichotomy. In Formula 10:5, however, the second ratio is based on the proportionate parts of the dichotomized whole, p and q.

---

* M. W. Richardson and J. M. Stalnaker, "A Note on the Use of Biserial r in Test Research," *Journal of General Psychology*, 8:463, 1933.

It should be emphasized that point-biserial r is written without a sign when the dichotomy is a true one and "higher" and "lower" therefore have no meaning. Under such circumstances, which part of the dichotomy is called P and which part Q is a purely arbitrary decision. If the value of the mean of part P of the variable distribution proves to be less than this value for the total distribution, the result will presumably be negative because of the order of M_P and M_T in Formula 10:5a. In such cases the sign is ignored and the result is interpreted by verbalizing the way in which the dichotomized attribute is related to the variable attribute.

Triserial, Quadriserial, and Quintiserial r

Jaspen * has recently developed formulas for *serial* correlation which are useful for determining the linear correlation between a continuously distributed variable and a variable segmented or classified into a few broad classes. Although he gives the formula for serial correlation in general, including biserial r, we shall present only his formulas for:

1. *Triserial r*, for the linear correlation of a continuously distributed variable with a trichotomized variable.
2. *Quadriserial r*, for the linear correlation of a continuously distributed variable with a variable segmented into four classes.
3. *Quintiserial r*, for the linear correlation of a continuously distributed variable with a variable segmented into five classes.

As in the case of biserial r, each of these coefficients is based on the assumption that the data of the segmented variable are derived from a variable that is normally distributed. We shall explain the computation of triserial r, using Cureton's data for the clerical proficiency test (dichotomized in Table 10:2), and then for reference present Jaspen's formulas for quadriserial and quintiserial r.

Triserial r

The procedure for computing triserial r is similar to that used for biserial r in that the means of each segmented class of the continuously distributed variable are compared. The greater the difference between these means, the greater the correlation; the less the difference, the less the correlation. The formula for triserial r is as follows:

$$r_{tri} = \frac{y_h M_h + (y_c - y_h)M_c - y_c M_l}{\sigma_t \left[\frac{y_h^2}{p_h} + \frac{(y_c - y_h)^2}{p_c} + \frac{y_c^2}{p_l}\right]} \qquad \begin{matrix}[10:6] \\ \text{Triserial } r\end{matrix}$$

* Nathan Jaspen, "Serial Correlation," *Psychometrika*, 11:23–30, 1946. I wish to thank the author, and Dr. H. O. Gulliksen, the editor of *Psychometrika*, for letting me see this article prior to publication and giving me permission to incorporate the formulas in this chapter. I have changed some of the symbols in Jaspen's formulas to make them uniform with those used in this book.

where

l, c, and h symbolize the three classes of the segmented variable, i.e., l = the lowest part, c = the central (or middle) part, and h = the highest part.

M_l, M_c, and M_h are the respective means of each of the segmented groups of the continuously distributed variable.

σ_t is the standard deviation of the continuously distributed variable.

p_l, p_c, and p_h are the respective proportions of the total group in each of the three segmented classes.

y_l, y_c, and y_h are the respective values of the ordinates on a normal curve at points that divide the total distribution into three parts, with the proportion of the area above the upper point of division equal to p_h, and the proportion of the area between the lower and upper point of division equal to p_c. (The ordinate value for p_l is always zero because this "point" is at the lowest end of the distribution.)

The computation of triserial r is illustrated in Table 10:6. The scores for the continuously distributed clerical proficiency test results are listed in the first column as they were for the computation of biserial r in Table 10:2. The trichotomized criterion ratings are given in columns 2, 3, and 4. The test results for persons with NO SKILL ratings are distributed in column 2; for those with ratings of SOME SKILL, in column 3; and for the SKILLED, in column 4. The distribution for the total group is given in column 5. The mean for the lowest group (NO SKILL) is obtained by the short method (see Table 7:6) and shown in columns 6 and 7. Similarly, the mean for the central (or middle) group (SOME SKILL) is shown in columns 8 and 9, and for the highest group (SKILLED) in columns 10 and 11. The mean and standard deviation of the total group are obtained by the short method (see Table 7:10) and shown in columns 12, 13, and 14. The ordinate values for y at the bottom of the table are obtained from Table 10:3 (or from Table I, Appendix B).

The value of triserial r for the correlation between the criterion ratings and the clerical proficiency test results is .71, which practically coincides with the value of biserial r for these data. In this case therefore, triserial r did not give a result any different from that obtained by biserial r; the mean difference between the NO SKILL and the SOME SKILL subgroups is not sufficiently great to make a difference in the value of r.

Quadriserial r

Jaspen's formula for quadriserial r is as follows:

$$r_{quad} = \frac{y_h M_h + (y_d - y_h)M_d + (y_b - y_d)M_b - y_b M_l}{\sigma_t \left[\frac{y_h^2}{p_h} + \frac{(y_d - y_h)^2}{p_d} + \frac{(y_b - y_d)^2}{p_b} + \frac{y_b^2}{p_l}\right]} \qquad [10:7]$$

Quadriserial r

where l, b, d, and h represent the four parts of the segmented variable: l the lowest, b the next lowest, d the next highest, and h the highest part. M, y,

Table 10:6. Triserial Correlation of Clerical Proficiency Test Results with Criterion Ratings Trichotomized

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
Test Scores	Criterion Ratings			Total Group	No Skill		Some Skill		Skilled		Total Group		
	No Skill	Some Skill	Skilled										
	f_l	f_c	f_h	f_t	l'	fl'	c'	fc'	h'	fh'	t'	ft'	ft'^2
51–52			1	1					7	7	9	9	81
49–50		1	5	6			8	8	6	30	8	48	384
47–48		0	7	7			7	0	5	35	7	49	343
45–46		3	8	11			6	18	4	32	6	66	396
43–44		2	4	6			5	10	3	12	5	30	150
41–42	1	1	8	10	7	7	4	4	2	16	4	40	160
39–40	1	0	9	10	6	6	3	0	1	9	3	30	90
37–38	2	4	7	13	5	10	2	8	0	0	2	26	52
35–36	2	5	2	9	4	8	1	5	−1	−2	1	9	9
33–34	3	3	3	9	3	9	0	0	−2	−6	0	0	0
31–32	1	3	2	6	2	2	−1	−3	−3	−6	−1	−6	6
29–30	4	4	4	12	1	4	−2	−8	−4	−16	−2	−24	48
27–28	2	4	1	7	0	0	−3	−12	−5	−5	−3	−21	63
25–26	6	3	2	11	−1	−6	−4	−12	−6	−12	−4	−44	176
23–24	3	0		3	−2	−6	−5	0			−5	−15	75
21–22	1	0		1	−3	−3	−6	0			−6	−6	36
19–20	3	3		6	−4	−12	−7	−21			−7	−42	294
17–18	3	1		4	−5	−15	−8	−8			−8	−32	256
15–16	1			1	−6	−6					−9	−9	81
	N = 33	37	63	133		Σ = −2		Σ = −11		Σ = 94		Σ = 108	Σ = 2700

$M_l = 27.5 + 2(-2/33) = 27.4$
$M_c = 33.5 + 2(-11/37) = 32.9$
$M_h = 37.5 + 2(94/63) = 40.5$

$M_t = 33.5 + 2(108/133) = 35.1$

$\sigma_t = 2\sqrt{(2700/133) - (108/133)^2} = 8.86$

$p_l = 33/133 = .25$
$p_c = 37/133 = .28$
$p_h = 63/133 = .47$

$y_l = 0$
$y_c = .318$
$y_h = .398$

From Table 10:3 (or Table I, Appendix B)

$$r_{tri} = \frac{.398(40.5) + (.318 - .398)32.9 - .318(27.4)}{8.86\left[\frac{(.398)^2}{.47} + \frac{(.318 - .398)^2}{.28} + \frac{(.318)^2}{.25}\right]}$$

$$= \frac{16.12 + (-2.63) - 8.71}{8.86(.34 + .02 + .40)} = \frac{4.78}{6.73} = .71$$

and p are the means, ordinates, and proportions, as in Formula 10:6, and σ_t is the standard deviation of the total distribution of the continuously distributed variable.

Quintiserial r

Jaspen's formula for quintiserial r is as follows:

$$r_{quint} = \frac{y_h M_h + (y_d - y_h)M_d + (y_c - y_d)M_c + (y_b - y_c)M_b - y_b M_l}{\sigma_t \left[\frac{y_h^2}{p_h} + \frac{(y_d - y_h)^2}{p_d} + \frac{(y_c - y_d)^2}{p_c} + \frac{(y_b - y_c)^2}{p_b} + \frac{y_b^2}{p_l}\right]} \quad [10:8]$$

Quintiserial r

where l, b, c, d, and h represent the five parts of the segmented variable: l the lowest, b the next lowest, c the central or middle, d the next highest, and h the highest. M, y, and p are the means, ordinates, and proportions as in Formula 10:6, and σ_t is the standard deviation of the total distribution of the continuously distributed variable.

C. TETRACHORIC CORRELATION

Purpose of the Method

A measure of linear correlation for the cross-tabulation of the data of dichotomized variables that are normally distributed is provided by the tetrachoric correlation coefficient. However, without the diagrams prepared by L. L. Thurstone and his associates * for determining the coefficient, the method is too arduous from the computational point of view for any practical use. The complete equation for r_t involves a series with many powers of r_t. The method is valuable in two types of situations in which measures of the degree of linear correlation are needed, and fortunately, Thurstone's diagrams make it easy to use.

Both situations concern bi-variates. The first situation involves bi-variates, either or both of which yield dichotomized data, or coarse groupings of results into a few classes that can be readily combined to make a dichotomy. These are the kinds of data characteristic of many market research investigations and of the social psychologists' studies of attitudes, preferences, etc. In the second situation, all the bi-variates yield continuous distributions of quantitative data, but a labor-saving device is required when many correlations are to be computed. Such occasions arise whenever all the inter-correlations between five or more variables must be obtained, or a series of test items must be correlated with a dichotomized criterion. For example, in a correlational analysis based on 10 variables, 45 inter-correlation coefficients must be computed (cf. Chapter 18). If each of these 10 variables is dichotomized near the median or mean of its distribution and the data are cross-tabulated into

* L. Cheshire, M. Saffir, L. L. Thurstone, *Computing Diagrams for the Tetrachoric Correlation Coefficient*, Univ. of Chicago Bookstore, Chicago, 1933.

2 by 2 (fourfold) tables, the coefficients can be readily obtained with Thurstone's diagrams. Also, in the item analysis of a test, n correlation coefficients will be needed; thus in a test of 100 items, 100 coefficients must be computed.

The Computation of Tetrachoric r (r_t)

If the data to be correlated are already dichotomized, the first step in computing r_t is the cross-tabulation, in a fourfold table, of the two variables to be correlated. If, on the other hand, r_t is being used as a labor-saving device in estimating r for two continuously distributed variables, the first step is to dichotomize each variable near its median or mean so as to put the data in a form suitable for cross-tabulation in a 2 by 2 correlation matrix.

We shall first discuss the computation of r_t in this latter situation, using the data in Table 9:1. The product-moment correlation of these data gave a coefficient of .67 for the relationship between the heights and the weights of 151 infants. The mean weight was 21.8 pounds and the mean height was 29.4 inches. Both distributions are first dichotomized at convenient points near their respective means. The height variable is dichotomized at the lower limit of the class interval which includes the mean, viz., 29.25 inches. The weight variable is dichotomized at 21.25 pounds. The results of cross-tabulating the data of these variables into a fourfold table are shown in Table 10:7.

Table 10:7. The Dichotomization for Tetrachoric Correlation of the Height-Weight Measurements of 151 Infants

(Original data in Table 9:1; $M_{\text{Weight}} = 21.8$ pounds, $M_{\text{Height}} = 29.4$ inches)

Weight (Variable 2) \ Height (Variable 1)	Below Average	Above Average	n_W
Above Average	a 16	b 60	76
Below Average	c 54	d 21	75
n_H	70	81	N = 151

[10:9]

$$r_t = \frac{bc - ad}{y_1 y_2 N^2} - \frac{z_1 z_2}{2} r_t^2$$

Tetrachoric coefficient for linear correlation of dichotomized bi-variates

where

a is the number of frequencies in a cell a (Quadrant II).

b is the number of frequencies in cell b (Quadrant I).

c is the number of frequencies in cell c (Quadrant III).

d is the number of frequencies in cell d (Quadrant IV).

y_1 is the ordinate value of the first variable at the point at which the distribution was dichotomized (from Table 10:3 or Table I, Appendix B).

y_2 is the ordinate value of the second variable at the point of dichotomization.

N is the total number of cases or instances cross-tabulated for correlation.

z_1 is the deviate distance, in terms of x/σ_x, of the point of dichotomization from the mean of the first distribution (from Table 10:3, or Table I, Appendix B).

z_2 is the deviate distance of the point of dichotomization from the mean of the second distribution.

For the height-weight data dichotomized and cross-tabulated in Table 10:7, the values for the solution of tetrachoric r are as follows:

$a = 16.$
$b = 60.$
$c = 54.$
$d = 21.$
$y_1 = .397$ (from Table 10:3, where $p_1 = \frac{81}{151} = .54$ and a therefore equals .04).
$y_2 = .399$ (from Table 10:3, where $p_2 = \frac{76}{151} = .50^+$ and a therefore equals $.00^+$).
$N = 151.$
$z_1 = .090$ (from Table 10:3, where $(a) = .036$).*
$z_2 = .010$ (from Table 10:3, where $(a) = .003$).*

Substituting these values in Formula 10:9 gives the following:

$$r_t = \frac{(60)(54) - (16)(21)}{(.397)(.399)(151)^2} - \frac{(.090)(.010)}{2} r_t^2$$

$$= \frac{2904}{3611.746803} - .00045\, r_t^2 = .804043 - .00045\, r_t^2$$

Expressed as a quadratic equation, this becomes:

$$.00045 r_t^2 + r_t - .804043 = 0$$

* These values of $z_1 = .090$ and $z_2 = .010$ are interpolated from Table 10:3 for (a) values of .036 and .003 respectively. It will be observed that if either z_1 or z_2 has a value of zero the second term of formula 10:9 will be zero, and consequently the solution for r_t is readily obtained from the simple equation. In the above example the value of the r_t^2 term is practically zero and its retention in the equation actually does not affect the value of r_t to two decimal places.

For this problem, the values of the coefficients are: $b = 1.0$; $a = .00045$; and $c = -.804043$. This quadratic equation must be solved for r. The r term of the usual formulation of a quadratic, viz.,

$$ar^2 + br + c = 0$$

is equal to:

$$r = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Substituting the values for a, b, and c, we find the r term is:

$$r_t = \frac{-1.0 \pm \sqrt{1.0 - 4(.00045)(-.804043)}}{2(.00045)}$$

One value for this equation is

$$\frac{-1.0 + 1.00072}{.0009} = \frac{+.00072}{.0009} = .80$$

The other value for this equation is

$$\frac{-2.00072}{.0009}$$

This value for r_t is absurd, since r cannot exceed ± 1.00.

Thus, the tetrachoric coefficient, r_t, is found to be .80. This is somewhat higher than the value of .67 obtained by the method of product-moment correlation in Fig. 9:13. A higher value is to be expected with Formula 10:9 which, as we pointed out, is a simplification of the complete equation for r_t. The value of r_t is even higher when Thurstone's diagrams instead of Formula 10:9 are used for these data; this is shown in the next section.

Estimating Tetrachoric Correlation with Thurstone's Diagrams

In order to use Thurstone's *Diagrams* in determining a tetrachoric coefficient, three proportions are needed: one for cell c of the 2 by 2 cross-tabulation, and one each for the total number of cases in the lower, or below-average part of each dichotomized variable. These are shown in bold-face type in Table 10:8 as (**a**), (**b**), and (**c**), where (**a**) is the proportion of cases for the first variable whose measures are below the mean; (**b**), the proportion of cases for the second variable whose measures are below the mean; and (**c**), the proportion of the total group whose cross-tabulated measures are in cell c, i.e., those below average on both variables. These symbols (**a**), (**b**), and (**c**) correspond to those used in Thurstone's *Diagrams*.

In computing the three proportions needed, sufficient accuracy is obtained by carrying the calculations to three decimal places and then rounding off to two places. In order to have an independent check on the accuracy of these computations, it is well to calculate the p values of all four cells and compare the sums of the proportions for the rows and columns with the marginal totals at the bottom and right of the fourfold table. This has been done in

Table 10:8, in which the frequency values in Table 10:7 have been converted into p values. The coefficient, r_t, is given as .74 in Thurstone's *Diagrams*—a somewhat higher value than the $r = .67$ obtained by Pearson's product-moment method.

Table 10:8. The Cross-Tabulated Data of Table 10:7 Expressed as Proportions of *N*

(For Tetrachoric Correlation by Thurstone's *Diagrams*)

		Infants' Height (Variable 1)		
		Below Average	Above Average	
Infants' Weight (Variable 2)	Above Average	*a* .10	*b* .40	.50
	Below Average	*c* .36 (c)	*d* .14 (c′)	.50 (b)
		.46 (a)	.54 (a′)	1.00 (*N* = 151)

$r_t = .91$ (from *Diagrams*)

The tetrachoric correlation method does not give a satisfactory estimate of linear correlation if the p value of either part of a dichotomized variable is less than .05. It is for this reason that Thurstone does not present diagrams for (**a**) values of less than .05.

Thurstone's *Diagrams* provide a separate page for different values of (**a**), beginning with $p = .05$ and ending with $p = .50$. If the value of (**a**) is greater than .50, then (**a′**) and (**c′**) are used in estimating r_t (cf. Table 10:8). Whenever this change to (**a′**) must be made, the sign for the value of r_t indicated in the appropriate diagram must be changed; if it is negative, it is written as positive. This is the case because the diagrams are set up for the p values of Quadrant III (positive), whereas the p value for cell *d* is for Quadrant IV, which is negative.

EXERCISES

1. Describe three different problems for which Spearman's rank-difference method would be appropriate.
2. Under what circumstances is Spearman's rank-difference method inadequate to measure the correlation between two variables?

3. Compute a rank-difference correlation for the bi-variate data correlated in Exercise 11, Chapter 9. Compare the value obtained for *rho* with that previously obtained for *r*.

1. With the data in Table 5:14, dichotomize the average grade scores of the college freshmen at the mean of the distribution, and obtain a biserial correlation between their intelligence test scores and their dichotomized grade scores.

5. Dichotomize the achievement ratings in Table 10:1 into two groups equal in size, and obtain the biserial correlation between the attitude test scores and the dichotomized achievement ratings.

6. Use the following dis ribution of data to compute the point-biserial correlation between marital status and the gross amount of life insurance sold over a period of a year:

Marital Status of Insurance Salesmen and Total Amount of Life Insurance Sold (in $1000's) over a Period of One Year (M = Married; S = Single)

Salesmen's Status	Insurance Sold	Salesmen's Status	Insurance Sold	Salesmen's Status	Insurance Sold
M	234	S	498	M	632
S	472	M	671	M	599
S	557	M	618	M	651
S	544	M	810	M	723
M	456	S	530	M	410
M	331	S	448	S	670
M	898	S	314	S	617
M	676	M	592	S	769
M	641	S	803	S	862
M	520	M	533	S	550
M	648	S	526	M	375
S	408	M	624	S	426
M	960	S	435	S	547
S	273	S	493	M	746
S	208	M	901	S	582
S	621	S	712	S	804
S	549	S	520	S	237
M	871	M	416	M	759
M	715	M	550	M	577
M	768	S	573	S	154

7. Using the intelligence test scores in Table 5:14, dichotomize each distribution at its respective mean, and obtain the tetrachoric correlation coefficient between the scores of the college freshmen and of their best friends. Compare this result with that obtained in terms of *r* in Exercise 10, Chapter 9.

8. What assumptions underlie the use of:
   a. serial *r* (biserial *r*, triserial *r*, etc.)
   b. point-biserial *r*
   c. tetrachoric *r*

PART TWO

Sampling and Analytical Statistics

CHAPTER 11

Samples and Sampling Techniques

A. INTRODUCTION

We pointed out in Chapter 1 that in the development of statistical method it is useful to make a distinction between *descriptive statistics*, on the one hand, and *sampling and analytical statistics*, on the other. We indicated there that sampling and analytical statistics consist essentially in the study of statistical populations or universes in terms of the data of *samples* derived from them. In purely descriptive statistics, no distinctions are made between the sample and the universe—the part and the whole—the data being treated as if they constituted a whole.

All statistics, however, are in a basic sense descriptive, regardless of whether methods for the data to be reduced are the data of samples or the complete data of a census. The fundamental methods of descriptive statistics that have been developed in the preceding chapters can be briefly outlined as follows:

I. *For the Data of Non-Variable Attributes*

1. The classification into categories of the data of non-variable attributes, yielding dichotomous or polytomous subdivisions.
2. The enumeration of instances by category, yielding the statistical frequency.
3. The development of proportions, percentages, and ratios for summary and comparative purposes.
4. The cross-tabulation of categorical data of two or more attributes, to show the relationships between attributes.
5. The determination of the *degree* of relationship between the cross-tabulated data of two attributes by a technique of correlation.

II. *For the Data of Variable Attributes*

1. The organization into class intervals of the data of variates, yielding the frequency distribution.
2. The summarization of variable data by the centile point method, yielding centiles, vigintiles, quintiles, deciles, terciles, quartiles, etc., and the derivation, from these values, of measures of dispersion and deviation.
3. The summarization of variate data by the algebraic method of moments, yielding the arithmetic mean, the standard deviation, and the Coefficient of

Relative Variation; the derivation from these values of z score and Standard score measures, and a psychograph or profile chart, for comparative purposes.

4. The cross-tabulation of the data of two variables, yielding the correlation chart or bi-variate distribution.

5. The determination of the degree of relationship between the data of bi-variates by appropriate techniques of correlation.

III. *Appropriate Graphic Methods*

The treatment and summarization of both non-variable and variable data by graphic methods.

In the remainder of this book we shall be concerned not only with the application of these methods of descriptive statistics to the data of samples, but also with the subject matter fundamental to sampling and analytical statistics:

I. *The Techniques of Sampling*

In using statistical methods with sample data, we must know the requirements in sampling in order to apply methods of statistical analysis correctly and to draw sound inferences from the results. This is the subject matter of the present chapter.

II. *Probability Theory and Statistical Inference*

In studying statistical universes or populations by means of analyzing sample data, we utilize the implications of the mathematical theory of probability. We need to know not only what these fundamental implications are, but how to apply them to particular problems that arise in psychology and related fields. We shall see that sound statistical inference is a form of probable inference and is developed statistically in terms of what is called a *Test of Significance.* These problems of sampling and analytical statistics will be considered in Chapters 12–15.

III. *Correlational Analysis*

We have already seen in Part I that the method of correlation is fundamental in statistics for the study of relations between phenomena. Correlation is essential to the discovery of law and order in the social sciences as well as in many aspects of the biological and physical sciences. In fact, Karl Pearson * took the extreme view that nature itself is essentially statistical in character: that law and order are formulations of what is empirically observed as characteristic of the average. Correlations between attributes or qualities are taken as the basis for law and order in nature. Irrespective of Pearson's point of view, the analysis of correlative relations has come to form the most

* Karl Pearson, *Grammar of Science*, Adams and Charles Black, London, rev. ed., 1900.

exploratory statistical technique of the biological and social sciences. Hence some of the basic aspects of correlational analysis will be considered in Chapters 16–18.

Census vs. Sample

Several distinctions in the terminology of descriptive and sampling statistics were suggested in Chapter 1. We need first a set of terms that will always clearly differentiate the part from the whole. In other words, we must be able unambiguously to distinguish a sample, or part, from the whole with which it is identified. The term *sample* is used to designate the part. The whole is called the statistical *population*, or the *collective* or *statistical universe*.

Whereas the observations or measurements of a sample yield *sample data*, the observations or measurements of an entire population yield *census data*. The analytical methods used in sampling statistics are developed for treating samples in the study of populations. Not only are the methods of descriptive statistics often sufficient for the summarization and presentation of census data, but they are indispensable for reducing the data of samples.

Were it possible and feasible to do so, a census rather than a sample of population would usually be obtained for study. The only errors in a census would be *errors of observation and measurement*, for a census would include all instances or members of the group to be studied. The statistical analysis of the results would be comparatively simple, provided the measurements were made under scientific conditions of control and standardization, for under such conditions the errors in the measurements would be distributed according to known laws of probability and hence could be taken into account in interpreting the results.

The normal, bell-shaped distribution is also the normal probability curve, and is often called *the normal curve of error*. It shows the way in which purely chance errors, or their effects, are usually distributed over an extended series of observations and measurements. The controls of scientific method, long employed in the laboratories of the physicist and chemist, are techniques that have been devised to eliminate bias in a series of results. In fact, scientific method per se can well be described in terms of the *operations* for observation and measurement that yield results whose values are just as likely to be affected positively as negatively by the errors that necessarily appear in any process of measurement.

However, for reasons indicated in the next section, it is usually either impossible or not feasible to obtain a census. A census is not necessary for the scientific study of a population; if it were, there would be very little in the way of scientific knowledge. Under well-designed and controlled conditions, a sample of observations or measurements can be obtained such that the nature or character of a statistical population can be logically and soundly inferred from the study and analysis of this sampled fraction of the whole. True, *errors of sampling* will enter, but again, if the method of sampling is properly

controlled, we can be confident that they will operate as *chance errors;* that is, they will be just as likely to affect the results positively as negatively. Knowing how such errors behave, we can allow for them in interpreting our results.

Sampling Is a Research Technique

Scientific method has until recently been portrayed in psychology and the social sciences essentially as a controlled and standardized procedure for making a series of observations or measurements under unbiased conditions such that the procedure could be repeated by two or more researchers working independently of each other. But scientific method in the biological and social sciences involves more than laboratory or field control over observation or measurement. It is a technique devised to obtain the correct *sample* for study in lieu of the data of a population or universe. It is just as important, for the sound interpretation of the data of a sample, to know the conditions under which it was obtained as to know the processes used for the observations or measurements. Scientific sampling techniques are as integral to the design of an experiment as are the processes used in observation and measurement.*

This chapter, then, emphasizes an aspect of scientific method—the technique of sampling—that is fundamental to research problems, whether they be the health value of a certain vitamin, an evaluation of different learning methods, or consumer opinion about a new kind of vacuum cleaner. Sampling is integral to scientific method whenever a census is impossible or not feasible.

A Gallup Poll

Some of the preceding points can be illustrated by an example from social psychology, viz., the increasingly important field of public opinion research. Let us say we wish to know the attitude of the voters in the United States toward Great Britain's request for a post-war loan of several billion dollars. A census may appear theoretically possible, but it is obviously not feasible to interview each franchised citizen in order to obtain his answer to this question—the United States Census of 1940 cost about 50 million dollars. So we shall take a sample from the people whose opinions are to be obtained. First, we need precisely to define the population or universe to be studied. Second, we need to obtain a sample of opinions that will be *representative* of the opinions of the whole group. This representativeness cannot be perfect; but if the sample is adequate in size and properly drawn, we can allow for the chance errors that will necessarily be present. Such errors include *sampling errors,* those inherent in the operation of sampling, and *errors of observation,* those inherent in the operation of getting each respondent's opinion in an interview. Although we are here concerned with reducing sampling errors to

* Cf. in this regard, R. A. Fisher, *The Design of Experiments,* Oliver & Boyd, London, 2nd ed., 1937.

a satisfactory minimum, errors of observation also affect the result and hence must likewise be kept to a minimum by scientific controls in the technique of interviewing, the formulation of the question, the recording of each answer, and the processing of the results.*

On October 1, 1945, Dr. George Gallup, Director of the American Institute of Public Opinion, reported the public's opinion on the above question worded as follows:

> "ENGLAND PLANS TO ASK THIS COUNTRY FOR A LOAN OF THREE TO FIVE BILLION DOLLARS TO HELP ENGLAND GET BACK ON ITS FEET. WOULD YOU APPROVE OR DISAPPROVE OF THE UNITED STATES MAKING SUCH A LOAN?" †

From a carefully constructed national sample of the voters in the United States, each member of which was personally interviewed about his opinion on this question, Gallup reported the following results:

Approve	27%
Disapprove	60%
No Opinion	13%

Three-fifths of the sample were opposed to such a loan, the ratio for those holding an opinion being more than 2 to 1 against the proposition. This is the sample result. Can we have confidence in it to the extent of concluding that all the voters in the United States had a similar division of opinion on the question? It is not feasible to check such sample results against a census of the opinions of all the voters; hence any confidence in the result must be based on our knowledge of the procedures employed and the reputation of the American Institute of Public Opinion, which has been making such surveys over a period of years. The sampling methods used by the Institute in the above survey were the same as those it used to forecast, with an error no greater than 2 percentage points, the outcome of five fairly recent national elections in five countries—Australia, Canada, Great Britain, Sweden, and the United States. In the 1945 election in Great Britain, the British Institute of Public Opinion forecast the gains of the Labor party with an average error of only 1%. Thus, although we have no independent criterion against which to check the public opinion poll on the loan to Britain question, we do have a basis for confidence in the result because of the Institute's previous outstanding successes in predicting election results from samples taken by means

* The techniques of public opinion research have been recently summarized by Hadley Cantril and his research associates in the Office of Public Opinion Research at Princeton University in *Gauging Public Opinion*, Princeton University Press, 1944.

Particularly detailed attention to questionnaire methods of public opinion and market research is given by A. B. Blankenship in *Consumer and Opinion Research: The Questionnaire Technique*, Harper, New York, 1943.

† George Gallup, "Loan to Britain Opposed," New York *World-Telegram*, October 1, 1945.

of the same technique. This technique is known as stratified sampling by the quota system, and will be described in a later section.

Errors of Sampling

How about the errors that necessarily enter into Gallup's sample results? Neither Gallup, Elmo Roper, nor the others doing research on public opinion ever claim that such a result is entirely free of sampling and measuring errors. Consequently, the problem is to determine the probable effect of these errors on the percentage results reported by Gallup. Is the effect of these errors likely to have been so great that we cannot be confident that at least a majority (50%+) of all the voters would have opposed such a loan? Or is it likely to have been so small that we can be confident that at least 55% of all the voters would have opposed such a loan? Is it likely that as many as two-thirds ($66\frac{2}{3}\%$) of all the voters would have opposed it?

It would be premature to describe the methods used in answering these questions (they are discussed in the following chapters), but it is in order to emphasize the following points:

1. When sample results are based upon a scientifically controlled sampling technique, both sampling errors and errors of measurement should operate as *chance errors*.
2. The effect of chance errors on a result can be dealt with satisfactorily by methods of analytical statistics that have been developed specifically for this purpose.
3. These statistical methods are based on the implications of the mathematical theory of probability.
4. Their application consists essentially in setting up appropriate statistical *Tests of Significance* and, in the light of such tests, formulating conclusions in which *confidence* is warranted.

A Test of Significance, like those developed in Chapter 13, Section D would indicate that we can be confident that at least a clear majority of the voters' opinions, as of the time the poll was made, was opposed to such a loan to England.

B. STATISTICAL POPULATIONS OR UNIVERSES

The Statistical Universe

A *statistical population* is not to be confused with a population of people. In the preceding sample, the statistical population under consideration was not a group of people; rather, it consisted of the *opinions* on the particular question held by all the voters in the United States. Opinions are obtained from people, just as are measures of aptitude, personality ratings, etc. It is the attributes, behavior, or traits of people, therefore, that constitute statistical populations or universes. Populations may also consist of the

attributes or qualities of other things, as for instance, the *behavior* of dice or coins, or the *behavior* of the molecules in a given volume of gas.

The concept *population* or *universe* in statistics thus denotes the whole which includes all the observations or measurements of a non-variable or variable characteristic. If the characteristic is non-variable and dichotomous, such as radios in U. S. homes, the statistical universe will be the *number* of all homes equipped with radios and the number of all homes not equipped with radios. If the characteristic is variable, such as the radio listening behavior of people, the statistical universe *for a given period of time* will be the different amounts of time all the people spend listening to their radios. The statistical universe could be the dichotomized behavior of a coin tossed an infinite number of times, and consisting of the number of times the coin landed heads up and the number of times it landed tails up. Or the statistical universe could be the variable behavior of a collection of coins tossed an infinite number of times, and consisting of a distribution of the frequencies with which the different possible combinations of heads and tails (including all heads and all tails) occurred.

The concept *population* or *universe* in statistics thus defines all the measurements or observations of the attribute or behavior of the phenomenon being studied. The actual behavior of dice as observed would constitute a sample from a universe whose instances theoretically comprise a population infinite in size. In microphysics, the behavior of a molecule in a given volume of gas could constitute a universe from which a very small sample of such behavior might be measured.

Finite and Infinite Populations

From the foregoing examples of statistical populations, or universes, it should be evident that they may be finite or infinite in size. The universes of public opinion and market research investigations are usually taken as finite, even though they may not be readily susceptible to an exact count. The fundamental calculus of probabilities has been developed for universes considered infinite in size, as for example, the behavior of dice or coins tossed an infinite number of times. Fortunately, from the point of view of applying the calculus of probabilities to finite universes, the latter can often be treated *as if* they were infinite in size, provided that the universe is large in relation to the size of the sample. Appropriate samples drawn from large finite populations will have statistical implications similar to those of similar samples drawn from infinite populations.

Actual vs. Hypothetical Universes

All infinite universes are necessarily hypothetical. An actual universe, on the other hand, is one such that the behavior of all its members or instances is susceptible to observation or measurement. Such a universe obviously must

be finite in size. A behavior or quality of the adults of a nation constitutes an actual universe provided we are referring to the behavior of those adults at a given time. But if we are speaking of a trait or characteristic of a racial group, such as stature or skin pigmentation, then presumably we are referring to a hypothetical universe that includes people still to be born, as well as those living today. Such a universe is part actual and part hypothetical. Obviously only its existing part can be sampled at any given time.

From the point of view of the statistical analysis of sample results, it makes little difference whether the universe is actual or hypothetical, provided each can be appropriately sampled. In one sense, this distinction between an actual and a hypothetical universe is exemplified by any situation in which an attempt is made to predict the future behavior of the phenomenon studied. The basis for present observations is a sample from a necessarily existing universe, where the predictions concern a universe not yet come into being. This contrast, interestingly enough, is characteristic of actuarial analysis, in which, for example, mortality tables are established and the attempt is made to predict the length of life of living people. Similarly, in polls of voters' preferences for candidates prior to elections, the attempt is made not only to ascertain their preferences at the time of the poll, but also to predict their preferences when the voters actually go to the polls. In other words, prior to the election the behavior of actual voting constitutes a hypothetical universe. It cannot be sampled directly, but people's attitudes and opinions about the way they expect to vote can be sampled and studied.

We cannot be *certain* that sample results in September or October will necessarily forecast the outcome of a November election, but we can obtain preliminary "straws-in-the-wind." Like the actuary, the researcher in public opinion can make estimates or predictions with considerable confidence, because of the ever-increasing, successful experience in this field. Unusual events that might radically alter the outcome of an election, such as the death of one of the candidates just prior to election day, could of course vitiate the prediction based on a poll of voters' preferences. *There are no certainties in the statistics of probability, but there are reasonable expectancies.*

C. SAMPLES AND THE TECHNIQUES OF SAMPLING

A sample is an actual collection of observations or measurements of an attribute, a behavior, etc. A sample is therefore necessarily finite. There are, however, different kinds of samples. They can be differentiated in terms of either (1) the way in which they are related to the universes of which they are a part, or (2) the technique or method used in obtaining them. Either or both of these criteria, considered together, are relevant to the way in which a particular sample is to be characterized. Samples may be *representative* or *biased; random* or *stratified-and-random; adequate* or *inadequate; controlled* or *uncontrolled; restricted* or *unrestricted*, etc.

Representative Samples

A sample of observations or measurements is representative of the universe of which it is a part if it is a replica of that universe. Thus, if a sample of information about the nativity of 100 people living in a given area shows that 75 are native-born and 25 are foreign-born, and a census of the nativity of all the people in that area reveals that 75% are native-born and 25% are foreign-born, then the sample of 100 observations is a replica of the statistical universe—the sample is truly representative. A sample of the intelligence test scores of 1000 salesmen would be representative of the intelligence test scores of all salesmen if the *distributions* for both sample and universe were similar in form and if their means and standard deviations were the same.

A representative sample, in other words, is one that yields a division or distribution of the attribute or behavior being studied that is the same as its division or distribution in the statistical population. The question of the representativeness of a sample is directly concerned with the behavior or trait under study, and only indirectly with some other behavior or traits Does a sample of measurements of variable x yield values that are the same as the measurements of variable x derived from the entire statistical universe? If $x_s = x_u$, the representativeness is perfect. In a market research investigation of consumer-use of a product, a sample is not necessarily representative of *consumer-use* in the universe, even though the sample is composed of the proportions of the sexes, of the different age groups, of the different income or economic groups, etc., that are truly characteristic of the proportions of these attributes in the total group of people in that particular market area. The sample is truly representative only if the proportions of those people using the given product and those not using it are identical with the corresponding proportions in the population. A sample is a replica of a universe, and therefore *representative*, provided that the distribution of the observations or measurements under investigation is identical in both sample and universe.

In order to determine whether a sample result is *truly* representative of a universe, we need to know certain facts about that universe. Ordinarily, however, we do not have the information about the universe necessary for an absolute check on the representativeness of the sample. If we did, there obviously would be no need to work with a sample; we could proceed with a summary and analysis of the data for the universe itself. In practice, therefore, it is rarely possible to describe a sample as truly representative of a population. Instead, the character of a sample is usually described in terms of the methods used in obtaining it rather than in terms of its representativeness of the statistical population. By this criterion of *method*, samples may be classified as follows:

1. Random samples.
2. Stratified-random samples.
3. Accidental or uncontrolled samples.

Either of the first two techniques, when used with samples of adequate size, should yield a result that is sufficiently representative of the statistical population to warrant confidence in the conclusion drawn from it. *Representativeness is a question of degree,* not of mutually exclusive alternates.

Before considering the methods and techniques of sampling, we shall consider the obverse of a representative sample, viz., a biased sample.

Biased Samples

A biased sample is definitely non-representative of the statistical universe of which it is a part. The scientist obviously uses every possible means to avoid biased samples. Consequently, when they occur, it is because of carelessness or ignorance of proper sampling methods, rather than any intent on the part of the investigator to obtain unsatisfactory results.

One of the most famous examples of a biased sample in public opinion research occurred during the presidential campaign of 1936. The *Literary Digest* had conducted mail-ballot polls of voters' preferences during several other presidential campaigns, and the predictions had been fairly successful. In 1936, however, its poll failed completely when it predicted the overwhelming defeat of Roosevelt and the election of Landon. It was found that the ballots mailed by the magazine in this poll were addressed only to samples of listed telephone subscribers and automobile owners throughout the United States. Unfortunately for the success of this poll, the voting preferences of people whose homes had telephones and who owned automobiles were not representative of the people generally. Many more people who had no telephones or automobiles voted for Roosevelt than for Landon in 1936. Interestingly enough, during the same campaign Dr. Gallup conducted a national poll by both mail-ballot and personal interview methods. Giving the factors of listed telephone subscribers and registered automobile owners, as well as other relevant factors, their proper weight in the total result, he predicted rather closely not only the outcome of the election but also the error in the *Literary Digest* poll.

Not all kinds of universes are equally difficult to sample without bias. An analysis of the red and white cells in a drop of blood is sufficient to give a satisfactorily representative sample of the distribution of all the red and white cells in a person's total blood supply. Although the population of cells may be somewhat variable in different parts of the circulatory system, the blood cells are sufficiently *homogeneous* in their distribution so that a drop taken from the finger tip will give a satisfactory *unit* for sample analysis. Here the sample is a cell count based on a small unit of the whole. But a single geographical unit of people in the United States, such as any city or any state, would in no way be satisfactory as a base for studying the anthropometric characteristics of all the people in this country, let alone their psychological traits and social attitudes. The homogeneity characteristic of

a drop of blood does not characterize the distribution of traits or behavior among people. That is, a group of people drawn from a small geographical unit of the whole will not be likely to provide a satisfactory base for a sample of such variable qualities as stature, aptitude, and opinion. In fact, as we pass from physical anthropology to psychological functions and social attitudes, the variability or heterogeneity of statistical universes increases. And, *pari passu*, they become more difficult to sample without bias.

Constant vs. Chance Errors

In contrast to *chance* errors, of either sampling or measurement, the errors that result from biasing factors affect the results of an experiment or investigation in a constant way. Hence, biasing errors are usually called *constant errors* of sampling or of measurement. How can these constant errors be avoided in sampling?

As the experience of research workers in a given field accumulates, it is increasingly possible to obtain samples that avoid bias to any disturbing or distorting extent. The accumulation of sampling experience, checked against census results or against the predicated implications of sample results, reveals various factors that make for bias and makes it possible to avoid them sedulously. There would obviously be bias in any attempt to sample voters' preferences with a group of registered Democrats as the only base.

Fortunately two sampling techniques are available that permit most, if not all, of the pitfalls of biased samples to be avoided. In fact, only with one or the other of them, or as close an approximation as is possible, can we hope to obtain satisfactory samples for the study of a universe. These two methods are random sampling and stratified-random sampling. Random sampling techniques yield a result for the universe being studied that becomes increasingly satisfactory as the size of the sample is increased. Stratified-random sampling also yields a result that becomes increasingly satisfactory not only as the size of the sample is increased, but also as certain additional control factors are introduced.

Bias from Inadequate Methods of Observation and Measurement

Bias in surveys of the character, behavior, or opinions of people is often greater because of constant errors of observation and measurement than because of constant errors of sampling. The science of sampling has been sufficiently well developed to enable the researcher (1) to reduce *chance* sampling errors to a satisfactory minimum and (2) to avoid serious biases in sampling.

Biasing errors of observation and measurement, however, are all too likely to plague any investigation in which the information sought is about or derived from people. W. E. Deming, of the United States Bureaus of the Census and the Budget, has listed thirteen factors that make for errors in

surveys.* Only a few of them lead to errors that are due to inadequate sampling, as for example, bias resulting from non-response or from late responses, from unrepresentative selections of respondents, or an unrepresentative date for a survey. The remaining sources of error are largely due to faulty procedures of observation and measurement, such as (1) variability in the answers or data furnished by respondents; (2) errors peculiar to the kind of method employed in canvassing respondents (mail vs. telephone vs. telegraph vs. direct interviews; intensive vs. extensive interviews, etc.); (3) bias and variations caused by faulty interviewing; (4) bias due to respondents' reaction to knowledge about the sponsor of the investigation; (5) bias from imperfectly designed questionnaires; (6) processing errors in the coding, tabulation, and statistical summarization of the data; and (7) errors due to misinterpreting the questionnaire data or the statistical results, or to personal bias in interpretation.

Biasing errors of observation and measurement resulting from one or more of the above causes are as common in a census of a population as in a sample. Nor does the taking of a census in itself eliminate biases of this kind. If a census is taken, but the methods of enumeration or measurement are faulty, only the sampling error will be minimized or eliminated. It is because of errors of observation and measurement that no categorical answer can be given to the question of the size of sample necessary to guarantee an adequate result or a result that will be accurate within a given margin of error. *Adequacy* depends not only on the character and size of the sample, but also on the techniques and procedures used to obtain the desired information from the respondents and to summarize it. The pre-testing of techniques and preliminary test-tube surveys are essential if biases in observation and measurement are to be avoided.

D. RANDOM SAMPLES—THE PRINCIPLE OF RANDOMIZATION

Definition

A random sample is one such that each instance or member of the universe being sampled has *an equal chance of appearing in the sample*. In other words, each observation or measurement of a random sample has the same opportunity, no more and no less, as all the other instances of the universe, of appearing in the sample. Sometimes random sampling is called "simple sampling." However, there is nothing simple about *applying* the method, for it involves the basic technique of control characteristic of all sound sampling procedures.

The theory and development of sampling statistics are based upon the assumption that a sample of a universe is a random sample. To the extent

* W. E. Deming, "On Errors in Surveys," *American Sociological Review*, 9:359–369, 1944.

that samples are not random, we cannot have confidence in generalizations about universes, made in the light of an analysis of sample results.*

The Technique of Random Sampling

How can a sample be established so that each instance or member of the universe to be studied will have an equal opportunity of appearing in the sample? In order to *guarantee* that a given sample will be a random sample, a careful control technique is necessary. It consists essentially in numbering each member or instance of the universe, and then drawing the desired size of sample by means of a lottery technique which in itself has been tested for randomness.

Random Numbers

Since mechanical lottery procedures are often inconvenient to use, tables of numbers that have been tested for randomness are available.† A table of truly random numbers is one such that any digit from zero to 9 has an equal chance of appearing in any position in the table (see Table II, Appendix C). Digits are combined by rows or columns to give two-place, three-place, or n-place numbers, according to the needs of the particular study. Hence, if the members or instances of a universe are consecutively numbered from 1 to N (the total size of the universe), a random sample of 100 or 1000, or whatever size is desired, can be obtained from the universe by means of a table of random numbers. The instances in the universe are chosen for the sample when their *numbers* come up in the table of random numbers.

It is often impossible to use a lottery method that will *guarantee* a random sample of a statistical universe. When a universe is taken as infinite in size, it is obviously impossible to number all its members and draw from it a random sample by the procedure just described. Also, it is usually not feasible or possible to assign numbers to all the members of a large finite population, such as all the voters in the United States. Ordinarily it is impractical to assign a number even to each voter in one state or one city in order to draw a random sample with the aid of a table of random numbers. Only in a great emergency, as in World War II and the Selective Service System, is it feasible to number large finite populations and draw samples by a truly random method, whether by a mechanical lottery device or by a table of random numbers.

In practice, research workers have therefore had to develop other techniques for the randomization of a sample. These can yield satisfactory results,

* A stratified-random sample is a special case of random sampling and hence does not constitute an exception to this basic assumption.

† R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural and Medical Research*, pp. 82–87, Oliver and Boyd, London, 1938, Table 33; M. G. Kendall and B. B. Smith, "Randomness and Random Sampling Numbers," *Journal of the Royal Statistical Society*, **101**:147–166, 1938; L. H. C. Tippett, *Random Sampling Numbers*, Cambridge Univ. Press, Cambridge, 1927.

even though the *guarantee* of randomness characteristic of a table of random numbers or a lottery device may be lacking.

Alphabetization as a Basis for Sampling

If there is available an alphabetical listing of the names of persons whose attributes or behavior is to constitute a universe to be studied by sampling methods, then each *i*th case (5th, 10th, 20th, etc., depending on the size of sample desired in relation to the size of the finite population) can be drawn, and the sample will yield a result fairly comparable to that obtained with a table of random numbers. Certain precautions in using such lists are necessary to avoid distortion of the samples. As Stephan* has pointed out, some of the names of the list or some records in a file may be missing or have been temporarily removed. If records are missing because they are being used, their absence may distort the sample because there may be some correlation between the active use of these records and the trait or behavior being studied.

One of the best examples of the randomization of samples in market research is provided by C. E. Hooper's radio research organization which conducts surveys of home listening to radio programs; the surveys are published at regular intervals and are widely used by the radio and advertising industries.† Hooper's organization draws random samples from the lists of telephone subscribers in the major population areas in the United States. The names for each sample are located on these lists by means of a mechanical lottery device. Hooper's statistical universe is strictly defined in terms of listed telephone subscribers and is sufficiently large that no subscriber will be included in a random sample more often than once a year. *Over the calendar year*, each listed subscriber has an equal opportunity with all other listed subscribers of being included in the samples chosen. The names drawn for a given sample are used only once during the year in order to avoid errors of *measurement* that might result from too frequent telephone calls to the same homes, and the consequent possible annoyance of the respondents.

In the United States Census of 1940, some of the information sought was obtained by sampling methods, for the first time in the history of the census.‡ A sampling technique somewhat analogous to taking every *i*th case from a file was employed by the interviewers. Each census taker secured information on certain questions (nativity, usual occupation, social security status, marital status, etc.) from each 20th person. The sample obtained was therefore based on 5% of all the people in the United States.

---

* F. F. Stephan, "Practical Problems of Sampling Procedures," *American Sociological Review*, 1:569–580, 1936.

† M. N. Chappell and C. E. Hooper, *Radio Audience Measurement*, Stephen Daye, New York, 1944.

‡ Cf. F. F. Stephan, W. E. Deming, and M. H. Hansen, "The Sampling Procedure of the 1940 Population Census," *Journal of the American Statistical Association*, 35:615–630, 1940.

Interviewing each *i*th case of a universe can yield a satisfactory sample provided the research worker uses a method that will avoid basing the selection of each *i*th case on any factors that could conceivably bias the results sought. If the interviewers adhere strictly to asking, say, each 20th case the sample questions and go from family unit to family unit in an order predetermined from their location on a map (the order must not be influenced by the appearance of the neighborhood, the loquacity of a respondent, the absence of a respondent, etc.), bias can usually be avoided. Generally, the randomization of people or families in house-to-house interviews in a given area must be accomplished by a systematic *design* of alternation or selection, laid out on a map or schedule in advance of the actual field work. When homes or families can be numbered, a table of random numbers or a lottery device can be used in selecting those to be included in the sample.

Principle of Inertia of Large Numbers

Some of the interviewers for the 1940 Census were evidently not too well trained for their jobs, and consequently certain biases may have entered *some* of the interviews. If this was true of only a small proportion of the interviews, the practical effect of such errors of measurement, or of sampling, would most likely be negligible because of the thousands of interviewers at work, and because of the operation of *the principle of the inertia of large numbers*. Thus, the 5% sample obtained in the 1940 Census amounted to more than six million cases. An extra million or two would not be likely to make any practical difference in the over-all results of the national sample, even though some interviewers may not have adhered too strictly to the rules for randomizing the sample on each 20th case, in order of interviews.

It should be emphasized in this connection that merely increasing the size of the sample will not eliminate any biases inherent in the general sampling technique used. Nor will the replication (or repetition) of a sample increase the soundness of the result if the same defective sampling method is employed for both samples. In other words, the principle of the inertia of large numbers does not mean that the addition of thousands or millions of cases to an already sizable sample will eliminate the effect of bias inherent in the sampling technique, but rather, that the effects of bias or constant errors may become negligible if such errors occur in only a small proportion of the observations or measurements.

The Sampling Unit

The sampling unit of an investigation is the basic identity whose characteristics or behavior is to be studied. The basic sampling unit in biology and the social sciences is usually the individual organism or person. However, the sampling unit is often the family in consumer research studies. It may be the household, a farm, a business organization, a school, a type of crop, even

a molecule. The sampling unit for a given investigation naturally depends on the nature and purpose of the study.

Initial or *primary* sampling units are sometimes used and are to be distinguished from the sampling unit per se. That is, the sampling units may lie within a large geographical area, and the initial sampling made with respect to geographical subdivisions prior to the sampling of the individual units that are to be studied.

In market and public opinion research, an initial random sample of geographical or areal units is sometimes obtained, such as townships or sections in sparsely populated areas, and blocks in cities. These units are numbered, and those to be used are then selected by means of a table of random numbers. The sample of people whose behavior or opinions are to be studied are those residing in each such geographical unit. Or random sub-samples within each geographical unit may be drawn. But the basic sampling unit is still a person, and not the geographical areas. The latter are units of the distribution of people and are used as the *initial* or *primary* sampling units when a universe is too large for feasible randomization of people from it as a whole.

The size and number of geographical units to be employed depend upon the purpose of the investigation, the heterogeneity of the people in the universe to be sampled, and their density of concentration per areal unit. Generally, smaller geographical units and a greater number of these units for sampling give a more dependable result than larger areal units and a correspondingly smaller number of them. This is the case because the homes of people in a city, state, or nation are not randomly distributed. That is, people of the same color group, or of similar educational or economic classes, tend to be located in one section--"the West Side" vs. "the other side of the tracks," etc. And for many types of market and public opinion investigations, there is a correlation between such factors as people's skin color, economic status, and education, and their buying habits, social attitudes, opinions on public questions, etc.

The difficulties encountered in obtaining random sampling units from areal or other initial sampling units are so great that the technique of stratified-random sampling is now usually employed. With this method, as we shall see in the following section, the results for any areal unit of the total sample can be prevented from overweighting the result. Thus, the difference in the concentration of family units in poor and in wealthy neighborhoods can be adjusted for and given the proper weight in the total sample.

Public school systems or schools, rather than geographical areas, are taken as the units for sampling in some types of investigations, as in educational or public finance research. All the public schools in a city, county, state, or nation can be numbered and a random sample drawn. The relevant aspects of each school (pupils, teachers, textbooks, financial records, etc., depending on the purpose of the investigation) will comprise the basic sampling units. If it is the average size of classes, and their variation, that is to be determined

from a sample of schools, the basic sampling units will be classes instead of pupils. If it is the ratio of average daily attendance to enrollment that is being investigated, the initial units, *schools*, will also be the basic sampling units.

When the initial sampling unit differs from the basic sampling unit, stratified-random sampling is usually better than to employ the initial units and depend upon relevant factors to randomize themselves among these units. Possible biasing factors can be controlled better by the stratification technique.

E. STRATIFIED–RANDOM SAMPLING

Definition

A stratified-random sample consists of two or more random samples drawn from two or more subdivisions (or strata) of the universe, each stratum having been established with respect to one or more secondary control factors. The size or weight of the sample from each stratum corresponds to the proportionate size or weight of the control factors in the universe being studied. The total sample result is therefore derived from a series of random samples, each of which is properly weighted for the proportionate incidence of the control factors in the universe. The secondary control factors in stratified sampling are traits or behavior that correlate to some degree with the attribute to be investigated. They are the criteria on the basis of which a universe is subdivided into two or more strata, or combination of strata, each of which is then sampled.

In the literature, stratified-random samples are often referred to simply as *stratified* samples. However, when randomization is employed as the primary control factor, as should be the case when the total result is to be treated *as if* it were a random sample of the universe studied, then it is important to emphasize the fact that randomization, the basic factor of control in all scientific sampling, has been utilized. If it has not been employed, as sometimes happens even though the sample is stratified, the result should obviously not be called a stratified-random sample. The stratified samples employed in many public opinion and market research studies are in fact not stratified-*random* samples but simply *stratified* samples.

Stratification consists in the drawing of samples in such a way that the representation of the stratifying factors in the total sample corresponds to their respective proportions in the universe studied. The principle is a useful control technique provided (1) that there is a significant correlation between the control factor, or factors, and the trait or behavior to be studied, and (2) that the necessary information about the universe is available so that the stratification can be based on facts and not merely on guesswork. The latter condition is the more difficult to satisfy in many investigations because of

the lack of up-to-date and dependable information about the characteristics of the universe to be studied.

Stratifying Factors

In public opinion and market research investigations in which the universes may be voters' preferences and opinions on public questions, or consumer preferences for and habits in the use of a brand commodity, etc., the most relevant factors for stratification are generally such attributes or characteristics as *sex*, *age*, *socio-economic status*, *education*, *residence* (urban vs. rural), *geographical* or *sectional region* in the United States, and *occupation*. In many such studies, additional factors, such as *skin color*, *nationality background*, *religious affiliation*, *political party affiliation*, etc., may be relevant for stratification.

Diminishing Returns

It usually is not feasible or helpful to stratify for all these various factors in an investigation, even when all the relevant information about the universe is available and stratification is possible. For one thing, the principle of *diminishing returns* operates quickly, both statistically and cost-wise. Furthermore, it becomes increasingly difficult to obtain random strata-samples for more than a few factors because of the inter-correlation of the factors themselves and the attendant difficulty of locating and identifying the sampling units when these units are people.

The Technique of Stratification

The stratification of a group of people whose attitudes or opinions are to constitute the universe under investigation consists, then, in drawing a series of random samples in such a way that the proportions of the stratifying factors in the sample will correspond to their proportions in the universe. If, for example, a researcher is to study the attitudes of the people in a city toward the prevailing policy of double-feature movies for the manager of a motion-picture theater, he can obtain a much more satisfactory sample, at the same unit cost, if the public is stratified for *sex* and *age*, provided there is a correlation between sex and such attitudes, and between age and such attitudes.

If a preliminary *test-tube sample* of 100 people showed the correlation between sex and attitudes given in Table 11:1, this result could well be used as the basis for a control factor in the main survey, since the double feature appeals more to women, and the single feature appeals more to men. Instead of one random sample taken without regard to the sex of the respondents, two random samples would be drawn, one for each subdivision of the stratum, viz., one of males and one of females. Each would be drawn to a size such that the respective proportions of males and females in the total sample would be similar to their proportions in the universe whose attitudes are to be studied.

If the proportion of males in the universe is 50%, the male stratum-sample would consist of a random sample equal in size to that of the female stratum-sample. (If the sample results were unequal in size, they could be given equal weights in the total result.)

Table 11:1

Type of Theater Operation Preferred	Sex: Male	Sex: Female	n_r
Double Feature	15	45	60
Single Feature	30	10	40
n_c	45	55	N = 100

Correlated Control Factors Reduce the Sampling Error

By this control device, the effect of some of the error attributable to chance factors can be reduced. A single random sample cannot be expected to give a truly *representative* result for the attitudes of the universe studied, nor would such a sample ordinarily give the same proportion of males and females as in the universe. The chance error effects caused by too many males or too many females in a single random sample can be avoided. On the basis of the correlation shown by the above cross-tabulation of the test-tube sample, an undue proportion of females in a random sample would overweight the result in favor of double-feature movies, and too many males would overweight the result in favor of single-feature movies.

But, it may be asked, why should not a single random sample of the universe being studied avoid the errors caused by an atypical division of the sexes? If, as assumed, the single sample were truly random, no *constant errors* of sampling would enter into the result, and it is these errors, as we have seen, that make for bias. In answer, it is to be emphasized that the failure to stratify for sex in the preceding example would not make for bias, so long as the single sample was a random sample. Rather, for a random sample of a given size, say 1000 cases, the effect of chance errors of sampling would be *reduced* (not eliminated) by stratifying the sample on the basis of a factor that correlates with the attitude or behavior to be investigated.

Depending on the size of the sample, the variation possible in the proportion of the sexes in a single random sample will vary purely on the basis of chance. The situation here is exactly analogous to the tossing of 10 coins. In the long run, we would expect the results to average 5 heads and 5 tails, if the coins themselves were fair. On any single toss, however, we might get

any one of 11 possible combinations, ranging from all heads, through 5 heads and 5 tails, to all tails. Purely on the basis of chance, the "long shot"—10 heads or 10 tails—could occur. Similarly, if the division of the sexes in a universe is equal, then truly random samples should in the long run yield an average of 50% males and 50% females. On the basis of chance alone, a single random sample could, however, yield a "long-shot" result.

The stratification of a sample on the basis of a factor that correlates with the trait or behavior to be studied thus reduces the variability of the error effects on the sample results. For a given size of sample, the results obtained with a stratified-random sample will therefore be more precisely representative of the universe than those obtained with an unstratified random sample. The two controls—the primary one of randomization, and the secondary one of stratification—will increase the accuracy of a result for a total sample of a given size. But if there is no correlation between a stratifying factor and the attribute or behavior, then the over-all random sample will yield a result that is equally as satisfactory—but no more and no less—as the stratified-random sample.

Stratified-random samples are thus a device for increasing the accuracy of a single random sample of a universe. For a given research budget, the researcher can obtain more adequate results; or, for a given size of sample, he can get more information at the same per capita cost. Stratification of course involves expenses that are not incurred in an over-all random sample, such as the cost of determining the divisions of a population with respect to factors known or thought to be relevant. Such data, however, are becoming increasingly available as a result of governmental and private censuses of cities, rural areas, counties, states, etc. Furthermore, by means of preliminary test-tube surveys, the possible relevance of one or more secondary control factors can be determined. Such preliminary surveys should be made for each specific investigation, so that the procedures of observation or measurement (the construction of the interview, the wording of a questionnaire, etc.) can be pretested for possible biases.

The Inter-Relation of Stratifying Factors

In the preceding example it was indicated that *age* as well as sex might be a second relevant factor in stratifying a sample for determining people's attitudes toward double-feature movies. If the preliminary test-tube survey showed a result like that in Table 11:2, a random sample would also be relevantly controlled if it were stratified for *age* as well as *sex*, because of the younger people's greater preference for double-feature movies and the older people's preference for single-feature movies. However, the two factors of sex and age are not independent, since all sampling units of the universe will be identified with both a sex group and an age group. Consequently, to stratify for both these factors, their inter-relation must be taken into account. That is,

Table 11:2

Type of Theater Operation Preferred	Age: Under 30	Age: 30 and Over	n_r
Double Feature	40	20	60
Single Feature	10	30	40
n_c	50	50	N = 100

the investigator would *not* proceed by drawing a random sample of 50% men and 50% women, and a second random sample of 50% "young" and 50% "old," according to the proportions of two such dichotomized age groups in the universe. Rather he will stratify for both factors simultaneously and draw four random strata-samples, one for each combination of the two factors, as indicated by the *cells* in Table 11:3. Such a division of the two combined

Table 11:3

Age	Sex: Males	Sex: Females	$\%_r$
30 and Over	25%	25%	50%
Under 30	25%	25%	50%
$\%_c$	50%	50%	100%

factors, sex and age, into *four equal-sized groups* would necessarily be based on knowledge that the sampling units (people) of the universe to be studied consisted of 50% males and 50% females, and that half of each of these groups were 30 years of age and over. In practice, such an even distribution of these two factors would not be typical for many universes.

The importance of defining the universe to be studied prior to the investigation was emphasized earlier, and is well illustrated by the above example. The researcher doing the study of people's attitudes toward double-feature movies for the motion-picture operator would not include everyone in the city in the universe to be sampled; he would eliminate infants and young children (perhaps under the age of 10) as well as the incapacitated. On the other hand, he would not restrict the universe solely to those who attend

movies regularly, because such a restriction would eliminate the people who might go more regularly if double features were eliminated. The sampling units for the universe might thus consist of anyone over 10 years of age, except the bedridden and those otherwise unable to attend theaters.

The Inter-Correlation of Stratifying Factors

Preliminary analysis of test-tube results often makes it possible to determine roughly whether the addition of control factors is likely to increase the representativeness and precision of a sample result, or whether a particular factor may be as adequate for control purposes as would its combination with another factor. How such an analysis is made is illustrated by Table 11:4,

Table 11:4

Type of Theater Operation Preferred	Male			Female			Total Male and Female
	Younger	Older	n_r	Younger	Older	n_r	n_r
Double Feature	13	2	15	27	18	45	60
Single Feature	9	21	30	1	9	10	40
n_c	22	23	45	28	27	55	N = 100

showing cross-tabulations of the hypothetical test-tube data for sex and age differences in attitudes toward double-feature movies. From this analysis with respect to *both* age and sex, it is evident that sex is probably a more relevant control factor than age, because a greater proportion of the females than males, regardless of age, prefer double features. It does not follow that age is entirely irrelevant as a control factor, as can be seen from the rearrangement of the same data in Table 11:5. In the older group, the correlation

Table 11:5

Type of Theater Operation Preferred	Younger			Older			Total of Young and Old
	Male	Female	n_r	Male	Female	n_r	n_r
Double Feature	13	27	40	2	18	20	60
Single Feature	9	1	10	21	9	30	40
n_c	22	28	50	23	27	50	N = 100

between sex and preference is high, the women preferring double features in a ratio of 18 to 9, and the men preferring single features in a ratio of 21 to 2. In the younger group, the women and girls prefer double features in a ratio of 27 to 1, but the men and boys also prefer the double feature, although the ratio is only 13 to 9.

Thus both age and sex would be relevant control factors in this investigation, although the principle of diminished returns operates when the age control factor is added to the sex factor.

Sub-Universes in Stratified-Random Sampling

Whatever the control factors used in stratifying a universe, and consequently in taking a random sample of that universe, the procedure of stratified-random sampling consists in sampling several independent parts of the whole, or universe, being studied. The whole is not randomized directly; rather, it is sampled by means of a number of random samples. Each part sample is a random sample of the members of a given cell, or multiple class—for example, all women over 35 years of age. The result obtained from each such sample is for a *part* of the whole. Hence, each random sample of an independent part of a universe may be called a random sample of a sub-universe. The process or technique of stratified-random sampling can therefore be looked upon as building up a sample for a universe by means of bringing together the results of a series of random samples for the sub-universes that make up the whole, each such part sample being appropriately weighted to yield a single sample result for the universe.

The sampling units of some sub-universes are relatively easy to identify, segregate, and sample independently of one another; this is true of universes that are stratified for geographical location or socio-economic status by residence areas. With other sub-universes, however, this is extremely difficult. For example, the four groups in the preceding example—older women, older men, younger women and girls, and younger men and boys—are intermingled in real life and cannot readily be approached as independent sub-universes of the whole. In such a case, the use of additional control factors in stratification is likely to complicate further the problem of locating the sampling units in each sub-universe.

When the sampling units are people, stratified-random sampling is particularly difficult to apply for more than but a very few control factors at a time. Where, for example, can a researcher locate the sampling units of a sub-universe defined as "white, female, white-collar workers, over 35 years of age, with a high school education, affiliated with the Democratic party, of the B socio-economic group, and residing in area X"? Such a sub-universe is not distributed in a manner that segregates it from other sub-universes, nor are lists of the members of such sub-universes usually available. As pointed out earlier, the principle of diminishing returns would make it un-

necessary to sample such a highly restricted sub-universe, even if all the factors mentioned were correlated with the trait or behavior to be studied.

Internal Controls in Sampling

In practice, when making the preliminary layout of his investigation, the researcher stratifies for very few factors if people are the sampling units. These factors are primarily external—density of population, geographical section or area, and residence—the latter especially in relation to socio-economic status. Sometimes, in city sampling, sections or neighborhoods are also stratified for nationality or cultural background when the city contains neighborhoods which have a relatively homogeneous concentration of people for this factor—"Little Italies," etc.

While he is obtaining a random sample of observations or measurements, the researcher brings in *internal controls* by determining not only the respondents' attitudes or opinions or consumer habits, but also their sex, age, occupation, education, etc. This additional information can be checked against the known distributions of these factors in the universe. By comparing the characteristics of the sample with the known characteristics of the universe, the investigator can determine whether he has in fact obtained a fairly typical cross-section as regards such factors as sex, age, education, occupation, etc. Where correlations are found between one or more of these factors and the trait or behavior to be studied, he can make adjustments in atypical samples for the purpose of reducing (not eliminating) the sampling error.

The possibilities of making such an adjustment by an analysis of a sample result are well illustrated by the *Literary Digest* poll during the presidential campaign of 1936. Each respondent in the mail-ballot poll of telephone subscribers and automobile owners was asked to indicate whom he had voted for in 1932. As Cornfield has pointed out,* it was possible to use the *published Literary Digest* results and predict the likelihood of a Roosevelt majority. The published results showed that 52% of those responding to the poll and reporting on how they had voted in 1932, had voted for Hoover; however, Hoover received only 41% of the vote that year. Thus the sample was heavily overweighted with 1932 Hoover voters, who in turn were overweighting the 1936 poll in favor of the Republican candidate.

Areal Sampling

Not all sampling problems are as difficult as those that arise in public opinion and market research surveys where the sampling unit is the citizen, the voter, the housewife, etc. Scientific sampling techniques to control the quality of production have come into widespread use in many manufacturing industries in recent years. The basic problem here is to obtain a random

* J. Cornfield, "On Certain Biases in Samples of Human Populations," *Journal of the American Statistical Association*, 37:63–68, 1942.

sample of the manufactured item as it comes off the production line. Similarly, the problem of developing norms for psychological tests is today rarely approached from the point of view of trying to obtain representative samples of the performance of *all* adults in the United States, of *all* 16-year-olds, etc. Rather, the universes are varied according to the particular purpose for which the test is designed. Thus, if a test is to be used in a given industrial situation, the norms will be developed on the basis of a random or stratified-random sample of test results derived from workers in that situation.

The fact remains, however, that a great deal of the empirical progress that has been made during the past ten years in sampling statistics in the social sciences has been in the fields of public opinion and market research. How are the complications of random sampling and of stratified-random sampling being surmounted in these fields? A most hopeful sign of real progress is the development of areal sampling, a method which, for public opinion and market research, promises to approximate most closely the strict conditions imposed by randomization in sampling.

Whether the sampling unit be people, particular kinds of people, family units, dwelling units, homes, or farms, the essential characteristics of areal sampling are as follows: (1) The minor civil units of a city, county, state, or nation are divided into small *area* units that are exclusive of each other. (2) Each sampling unit is associated with only one such area unit. (3) Some of the relevant characteristics of the sampling units of the area are known. Hence (4) a sample of areas can be used to establish stratified-random samples for a city, county, larger rural area, state, geographical section, or an entire nation.

The Bureau of the Census and the Bureau of Agricultural Economics have for some time given serious attention to the compilation of information on the relevant characteristics (or factors that can be used as secondary controls in stratified-random sampling) of small areas.*

On the basis of the 1940 census data (much of which is now out of date), *Block Summary Cards* for the 191 U.S. cities with 50,000 or more inhabitants in 1930, and *E.D. Summary Cards* for each of the 154,000 Enumeration Districts of the Census not included in the areas covered by the Block Summary Cards, were made available several years ago for use in sampling. The E.D. Summary Cards are identified by state, county, and minor civil division and include such information for each Enumeration District as the following: total population (usual range, between 500 and 1500), native white population; percentage native white male; Negro population and percentage male; number and percentage farm population and total number of farms; total number of dwelling units; number and percentage of owner-occupied dwelling units; and average rent per dwelling unit (including estimated rental values

* M. H. Hansen and W. E. Deming, "On Some Census Aids to Sampling," *Journal of the American Statistical Association*, **38**:353–357, 1943.

of owner-occupied units). The Block Summary Cards identify the location of each block in the 191 cities and give such information as number of structures; total dwelling units; vacant and occupied dwelling units; owner-occupied and tenant-occupied dwelling units; dwelling units occupied by non-white household; and average monthly rent per dwelling unit. The city block is usually smaller than the Enumeration District and is consequently a better areal unit for general sampling purposes.

When such information as the preceding is up to date, a sample can be stratified not only with respect to geographical section, urban and rural communities, and density or concentration of dwelling units and of population, but also for socio-economic status in terms of average monthly rental; for white, Negro, and mixed area units; for tenants and owners of dwelling units, etc. Within each stratum, a random sample of blocks or Enumeration Districts can be drawn systematically by means of random numbers or a tested lottery device, and the sampling units (people, family units, etc.) within each of the area units can be interviewed. Or sub-samples of the family units, etc., within the area units may be drawn by a random method. The latter procedure is particularly practical for area units with a large number of sampling units, especially if they are fairly homogeneous in the relevant stratifying characteristics.

If the research problem were to determine the distribution of television sets in dwelling units, *all* the dwelling units in each of the areas (blocks and Enumeration Districts), stratified and drawn in the random sample, would constitute the basic sampling units. The results could be weighted for such known control factors as concentration of dwelling units and average rental, as well as analyzed in terms of such factors. If voters' opinions on public questions or governmental policies were to be studied, opinions could be obtained from all the qualified voters in the stratified-random sample of areas drawn for the survey.

The maintenance of up-to-date information on scores of thousands of area units in the United States obviously poses something of a census problem as well as a financial problem. The project is too costly for any private organization to undertake; but because of the value of such information for scientific sampling in market research, private business and industrial organizations might well consider supporting such a project on a subscription basis. The addresses, names, ages, and sex of everyone residing in each areal unit, if kept up to date on an annual basis, would be particularly valuable supplementary information to that already considered. Furthermore, when people constitute the basic unit of sampling, the technique of randomization by tables of random numbers or a lottery device can be applied to the people themselves.*

* Cf. J. N. Webb, M. S. Northrop, and S. L. Payne, "Practical Applications of Theoretical Sampling Methods," *Journal of the American Statistical Association*, **38**:69–77, 1943. These authors would perhaps consider an annual enumeration of all small areal units in

There is a possible source of error in the researcher's knowing the respondents' names and addresses, especially in public opinion research. As Stock has pointed out, "There is evidence that people do not always give their true opinions even to a stranger who does not know their names or addresses and has no way of checking up on them again." * Errors that arise from non-anonymity are errors of measurement rather than of sampling, and can perhaps be overcome by a *secret-ballot technique*.†

The Technique of the Master Sample

Areal sampling on both the *extensive* scale (for the United States as a whole) and the *intensive* scale (by Enumeration Districts and city blocks) is for the future. In the meantime, the foundation for a technique for agricultural surveys has been laid in the development of the *Master Sample*, suggested in 1943 by Rensis Likert of the United States Bureau of Agricultural Economics.‡

The Master Sample technique attempts to achieve the advantages of small-unit areal sampling by means of a *cross-section* of areal samples for the universe to be studied, rather than by the subdivision of the complete universe into areal units. The Master Sample developed for agricultural surveys consisted in 1945 of 67,000 sample areas, selected from practically all the 3070 counties in the United States. They averaged about 2.5 square miles and about 5 farms per areal unit. Their actual size varies with the density of the population: they are largest in Nevada, where they average 108 square miles, and smallest in Indiana, where they average 0.71 square mile. The total 67,000 sample areas contain 1/18 of the land area of the United States, 1/18 of the farms (about 300,000), and 1/18 of the population. Thus the Master Sample is a cross-section sample that is equal to 5.5% of the whole. It is stratified for the civil character of the areas—Incorporated, Unincorporated, and Open Country—as well as for density of population. By means of aerial photographs of all the counties, *work maps* have been set up on which each areal unit of the Master Sample is identified in relation to many more landmarks than can be indicated on an ordinary map. In fact, the initial definition of each areal unit was considerably facilitated by the use of aerial photographs.

The data of the 1945 Agricultural Census will be drawn upon to establish relevant characteristics for the farms and people of each area in the Master Sample. But, even without advance knowledge of the population character-

---

the United States both extremely impractical and unnecessary. Certainly such a program would be impractical unless, as intended, the data were conveniently available on machine cards for use by many government agencies as well as private organizations. The Master Sample plan, described in the next section, may prove to be sufficiently satisfactory to constitute an adequate body of information for all purposes.

* J. S. Stock, "Some General Principles of Sampling," chap. 10 in Hadley Cantril, *op. cit.*, p. 139.

† W. Turnbull, "Secret vs. Nonsecret Ballots," *ibid.*, chap. 5.

‡ A. J. King and R. J. Jessen, "The Master Sample of Agriculture," *Journal of the American Statistical Association*, 40:38–56, 1945.

istics, the Master Sample can be used for many research problems in agricultural economics. Plantings, estimates of yield, and actual yield for various types of crops can be calculated quickly and efficiently.

The Master Sample technique will undoubtedly be extended for use in all types of surveys, especially labor statistics, and may well prove a satisfactory alternative to areal sampling of the total population. If it is generally satisfactory from a sampling point of view, particularly with respect to the avoidance of biasing factors, the Master Sample plan will certainly be more efficient per unit cost. As in total areal sampling, it has the advantage of eliminating freedom of choice on the part of the field workers, for the sampling for an investigation can be completely designed in advance in the central office by means of appropriately stratified and randomized techniques.

The Random-Point Method of Sampling

The random-point technique of sampling, sometimes included in stratified sampling methods, consists in the random location of points on a map from which the sampling units are chosen, so many of the nearest households, farms, people, etc., being taken as the sample. The biases that may enter are unfortunately difficult to control. It is difficult, for example, to design a truly random method for selecting the specified number of sampling units at each point on the map. And, as was said earlier, a random selection of geographical units whose basic characteristics are unknown may result in an over-concentration of certain characteristics that will make for bias in the results. The stratification of factors *among* geographical units, as in areal sampling, is practically a "must" type of control for any sampling procedure in which the initial sampling units are located on a map.

The Stratified-Quota Method of Sampling

The stratified-quota method is the sampling technique that has been successfully employed by Gallup and others in public opinion and market research studies. It consists essentially in the stratification of the universe to be studied so that the people in the sample will be in the same proportion as they are in the total population sampled. Four initial strata controls are generally used in a national sample: geographical section of the United States, rural-urban distribution, economic status, and color of respondent. Additional strata controls, such as age and sex, are usually employed during the interviewing process itself. Each interviewer in a specified area is given a definite quota of interviews for each cell of the inter-related strata. Thus a quota may be "ten white men, over forty, in the highest income group." The interviewer attempts to avoid any biases in selecting these ten men, and the persons in the other cells in his quota schedule. The success of this technique depends to a great extent upon the interviewer's ingenuity in avoiding biasing factors when he selects the people to be interviewed.

This method is an approximation of the stratified-random sampling technique, and is often so characterized because of the efforts to obtain random samples of people in filling the quotas. Each quota for an area is set up so that its sample proportions correspond to the sub-universe proportions. If 5% of a sub-universe are in the highest socio-economic group, then 5% of the quota sample will include members of this group, or the sample result will be weighted according to this proportion.

As in stratified-random sampling, the feasibility of stratified-quota sampling is dependent upon knowledge about the distribution of the stratifying factors in the universe to be studied. The stratified sample cannot be a "typical cross-section" (the usual descriptive phrase) unless the distribution of such factors as population concentration in the major geographical sections and in rural and urban areas, socio-economic status, color, sex, age, etc., is known for the universe. An internal control technique is also usually employed by a comparison of additional characteristics of respondents, obtained during the interview, with their known distribution in the universe; such characteristics include political party affiliation, occupation, education, nationality, etc.

That the stratified-quota method has been generally successful in the prediction of election returns is empirical testimony to its usefulness. It has the advantages of being relatively inexpensive and of quickly yielding the sample result. It is not, however, as fool proof a sampling method as is strict randomization of the basic sampling units of a universe or of a series of sub-universes established by stratification.

Chief Source of Error in Stratified Sampling

The chief source of error in stratified sampling is the failure to obtain a truly random sample of observations or measurements for each control stratum of the universe sampled. We cannot overemphasize the fact that *randomization is the primary control factor in all sampling.* Stratification is a secondary control factor, even though random samples are often referred to in the literature as "simple samples," whereas stratified samples are commonly called "controlled samples."

We have seen that randomizing a population by a lottery method which guarantees the randomness of the sample is often not feasible, and that stratified-random sampling might be utilized in such a situation, with even more satisfactory results. But, we may ask, if we have to draw a random sample for each stratum, why not just draw one large random sample at the beginning and be done with it? The answer is that in consumer and public opinion research, and similar fields, it is usually easier to apply to the sampling the secondary control, stratification, than the primary one, randomization. We randomize the best we can; and to the extent that we fail, we hope that stratification will compensate for the errors.

It is because of this compensation effect that the measures of chance

error in stratified-random sampling are only infrequently adjusted to indicate the greater accuracy or precision that should characterize samples of a given size. The usual estimates of error employed in random sampling are also used in stratified-random sampling in the hope that the stratification has compensated for any failures in true randomization. Only when the method used to randomize stratified samples *guarantees* randomness within each stratum are we justified in *reducing* the statistical estimates of *chance error* in the results. This, however, does not preclude making an internal analysis of control factors and results in order to eliminate possible biases, or *constant errors.*

The "Representativeness" of Stratified Samples

Current terminological emphases often imply that a well-stratified sample is necessarily representative of the universe sampled. Unfortunately, confusion in interpretation has arisen in this regard. We earlier defined a representative sample as one that is a replica, at least to a satisfactory degree, of the universe sampled. This *representativeness* refers to the trait or behavior being studied, and not to some other factor or characteristic. A sample of consumer purchases of a given product is representative of the universe sampled if the sample *result* corresponds to the consumer purchases of the product made by all the members of a universe.

Yet, because a stratified sample is chosen so as to be a cross-section of a universe in regard to several control factors, it is often referred to as a "representative sample." The confusion arises from the use of "representativeness" in both these situations. The stratified sample may indeed comprise sampling units that are a typical cross-section of the universe in secondary control factors such as residence, socio-economic status, age, sex, etc. But it does not necessarily follow that the sample *result* will be representative (within measurable units of error) of the universe. Even when correlation is known to exist between the secondary control factors of stratification and the trait or behavior being sampled, we cannot be as confident of the representativeness of the sample result as when randomization has been employed in selecting the sampling units.

It would be better to call a stratified sample a *typical cross-section* of the sampling units of the universe, and then describe the control factors used in the stratification, than to refer to such samples as *representative samples.*

Finally, it should be emphasized that random-sampling is not necessarily inferior to stratified-random sampling. The latter is often markedly superior per unit cost; and it is more efficient for a given size of total sample, particularly when the sample is relatively small in relation to the size of the universe. However, when a random sample of a universe is feasible, these advantages of stratified-random sampling diminish, particularly if large samples are employed, since the precision of a result varies little in large samples.

F. SOME FURTHER CONSIDERATIONS ABOUT SAMPLING

Precision and Adequacy in Sampling

In the preceding discussion of random and stratified-random sampling we referred to the precision and adequacy of samples, as well as to their representativeness. These terms as used in sampling statistics have rather specific meanings. We earlier emphasized that the representativeness of a sample is rather a matter of degree than of all-or-none, and that the sample is representative of the universe sampled if only *chance* errors appear in the result. If the sampling techniques employed warrant confidence in the likelihood that a sample result is random for a particular universe, the degree to which it is representative is denoted in terms of the precision of the result.

The *precision* of a sample result is evaluated in terms of the extent to which any measure derived from it agrees with the value of that measure for the universe sampled. Thus, if a universe mean is 50.0, the mean of one random sample of that universe is 45.0, and the mean of a second random sample is 49.0, both sample means are representative of the universe mean in that they are derived from random samples of that universe, and the differences can therefore be expected to occur on the basis of chance errors of sampling. The mean of the second sample, however, is a more precise measure of the universe mean than is that of the first sample; hence, the second mean is a more representative measure.

The *adequacy* of a sample result, on the other hand, is based on the concepts of both representativeness and precision. Adequacy is a function of a particular investigation. Thus, a random sample of clerical workers in a large business organization may give results on a clerical efficiency test that will constitute a sample sufficiently representative and precise for that universe to be an adequate basis on which to develop test norms for all the clerical workers in that organization, but which will be inadequate for the development of test norms for clerical workers generally, or in another business organization.

It should be evident that the adequacy of a sample result is contingent upon the methods used in sampling a given universe. Random samples and stratified-random samples of a given universe should always yield sample results that are to some degree representative of that universe. Precision, however, is a function of the size of the sample. Measures of the precision of a result in sampling statistics are developed on the assumption that only chance errors affect the result. The precision ordinarily increases in proportion to the square root of the increase in the size of the sample (cf. Chapter 12, Section E). The effect of increasing the sample size is to reduce the effect of chance errors of sampling.

A sample result based on very few observations or measurements may not be adequate because it is likely to lack the precision necessary for an

investigation. Likewise, it is impossible for a biased sample to be adequate, for it cannot be representative of the universe it was intended to sample.

It should be evident from the foregoing that an attempt to evaluate the precision of an unrepresentative or biased sample is analogous to testing the sharpness of pieces of steel that have not yet been ground into knife blades. The first requisite in evaluating sample results is that the sampling and measurement techniques used are such as will yield a representative result for the universe. Once such results are obtained, they are then statistically evaluated for precision. If a representative sample is sufficiently precise to satisfy the purposes of the investigation, it can be characterized as adequate.

This relationship between the representativeness, precision, and adequacy of a sample result is analogous to that between the validity, reliability, and adequacy of a psychological test (cf. Chapter 17). If a test measures or predicts the kind of behavioral functions it is intended to measure or predict, it is said to be a valid test of those functions (at least to some degree). If, furthermore, it differentiates with a relatively high degree of accuracy the extent to which a group of people manifest such functions, it is said to be a reliable test. Hence it will be adequate for the intended purpose. Thus, if an intelligence test is found empirically to predict the success and failure, at least to some degree, of children in their later academic work, it may be considered a valid test of academic aptitude. If, furthermore, the predictions made from the test results hold with but few exceptions (or errors of prediction), the test is highly reliable. And the more reliable a valid test is, the more adequate it is. But the converse is not true; i.e., it does not follow that an invalid test will become more adequate for the intended purpose if its reliability is increased. The latter would be analogous to attempting to increase the adequacy of a biased sample merely by increasing its size.

The Character of Samples vs. the Size of Samples

These considerations of representativeness, precision, and adequacy in sampling, as well as our earlier discussion of the importance of randomization in scientific sampling, should make it evident that the character of a sample is more important than its size. Not that both are unimportant; rather, *character* denotes the stuff out of which samples are made. However, if the blend is wrong, the character cannot be changed simply by increasing the size of the sample.

The character of a sample, whether or not it is representative of the universe to be studied, ordinarily must be evaluated in terms of (1) a knowledge of the methods used to obtain it, or (2) the extent to which it predicts the behavior expected of the universe, or both.

Once we are warranted in assuming that the character of a sample is satisfactory, we can consider the size of sample necessary for the degree of precision that will be adequate for the purposes of our investigation. If, in a poll of voters' preferences for candidates A and B, we employ a technique that

we can be certain yields a random result for the universe sampled, we must determine how large a sample we need for a result in which we can have sufficient confidence to make a prediction. We can have more confidence in the results obtained with a sample of a given size if 75% favor candidate A and 25% candidate B, than if 52% favor candidate A and 48% candidate B. In other words, the size of sample required for such an investigation depends upon the closeness of the poll result. Larger samples are required in order for smaller differences to be significant.

When the universe is finite, a random sample of a given size will yield a more precise result for a small than for a large universe. Similarly, the more heterogeneous the character of the universe to be sampled, the larger will be the sample required; and conversely, as in the case of the drop of blood, the more homogeneous the character of the universe, the smaller will be the sample required.

The possible usefulness of relatively small but carefully stratified samples in public opinion research is illustrated by the following research from Cantril.* The Office of Public Opinion Research at Princeton University attempted to predict the outcome of the New York gubernatorial election in 1942 by means of a sample of only 200 voters in that state who were selected to provide a cross-section of the voters with respect to color, economic status, and age. All the interviewing was done by one person in the week before election day, the interviews being limited to registered voters. Table 11:6 shows a comparison of the OPOR results with actual election results and with the larger samples used in the American Institute of Public Opinion poll and the New York *Daily News* poll.

Table 11:6

	Gubernatorial Candidates			Number of Cases
	Dewey	Bennett	Alfange	
Actual election results	53%	37%	10%	4,112,000
OPOR prediction	58	36	6	200
AIPO prediction	53	39	8	2,800
N.Y. *Daily News* prediction	57	37	6	48,000

Thus the carefully stratified OPOR sample of 200 voters yielded a prediction almost as precise as the large sample of 48,000 voters used in the *Daily News* poll. Much more consideration was given to the *character* of the small sample than the very large. It will be noted that Gallup's poll for the AIPO, in which the quota sample was carefully stratified but was 14 times larger than that used for the OPOR, predicted the percentage of Dewey's vote without error.

* *Op. cit.*, chap. 12, "The Use of Small Samples."

It must not be concluded from the foregoing example that a small sample of 200 cases for a universe of more than 4 million will generally yield such an adequate result. Rather, this example illustrates the relatively greater importance of the *character* of a sample as against sheer size.

Accidental Samples

The accidental in sampling is often confused with randomness. If a sample is drawn in an unplanned or *haphazard* fashion, if the cases that happen to be conveniently available are used, the sample is most likely to be an accidental, not a random, sample. The technique for the latter type of sample, as we have seen, usually requires a great deal of systematic planning and control in order to guarantee that each member of the universe to be studied will have an equal chance of appearing in the sample.

Accidental samples are *ignorant* samples; they represent no known universe. The methods of sampling statistics have no logical application in an analysis of the results of accidental sampling. Yet this non-sampling device (it really should be characterized negatively, since it is not a technique of sampling) has been all too prevalent in academic research. McNemar * emphasizes this point ironically in his reference to the samples so often used for doctoral dissertations, test standardizations, etc.: "It is here that the college sophomore has an advantage in being the raw stuff out of which psychologists build a science of human behavior."

Following his work with Wundt in the "new experimental psychology" at Leipsig during the last quarter of the 19th century, James McKeen Cattell began to pay some attention to the facts of individual differences. Cattell was not satisfied with the attempt to develop a science of mind by means of intensive observations and measurements with one or two students as subjects, the characteristic approach of the Wundtian experimentalist. As a result, the *number* of subjects was increased. But only recently have psychologists and other social scientists generally realized that sheer numbers are oftentimes not sufficient. The *character* of the sample is of prime importance. Therefore accidental samples, being without character, have no place in scientific research.

Restricted Universes and Partial Investigations

The importance of defining the universe to be sampled has already been emphasized. Adequate samples are not in themselves ever restricted, but they may be samples of restricted universes. Thus, in standardizing psychological tests for personnel work in business and industry, the empirical process of trial and error has made it clear that there is little sense in trying to sample the *generality* of mankind (least restricted universe), or of adults in the United States (somewhat more restricted), or of all adults in industrial work

* Q. McNemar, "Sampling in Psychological Research," *Psychological Bulletin*, 37:331–365, 1940.

(even more restricted), . . . etc. Rather, the relatively restricted universes of a given business or industrial organization are used to develop the norms that will prove most useful for its practical problems of employee selection, upgrading, etc.

The extent to which a universe is restricted is thus relative. The universe of housewives in Colorado is relatively restricted as compared with the universe of housewives in the United States; but compared with the universe of housewives in Boulder, Colorado, it is relatively unrestricted.

A series of observations or measurements drawn from the lower or upper part of a distribution derived from *part* of a universe is sometimes referred to as a restricted sample. The high I.Q. children in a school population would constitute such a group. However, it is better not to employ the concept *sample* in this connection unless the universe to be studied is similarly restricted. Such restricted parts of a whole are better described as *partial groups*, and the study of their traits and behavior as *partial investigations.*

The Analysis of Intra-Group Differences in Sampling

The analysis of similarities and differences in the sample results of subuniverses, particularly in public opinion and consumer research surveys, often gives valuable information as well as considerable insight into the basis for the character of the total result. This is why additional information about the characteristics of respondents is obtained during interviews. Such information not only serves the purpose of determining whether a sample is a typical cross-section in certain respects, but also provides relevant data for the analysis of intra-group differences.

The Gallup poll on the loan to England, mentioned earlier, made the following breakdown of the results for three characteristics of the respondents, viz., occupation, education and political party affiliation:

"England plans to ask this country for a loan of three to five billion dollars to help England get back on its feet. Would you approve or disapprove of the United States making such a loan?"

	Approve	Disapprove	No Opinion	Total
National sample result	27%	60%	13%	(100%)
By occupation				
Business and professional	37%	55%	8%	(100%)
White collar	35	54	11	(100)
Farmers	26	62	12	(100)
Manual workers	20	65	15	(100)
By education				
College	45%	50%	5%	(100%)
High school	32	58	10	(100)
Grammar and no school	22	63	15	(100)
By party				
Republicans	28%	62%	10%	(100%)
Democrats	28	59	13	(100)

Apart from the over-all result, it is evident that at least a majority of each of the three strata groups (occupation, education, and political party affiliation) and the subgroups within each stratum disapproved the loan. Republicans and Democrats alike disapproved in a ratio of more than 2 to 1; in other words, "politics" did not enter into the result. However, there were rather marked differences in the opinions of those with a college education and those with less education; of manual workers and white-collar workers; of farmers and those in business and professional occupations. The more than 3 to 1 opposition of manual workers is particularly interesting in view of the fact that England's Labor government was seeking the loan for the purpose of improving the living conditions of "the working classes."

We have already seen that cross-tabulation of the data of different sampling strata reveals the relative importance of the relations between the stratifying factors and the opinions or behavior studied. Although the above breakdowns of Gallup's results are not in a form suitable for analysis by correlation—unless it can be assumed that the sizes of the subgroups of a stratum are equal—the percentage comparisons of the opinions of each subgroup indicate no correlation between political party affiliation and attitude toward the loan, some correlation between occupation and attitude, and more marked correlation between education and attitude. From the discussion of stratification, it should be clear that occupation and education are relevant for stratification because of the correlation of these characteristics with the respondents' attitudes toward the proposition. Political party affiliation would not be relevant for stratification in this particular case. However, for a question of such public importance and one that would require congressional action, the knowledge as to whether there are or are not differences in the opinions of Republicans and Democrats is relevant.

It is not possible, from these data as presented, to determine whether there is much or any correlation between education and occupation of the respondents. Presumably there is some, because of the usually greater incidence of the high-school and college educated among the white-collar and professional workers and the lower incidence of respondents with such education among farmers and manual workers. The divisions of attitudes among the occupational and educational strata are certainly not incompatible with the possibility of such correlation.

Finally, in analyzing intra-group differences in sample results, it should be emphasized that not only the character but also the size of each such part of a sample must be considered if the differences are to be taken seriously. Subdivisions of a stratum that include only a few cases may be inadequate for satisfactory precision for the particular result.

Sampling in the Experimental Method of Equated Groups

The discussion of sampling in this chapter has been oriented primarily toward the problem of sampling universes in an effort to study them and

ascertain their nature. Sampling has been developed as a useful alternative to a *census* of finite and existent universes, and is the only possible technique for the study of infinite and hypothetical universes. Problems of sampling, however, arise not only in studying the nature of universes and in analyzing intragroup differences, but in using experimental method. Whether the experimental problem is in the field of psychophysics or involves determining the possible effect on working efficiency of such a factor as method of work, frequent rest periods, a financial-incentive plan, or the effect of a drug, consideration must be given to the character of the sample, not merely its size. A particular method of work may improve the efficiency of some people but not others. Homo sapiens, in *general*, has a brain: he thinks and remembers; he sees, hears, feels, tastes, and smells. But far from functioning homogeneously in these respects, homo sapiens is essentially *variable* in the extent to which and the manner in which he utilizes his psychological and social capacities.

The only empirical means of determining the extent to which a psychophysical principle holds for people in general is to test it with many different kinds of people—with men and women, with individuals of different chronological ages, etc.

Determinations like the preceding can often be made by the use of *independent samples*, i.e., samples chosen from different universes or sub-universes, as well as from the same universe in the replication of an experiment. Independent samples are those chosen in such a way that the selection of the units in one sample is in no way affected by the selection of the units in the other. Thus two or more *random samples* are, by definition, independent samples.

Samples matched to establish equated groups, on the other hand, are dependent samples. They are of great value in many experimental problems, particularly in controlled experimentation, which requires the use of a control group and of one or more experimental groups, depending on the problem. Matched sampling of two equated groups consists in pairing the sampling units of each group with respect to factors that are known or thought to be correlated with the experimental variable. Thus, if we wish to determine whether vision affects the quality of work done in a given industrial situation, we might have a sample of these workers examined for vision, and supply glasses where indicated. We could rate the work of those with glasses before and after their vision was corrected, and analyze any differences.* However, as is well known, individual output varies for many reasons and there is often an increase in efficiency with increased experience, for example, regardless of other factors. Therefore, in order to *isolate* the possible effect of the factor of

* The importance of "corrected vision" in certain types of industrial work was well established during the past war at the Sperry Gyroscope Co. in New York. Cf. J. H. Coleman, "The Visual Skills of Precision Instrument Assemblers," *Journal of American Psychology*, 9:165–170, 1945.

corrected vision, the scientific procedure is to match two groups in relevant respects, use one as a control, and subject the other to the conditions of the experimental variable. In this case, relevant factors for matching might be (1) initial condition of defective vision (determined for all subjects by routine eye examinations), (2) length of time on the job, (3) skill or proficiency in work, and possibly (4) age.

The possible effectiveness of a paired matching procedure is analogous to the possible effectiveness of stratified sampling. The factors chosen are *controls* to the extent that they correlate with the experimental variable. If there is little or no correlation, they cannot function as controls. In other words, if age is not correlated with efficiency on the job, it would not serve as a control factor. Furthermore, as in stratified sampling, (1) the principle of diminishing returns is likely to operate as the number of control factors is increased, and (2) it becomes increasingly difficult to match two groups, pair by pair, for more than a few factors.

Matched samples of equated groups are *dependent* samples since the selection of each case in one sample is dependent on a case in the other sample. Such samples are often established as partial rather than purely random samples. That is, a sample of 300 workers doing a given type of job in an industrial situation might be initially selected by a random method from a restricted universe of 3000 workers with *defective vision.* But in attempting to divide and *pair* the 300 subjects in respect to the four control factors mentioned above, a number of them would probably have to be eliminated because of the absence of "mates." In fact, the difficulties of pairing for four factors are such that an experimenter is fortunate if he can establish two samples of 100 cases each, matched pair by pair. The matched groups would thus be partial samples of the original random sample of 300 cases. One group would serve as the control, and the other would be "exposed" to the experimental factor, "corrected vision."

In a well-designed investigation, the working and other conditions for both the control and the experimental groups would be kept as much alike as possible at least pair by pair, if not for each group as a whole. Only the experimental group, of course, would be exposed to the experimental factor, "corrected vision." After an appropriate length of time (this is in itself a variable for different kinds of work and work situations), the efficiency of the two groups would be measured and compared. If there were a *significant* difference in performance in favor of the experimental group, this would indicate a difference greater than could be expected on the basis of chance. Nor could it be attributed to (1) the initial defective vision, (2) length of time on the job, (3) skill or proficiency at the beginning of the experiment, or (4) age, because these factors were controlled by the pairing procedure in matching the two groups. Therefore, the difference would be explained in terms of the experimental variable, "corrected vision," and the importance of this factor in the particular situation would be established.

The function of the control group is thus to give a basis for measuring what the experimental group would have done had it not been subjected to the experimental variable. Except for chance errors, the results yielded by two relevantly matched groups continuously exposed to similar conditions should be similar to each other. But if one condition is changed for one group (its vision is corrected) and it is therefore called the experimental group, the other group (the control) will serve as the yardstick on which to measure the behavior of the experimental group had it not been exposed to the experimental factor (corrected vision).

It should be clear from the preceding that there could be only a *presumption* of the *generality* of the effectiveness of corrected vision in *all types* of working situations. Whether such a factor would be effective in working situations other than the type used in the particular experiment could be definitely determined only by sampling other types of universes (i.e., other types of working situations).

Experimental Method with Random Samples

We said above that the equated group technique in experimental method is analogous in *control* respects to the stratified sampling technique. There is also an experimental technique that is analogous to purely random sampling, in which the basic control is the randomization of a universe. Such samples are independent rather than dependent, and hence are uncorrelated.

A satisfactory control and experimental group can be set up by drawing two random samples from the same universe. They could be "matched" in only one basic factor, which, however, is the primary *control* factor of *randomization.* Any difference between two such groups at the beginning of an experiment should be no greater than would be expected on the basis of *chance* itself. Hence, any difference between them at the conclusion of the experiment, greater than could be expected on the basis of chance, would be attributable to the experimental variable, provided other conditions for both groups were kept similar during the experiment.

This procedure is often used to establish groups matched only in the sense that they are random samples from the same universe; both are then exposed to experimental variables. Thus, in "split-run" copy testing of two advertisements, in which one ad is printed in half the press run and the other ad is printed in the other half, both ads being identical in all respects except for one experimental variation—the headline, for example—each is exposed to two random samples of a universe.* The universe may be the entire circulation of a magazine. It can be randomly divided into halves by *alternating* the two advertisements in the particular issue of the magazine. By means of a free-sample device "buried" in small print in a relatively less important part

* J. Zubin and J. G. Peatman, "Testing the Pulling Power of Advertisements by the Split-Run Copy Method," *Journal of Applied Psychology,* **29**:40–57, 1945.

of the ad, the relative pulling power of each ad can be evaluated in terms of the number of replies to each one. Any difference in the number of replies, greater than would be expected on the basis of chance, can be attributed to the difference in the particular copy that was experimentally varied. The copy "pulling" the greatest number of replies is more effective. However, whether it would also be more effective for other universes (i.e., other types of magazines, newspapers, display advertising, etc.) could be determined only by sampling and testing it with these other universes.

G. SOME TERMINOLOGICAL DISTINCTIONS FOR SAMPLING AND ANALYTICAL STATISTICS

In addition to the concepts and techniques for the sampling of statistical universes which have been developed in the preceding sections, there are several distinguishing concepts and symbols in sampling and analytical statistics that should be noted at this point.

Parameters and "True Measures"

Any measure of the distribution of a *statistical universe* or *population* is called a *parameter*. Percentages, means, standard deviations, centile values, correlation coefficients, and similar measures derived from the data of universes are parameters.

In the case of infinite populations, parameters are purely hypothetical measures, but their value can at times be closely estimated from large samples of observations. For some finite populations, actual parameter values can be computed. However, it should be borne in mind that such values are always subject to errors that occur in the process of observation or measurement itself. Consequently, although by definition they are parameters, such values can hardly be described as "true measures." The concept *true measure* implies an errorless parameter value. Although such measures are obviously hypothetical, statistical techniques have been developed that will yield an estimate of them, with the attenuating effect of errors of observation or measurement theoretically eliminated.

Some statisticians use the terms *parameter values* and *true measures* synonymously. However, as the two terms are defined here, a parameter value is not necessarily a true measure. Only if it is errorless is the parameter, by definition, *a true measure*.

Statistics

Any measure derived from the data of a sample is called *a statistic*. In practice, statistics are often referred to as *obtained measures*. However, not all obtained measures are necessarily statistics, for a measure may be obtained for an entire finite population, in which case it is a parameter rather than a statistic.

The measures obtained in the reduction of *sample data*, such as percentages, means, standard deviations, centiles, correlation coefficients, etc., are statistics. The concept *statistic* is used to differentiate sample measures from those of universes.

Many of the techniques of sampling and analytical statistics have been developed in order to enable the estimation of the values of parameters from known values of statistics. As we have seen, it is of the essence of sampling and analytical statistics that universes be studied by means of the "evidence" of sample data.

Symbols for the Differentiation of Parameters and Statistics

Since most of the measures actually used in sampling statistics are *statistics* rather than parameters, it is usually sufficient to differentiate only the latter. For this purpose we shall employ the subscript u, for *universe*. Thus, the mean of a statistical universe, and hence its parameter value, will be signified by M_u. Similarly, parameter values of centiles, standard deviations, and similar measures, will be signified by C_u, σ_u, etc. The subscript h will be frequently used to denote a *hypothetical* value of a parameter. When it is necessary to distinguish a *statistic*, the subscript s will be employed.

Some authors have drawn on the Greek alphabet for symbols for parameters and on the English alphabet for symbols for statistics. This practice, if universal, would no doubt offer the easiest way of clarifying the present situation in statistical terminology, which is admittedly confused. However, many Greek symbols already have established usages in descriptive and sampling statistics. Thus, the Greek *rho* has long been used to symbolize Spearman's rank-difference correlation coefficient. Its use to represent a universe r would be confusing, as would also the use of σ to represent only the standard deviations of universes.

Sometimes a sample of a variable must be differentiated from the universe of the variable. Thus, if the variable under consideration consists of intelligence quotients and the variable itself is symbolized by *I.Q.*, then a sample of the variable will be differentiated from the universe by the same subscript, s, as is used to distinguish a *statistic*. $I.Q._s$ will refer to the sample of the variable, and $I.Q._u$ to the universe. If a variable in a formula is algebraically symbolized by x or y, a sample will be designated by x_s to differentiate it from x_u, or x_h. Similarly, the number of cases in a sample will be differentiated from the number of cases in a universe by N_s and N_u.

Sampling Distributions

Just as the data of a sample are called sample data, so the distribution of a given statistic, such as a mean obtained from a series of samples, is called a sampling distribution. Sampling distributions are composed of the values of a given type of statistic derived from a series of random samples of uniform

size from the same universe. If we drew from a given universe 100 random samples, each consisting of 1000 cases, a distribution of the 100 means of these samples would be an empirical sampling distribution of the particular statistic, namely, the *mean*.

A sampling distribution of a statistic is different from a distribution of the data of a sample. The latter is the ordinary distribution of the *frequencies* of a single sample.

The concept *sampling distribution* is integral to the methods of sampling and analytical statistics because the study of universes from the data of samples is based on definite assumptions about the form of sampling distributions, as well as upon estimates of the extent of their variation. For example, many of the computations and estimates of analytical statistics are based on the assumption that the form of the sampling distributions of various statistics is similar to that of the normal probability curve. In other cases, especially in very small samples, sampling distributions have a different form. Thus, for a series of samples of less than 25 or 30 cases each, the form of the sampling distribution of any statistic derived therefrom will skew more and more from the normal probability curve, the smaller the size of the sample. Sampling distributions of product-moment correlation coefficients also are extremely skewed as this statistic approaches values of 1.0 or −1.0.

Small Sample Theory vs. Large Sample Theory

The concept *large sample theory* is used in statistical method to designate sampling distributions, and the techniques developed for them, that are sufficiently large and of a character to yield distributions approximating the normal probability curve. The concept *small sample theory*, on the other hand, is used to describe the sampling distribution and related concepts derived for statistics of samples that are so small in size (less than 25 or 30 cases) as to yield sampling distributions which definitely diverge from the normal probability type. It should be observed, however, that not all sampling distributions for statistics derived from large samples necessarily yield sampling distributions that are normal in form. Correlation coefficients that approach 1.0 or −1.0, just cited, are an example. Such statistics as these are often derived from very large samples, but the parameters themselves are of such a character as to yield sampling distributions of a form other than the normal type.

In practice, sampling distributions are usually hypothetical concepts. In other words, it frequently happens that only the data of one or two samples are available in a research investigation, and consequently, both the form of the sampling distribution and the measure of the extent of its variation have to be estimated. Many of the mathematical procedures in sampling and analytical statistics have been devised for just this purpose, on the assumption that the sampling distribution has a definite form and is based on *random sampling* or *stratified-random sampling*—another reason for emphasizing the

importance of random sampling in analytical statistics. In fact, no sound inferences or conclusions about specific universes are possible from a study of samples derived therefrom unless the latter have been obtained by the method of random or stratified-random sampling. Only when a sample is known to have been drawn randomly and to have been derived from the kind of universe which yields a definite type of sampling distribution can inferences be made about the universe *with confidence.*

The Standard Error of a Statistic

The standard error of any statistic is the standard deviation of a sampling distribution of that measure. In practice, it is often necessary to estimate this deviation from only one or two samples. For some statistics, however, relevant hypotheses about the universes investigated are of such a kind as to enable a rather precise estimate of the standard deviation of the hypothetical sampling distribution of a given statistic.

Statistical Hypotheses

The mean of a sampling distribution is, as we shall see in Chapter 13, usually established by hypothesis, in which case it is a parameter value. For example, in an attempt to determine from a random sample of purchasers of X cigarettes whether the buyers are evenly divided between men and women, we test the statistical hypothesis of a mean proportion of .50 for each sex group. These hypothetical proportions would be the parameter values of the universe studied if the purchasers of X cigarettes were in fact evenly divided between men and women. The research problem is to determine whether a random sample yields a result that differs significantly (greater than that to be expected on the basis of chance) from the hypothetical values of the universe sampled, viz., a proportion of .50 for each sex group. If the random sample of purchases, divided among men and women, does differ significantly from these parameter values, we can confidently reject the hypothesis that the universe of men and women purchasers of X cigarettes is evenly divided.

In this type of test for a statistical hypothesis, we said that *mean* values are often established by hypothesis. In the preceding example, the statistical problem of estimating the extent of the variation in the hypothetical sampling distribution of mean proportions is based on equations derived from the hypothesis. However, in other types of problems in sampling statistics, the sample data themselves must be utilized in estimating the degree of this variation. In any event, the measure of variation used is taken in terms of the standard deviation, and the standard deviation of any sampling distribution is called the standard error of the statistic in question.

The standard error of a measure is also symbolized by σ with an appropriate subscript. Thus, σ_M symbolizes the standard error of a mean; σ_p of a proportion; σ_r of a correlation coefficient, etc.

The Probable Error of a Statistic

The measurement of the variability of a sampling distribution is sometimes taken in terms of *P.E.*, the probable error. This measure, however, is based on the standard error, and for sampling distributions that are distributed according to the bell-shaped, normal probability curve, *P.E.* is always equal to .6745 of the standard error, i.e., about two-thirds as large. Historically, *P.E.* has been used as a measure of the variability of normal, bell-shaped sampling distributions because the limits of $M \pm 1$ *P.E.* mark the range of the middle 50% of the distribution. Hence, in such a distribution a statistic will have an equal chance of being within or beyond these limits.

Sampling Error and Error of Measurement

The concept *sampling error* is technically used in statistics to denote the difference between the value of a parameter of a universe and the value of the statistic derived from a sample of that universe. Since truly random samples of a universe are affected only by chance errors of sampling, any measure of the sampling error is a measure of the probable effect of chance errors on a statistic. As previously emphasized, a measure of sampling error is based on the assumption that the sample itself is unbiased.

On the other hand, *errors of measurement* are analogous to the physicist's errors of observation. They are the errors that occur in connection with the procedures of observing, counting, making measurements, etc. Their effect on a sample result can often be estimated, and hence allowed for, if it can be assumed that they are randomly distributed over a series of observations. If these errors of measurement occur randomly, then a sampling distribution of them will tend to be distributed according to the normal curve of error (the normal, bell-shaped probability distribution). The larger the series of observations or measurements, the less the effect of such errors on the sample result and on most statistics derived from it. This is so, because if they are distributed randomly, they will tend to balance each other and hence cancel out in their net effect on the value of such a statistic as a mean or a proportion. But in a few cases, chance errors of measurement distort the value of a statistic; for example, they tend to decrease the size of a correlation coefficient. However, on the assumption that such errors are randomly distributed, mathematical techniques are available that will permit an estimation of the attenuating effect of such errors on a result; and therefore their effect can often be allowed for in evaluating a result.

EXERCISES

1. What is the difference between a census and a sample? What circumstances require the use of samples rather than of censuses?
2. Define a statistical population or universe and give several examples in research of (a) finite and (b) infinite universes; of (c) actual and (d) hypothetical universes.

3. What kinds of errors necessarily are present in all sampling? Why?
4. What kinds of errors should be avoided in sampling? Why?
5. Define a *representative* sample result. Are sample results derived from a cross-section or stratification of a universe necessarily representative of the universe sampled? Why?
6. Define a biased sample and state the kinds of factors or circumstances that make for bias in sampling. Give several examples of these factors.
7. Define a random sample and a stratified-random sample. What do these two types of samples have essentially in common? What is the basic difference between them?
8. Describe the techniques employed to insure the randomization of samples.
9. Define a sampling unit and distinguish between initial or primary sampling units and the sampling unit per se.
10. State the procedures by which a stratified-random sample of a given size can yield a more adequate result than a random sample of the same size.
11. What kinds of research information can be derived from a stratified-random sample that ordinarily are not obtainable from a purely random sample?
12. What are *internal controls* in sampling and how are they utilized to make the result more adequate?
13. On what fundamental assumptions and principles is the technique of areal sampling based? What is the essential difference between areal sampling and master sampling, and under what circumstances can the latter produce adequate results?
14. Define the random-point method of sampling and describe the difficulties inherent in it.
15. Define the stratified-quota method of sampling and discuss its advantages and disadvantages. Is this method the same as stratified-random sampling?
16. What is the primary *control* factor in all sampling techniques?
17. Distinguish between the *representativeness* and the *precision* of a sample result, and define an *adequate* sample.
18. Why is the character of the sample a more important consideration than its size?
19. Distinguish between random samples and accidental or ignorant samples.
20. In sampling, why is it necessary to study restricted universes? What is the difference between the sampling of a restricted universe and a partial investigation?
21. Describe the nature of the sampling procedures used in the experimental method of equated groups. What is the essential difference between dependent and independent samples?
22. What is the technical distinction between a parameter and a statistic? In what sense is it misleading to describe a parameter as a " true measure " ?
23. What is the difference between a distribution of a sample and a sampling distribution?
24. Under what circumstances is the standard deviation of a distribution called the standard error?
25. What is the difference between sampling errors and errors of measurement?
26. Contrast the research procedures used to control errors of measurement and sampling errors.

CHAPTER 12

Probability and Statistical Inference

A. THE STATISTICAL CONCEPT OF PROBABILITY

If the individuals of a large universe are distributed in the proportions of exactly .50 females and .50 males, the *most likely* result of many *random* samples of, say, 1000 persons per sample drawn from that universe will be 50% females and 50% males. These are the probability values of the occurrence of two kind of events, females and males, in this particular universe. The *P* (probability) value for males (or for females) is expressed as .50, or 50 in 100, or as the ratio 1/2.

We would not expect a single sample, or only a few samples, to give exactly 50% females and 50% males, because of the operation of *chance* errors in sampling. In fact, some result other than .50 and .50 is more likely, because many different proportions of males and females are possible, purely on the basis of *chance* in random sampling; .50 and .50 represent but one combination of these possible results. Although this particular combination is more likely to occur than any other *one* combination (say, .45 females and .55 males), it is not more likely than all the other possible combinations taken as a whole.

Definition of Probability

The probability of an event is defined as the relative frequency of that event in all possible events of the class or universe under consideration. This is the *frequency theory* definition of probability. Although historically there have been other definitions, the above is basic to the statistical utilization of the implications of probability. It is based on the assumption that the instances or events of a universe are *indefinitely repetitive*. As von Mises says, "We call the probability of an attribute (experimental result) in a collective [universe] the limiting value of the relative frequency with which this attribute recurs in an indefinitely prolonged sequence of observations." *

In terms of the preceding example, the limiting value of the relative frequency of males to persons (both males and females) should be .50, or 1 in 2, as a sequence of random samples is indefinitely prolonged for a universe of people evenly divided with respect to sex. If the parameter proportions of the attributes of a particular universe are not known (as is ordinarily the case), the proportion of an attribute (males) would be empirically determined by

* R. von Mises, *Statistics, Probability and Truth*, Macmillan, New York, 1939, p. 308.

drawing a prolonged series of samples (of persons) at random from that universe. If in this series the proportion of males approached .50 *as a limit* (the limiting value), the P value for males would be taken as equal to .50, or 1/2.

The probability, P, of an event (or attribute) is expressed algebraically as the following ratio:

$$P = \frac{p}{p+q} \qquad \text{[12:1] Probability ratio}$$

where p equals the relative frequency of occurrence of the given type of event (or attribute) in a class of events; q equals $1 - p$; and $p + q$ is equal to unity, the total of all possible events in the class, ordinarily represented by 1.00 or 100%.

If the universe of persons referred to above is taken as a finite universe consisting of 100,000,000 members, and the P value for males is .50, this will mean that

$$P = \frac{f_p}{f_p + f_q} = \frac{50{,}000{,}000}{50{,}000{,}000 + 50{,}000{,}000} = \frac{1}{2} = .50$$

where 50,000,000 in the numerator equals p, the frequency of males, and the sum of the two figures in the denominator represents the total frequencies for the universe.

If the universe of persons is taken as infinite, and an indefinitely prolonged series of observations (samples) of that universe yields .50 as the *relative frequency* of males,

$$P = \frac{p}{p+q} = \frac{.50}{.50 + .50} = \frac{1}{2} = .50$$

where $p = .50$ is the relative frequency of males in all possible events or attributes (males plus females) in the class of events (persons) in the universe. And $P = .50$ is the limiting value of the relative frequency with which males recur in an indefinitely prolonged series of observations of the sex of persons in a theoretically infinite universe.

A Single Event Has No *P* Value—The Concept of Likelihood

When we speak of the probability of a single event, we use the singular only metaphorically. From the point of view of the frequency theory of probability, a single instance or occurrence has no probability value. The probability of a single occurrence cannot be determined by the probability calculus. The implications of the theory of probability concern what happens *on the average*, or *in the long run*.

It might be argued that we can indicate the probability of rain on a given afternoon on the basis of our having observed weather conditions only a few hours earlier. However, if it looks like rain and we say the probabilities are that it will rain, we mean that rain appears to be very *likely*, i.e., that all the relevant evidence (observations, etc.) warrants the judgment or con-

viction that this event will occur, rather than that the P value of rain is 90 in 100, or any other figure indicative of high probability. If, in tossing a coin, we say that the probability of its landing heads up is 1 in 2, we are using the probability concept to express our ignorance of how it will land. *In the long run* we can expect heads half the time, but we cannot predict what will happen in a single toss. The laws of chance cannot be used to compute a probability value for the behavior of a single occurrence or unique event; rather, they can be used to forecast what will happen on the average, or in the long run, for mass phenomena. And they may be utilized to judge what is *likely* or *unlikely*, on the basis of chance, for a single sample result.

Unfortunately, the lay use of the concept probability is confused with the concept "likely." In statistical inference, we shall see that both these concepts are needed, that they have distinctive meanings, and that the implications of each are essential to many types of problems. A single event has no calculable probability value, but its possible occurrence may be judged to be *likely* or *unlikely* on the basis of knowledge of the behavior of and antecedent conditions in that particular class of events of which it is a member.

Strict Causality vs. Statistical Relations

On the other hand, the occurrence of a single event that is strictly determined by attendant or antecedent conditions is considered to be *certain*, rather than *highly probable* or *very likely*. Strict causal relations are characterized as *If-x-then-y* relationships, whereas statistical relations (and by definition they are not strictly determined) have a probability value for what is expected to happen on the average, or in the long run. Statistical relations may be characterized as *If-x-then-y-is-likely* relations. In statistical inference it is usually necessary to judge for a particular event or sample of observations whether, under circumstances x, y is or is not likely to occur.

Many physical laws are expressions of *If-x-then-y* relations; an example is the relation between the temperature at which water boils and atmospheric pressure. Statistical laws or relations, on the other hand, are expressions of *If-x-then-y-is-likely* relations. Thus, a sample of persons may yield a proportion of males (y) that is *likely* to be *representative* of (not identical with) the parameter proportion of males, *if* conditions x are satisfied, viz., a large sample is drawn randomly from the universe under consideration. Or, in the correlation of bi-variates, such as height and intelligence, we may find little or no relationship in a sample result and therefore be warranted in concluding that *no relationship* for the universe is *likely*.

The regression equations of correlation are a direct expression of *If-x-then-y-is-likely* relations:

$$y = (f)x$$

where the function is calculated in terms of the regression coefficient, $r_{yx}\frac{\sigma_y}{\sigma_x}$. Only if r_{yx} were 1.00 would there be an *If-x-then-y* relation, as in the relation

between the diameters and areas of circles. In natural and social phenomena, however, the correlation of bi-variates yield *If-x-then-y-is-likely* relations. The statistical problem is to determine the *degree* of the relation for a sample result, estimate the probable effect of chance errors on the result, and then judge what is or is not *likely* to occur in the light of all relevant information about the result and of experience with the phenomena under consideration.

B. THE BINOMIAL DISTRIBUTION AND THE NORMAL PROBABILITY CURVE

Normal Sampling Distributions

Sampling distributions for many statistics are normally distributed. The normal probability curve (Fig. 12:2) describes the way in which *chance errors* of observation and measurement affect a result. Such errors are not mistakes; rather, they represent the effect, on the behavior or quality being studied, of innumerable factors that are as likely to affect a result positively as negatively, favorably as adversely.

The normal curve describes the form of the distribution of many qualities or traits of organisms, particularly those that are the consequences of innumerable determiners in genetic development and growth. It is as if the determiners operate to produce a *combination of results* that are distributed according to the laws of chance—some favorably, others unfavorably. Some combinations are rare; others are fairly frequent; still others are very frequent at the average. Distributions of *I.Q.* scores and of less generalized abilities, and distributions of such anthropometric traits as height and weight have been found in certain types of populations to resemble the normal probability curve, provided the populations are not too heterogeneous as regards such factors as race, age, and sex.

Some insight into the relationship between the laws of chance and the normal probability curve can be obtained from a consideration of the point binomial $(p + q)^n$, in which p represents the probability of the occurrence of an event in a class, q represents the probability of its non-occurrence, and n is the size of the sample, N_s. For illustrative purposes, we shall continue to use the above example of a universe in which the sexes are evenly divided. The probability, P, of the occurrence of males, p, in random samples will be equal to the ratio 1/2, or $P = .50$; and q, the probability of the non-occurrence of males, will also be 1/2, or $P = .50$. Because the only alternative to the occurrence of males is the occurrence of females, q of the binomial represents females.

Equiprobability

Since the probability of either p or q is 1/2, or .50, p and q are *equiprobable.* In other words, males are just as likely to be drawn in the samples as females, and vice versa. The *equiprobability* of events underlies the normal probability

distribution. As was said in the preceding chapter, *chance errors* are errors that are just as likely to occur as not to occur; hence they are equiprobable. In drawing samples of persons from a universe composed of an equal number of males and females, the principle of equiprobability should operate provided truly random samples are drawn. Let us see, first, what the results will be if small samples of 2 persons each are drawn, i.e., $N_s = 2$.

Binomial for Samples of $N_s = 2$

When $N_s = 2$, the result of any random sample of persons will consist of one of the following three alternatives:

2 males
1 male and 1 female
2 females

These three alternatives or variations are the only results possible with the universe sampled, 2 people per sample. Any of them may be obtained on the basis of *purely chance factors* in sampling. However, the combinations themselves are not equiprobable even though the probability of the occurrence of a male is 1/2 and of a female is 1/2. The three combinations are not equiprobable because one of them can occur in two different ways. In drawing a random sample of 2 persons, any one of the following four arrangements may be obtained: (1) m and m; (2) m and f; (3) f and m; (4) f and f. The *combination* of 1 male and 1 female can thus be obtained in two different ways, whereas the combination of 2 males or of 2 females can be obtained in only one way. The different ways in which a given combination can occur are known as its *permutations*.

When p and q are equiprobable, the probability of any particular combination is therefore the ratio of the number of different ways it can occur *to* the total number of ways all the different combinations can occur. In this case the probability of the combination 1 m and 1 f is 2/4 or 1/2, whereas the probability of 2 m is 1/4, and that of 2 f is also 1/4. These values are readily obtained by the following expansion of the binomial:

[12 : 2]

$$(p + q)^n = (p + q)^2 = p^2 + 2pq + q^2$$ Binomial expansion when $n = 2$

where n, the power of the binomial, is equal to the size of the sample, N_s.

We have assumed that the p and q events in the sample are equiprobable. Their respective P values are therefore 1/2. Hence, if we substitute their P values in the preceding equation, we have:

$$(\tfrac{1}{2} + \tfrac{1}{2})^2 = (\tfrac{1}{2})^2 + 2(\tfrac{1}{2})(\tfrac{1}{2}) + (\tfrac{1}{2})^2 =$$
$$\tfrac{1}{4} + 2(\tfrac{1}{4}) + \tfrac{1}{4} = \tfrac{1}{4} + \tfrac{1}{2} + \tfrac{1}{4}$$
$$P = \underset{(2\,\mathrm{m})}{.25} + \underset{(1\,\mathrm{m},\,1\,\mathrm{f})}{.50} + \underset{(2\,\mathrm{f})}{.25} = 1.0$$

The expansion of the binomial thus expresses the probability of any possible combination of result occurring in a prolonged series of random samples. The *number* of different possible combinations is equal to $n + 1$. The total number of different ways (permutations) in which all possible results can be obtained, when p and q are equiprobable, is equal to 2^n. And the probability of any particular combination is the ratio of its number of permutations to all possible permutations, the number of permutations for any combination being indicated by the coefficient of the binomial term for the particular combination, provided $p = 1/2$, and the number of all possible permutations being indicated by 2^n. Where N_s, the size of the sample, is equal to 2, the number of different *combinations* of possible results is

$$(n + 1) = 2 + 1 = 3$$

the total number of possible permutations, when p and q are equiprobable, is

$$2^n = 2^2 = 4$$

And, as already indicated, the probability value of each of the three combinations is as follows:

2 males, $P = .25$; 1 male, 1 female, $P = .50$; 2 females, $P = .25$

In the long run, for an indefinitely prolonged series of random samples, 2 persons per sample, we would expect to obtain the results shown in *A* in Fig. 12:1. This is the theoretical *sampling distribution* of a point binomial in which the p and q events are equiprobable and the size of the sample is 2. One-quarter, or 25%, of the sample results will consist of 2 males (or $2p$); one-half, or 50%, will consist of 1 male and 1 female (or pq); and one-quarter, or 25%, of 2 females (or $2q$).

It will be observed that the distribution for $N_s = 2$ and $p = 1/2$ in Fig. 12:1 is uni-modal and bilaterally symmetrical. This distribution thus has two of the essential properties of the standard, normal probability curve. However, it is a *discrete* rather than *continuous* distribution, and is considerably more *peaked* than the normal probability curve.

The mean frequency of males (or p events) for a binomial distribution is as follows:

$$M_f = N_s p \qquad [12:3]$$

Mean frequency of p events in a binomial distribution, $(p + q)^n$

where M_f is the mean frequency; N_s is the size of the sample; and p is, by knowledge or *by hypothesis*, the *proportion* (or probability value) of p events in the universe under consideration. In this case, $p = 1/2$, or .50. Hence, the mean frequency of males in the sampling distribution in Fig. 12:1 is 1.0:

$$M_f = 2(.50) = 1.0$$

The standard deviation of a distribution of *frequencies* of p events for a binomial distribution is as follows:

[12 : 4]

$$\sigma = \sqrt{N_s pq}$$

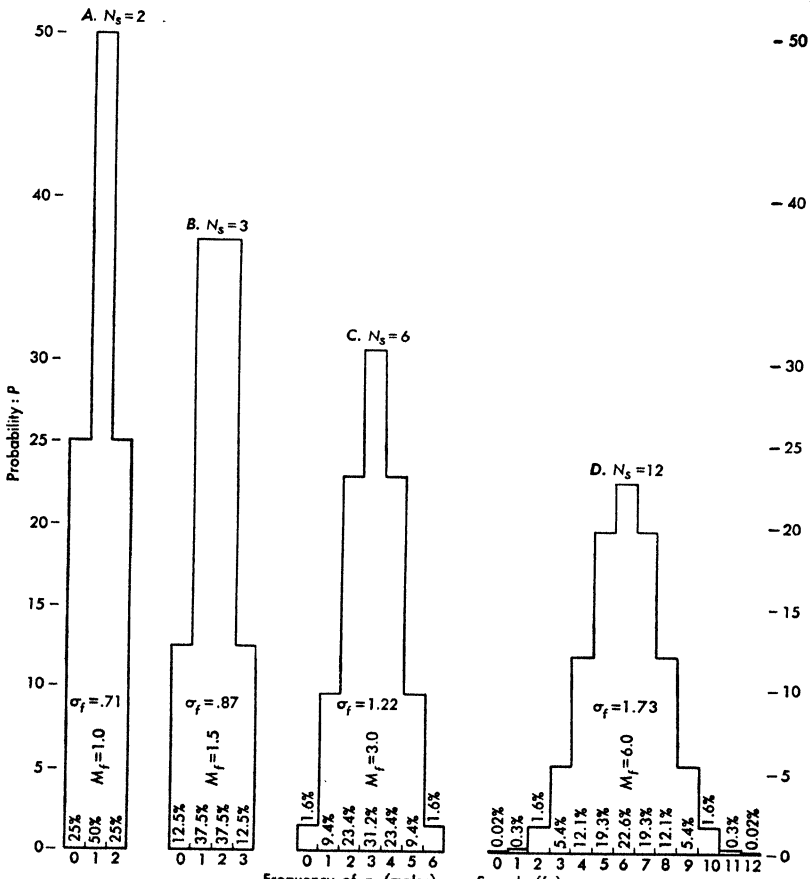
Standard deviation of the frequency of p events in a binomial distribution

where N_s is the size of the sample, p is the proportion (or probability value) of p events, and q is $1.0 - p$.

When $N_s = 2$ and $p = .50$,

$$\sigma = \sqrt{2(.50)(.50)} = \sqrt{.50} = .707$$

Fig. 12:1. The Theoretical Sampling Distributions of the Binomial $(p+q)^n$, When p and q Are Equiprobable; with the Size of the Samples, N_s, Equal to 2, 3, 6, and 12, and with the Total Areas of Each Distribution the Same. (Plotted from data in Table 12:1)



Probability Distributions

The four binomial distributions in Fig. 12:1 have been drawn so that the effect of an increase in the size of the sample, N_s, on the *form* of the sampling distribution may be seen. The possible frequencies of p events (males), per sample result, are scaled on the abscissa. The ordinates are scaled in *percentage* of sample results. As indicated in Table 12:1, which contains the data for the distributions in the figure, the percentage of sample results for each possible frequency of p events (males) in random samples coincides with the P value (probability value) of each type of possible sample result. The distributions in the figure are thus probability distributions, and they are scaled so that

Table 12:1. Theoretical Percentage Distributions and P Values of Sample Results for the Binomial When Males (p) and Females (q) Are Equiprobable and the Size of Random Samples, N_s, Is 2, 3, 6, and 12

Frequency of Males per Sample	% Frequency of Sample Results	Probability Value (P)
A. (When $N_s = 2$)		
2 males	25%	.25
1 male	50%	.50
0 male	25%	.25
Total =	100%	1.00
B. (When $N_s = 3$)		
3 males	12.5%	.125
2 males	37.5%	.375
1 male	37.5%	.375
0 male	12.5%	.125
Total =	100.0%	1.000
C. (When $N_s = 6$)		
6 males	1.56%	.0156
5 males	9.38%	.0938
4 males	23.44%	.2344
3 males	31.25%	.3125
2 males	23.44%	.2344
1 male	9.38%	.0938
0 male	1.56%	.0156
Total =	100.0 %	1.000
D. (When $N_s = 12$)		
12 males	0.024%	.0002+
11 males	0.29 %	.0029
10 males	1.61 %	.0161
9 males	5.37 %	.0537
8 males	12.08 %	.1208
7 males	19.34 %	.1934
6 males	22.56 %	.2256
5 males	19.34 %	.1934
4 males	12.08 %	.1208
3 males	5.37 %	.0537
2 males	1.61 %	.0161
1 male	0.29 %	.0029
0 male	0.024%	.00024
Total =	100.0 %	1.000

each has the same total area on the chart. The total area is taken to a base of 100 (for per cent), or 1.0 (for P values expressed as proportions).

The Product and Addition Theorems of Probability

Before considering further the effect of an increase in sample size on the form of the binomial distribution, we should be familiar with two of the theorems of probability that underlie the general theory of probability. They are the product theorem and the addition theorem.

The *product theorem* states that the probability of the joint occurrence of two or more independent events in a class is equal to the product of their respective probabilities. Thus,

$$P_{(a \cdot b \cdot c \ldots n)} = P_a \cdot P_b \cdot P_c \cdot \ldots \cdot P_n \qquad [12:5]$$

Probability of the joint occurrence of independent events

In the foregoing example the probability of a single p event (the occurrence of a male) is 1/2, or .50. Hence in random samples when $N_s = 2$, the P value of the joint occurrence of $2p$ events (2 males) is $(1/2)(1/2) = 1/4$, or .25. This was the value obtained for $2p$ (males) in the expansion of $(p + q)^n$, when $p = 1/2$ and $n = 2$. Similarly, the P value of zero males (2 females) is .25. The P value obtained for any single combination of results in the expansion of the binomial is, therefore, based on the assumption that the members of a sample result are drawn independently of each other. It makes no difference whether they are drawn simultaneously or successively (as in most sampling), so long as they are drawn independently. This independence is implicit in the principle of randomization. However, the P value of any given combination of results is also based on the addition theorem for the probability of alternative events.

The *addition theorem* states that the probability of two or more alternative (or disjunctive) events is equal to the *sum* of their respective probabilities. Two disjunctive events are mutually exclusive; i.e., they cannot occur simultaneously in a sample result; they are "either-or" events.

$$P_{(a+b+c+\ldots+n)} = P_a + P_b + P_c + \ldots + P_n \qquad [12:6]$$

Probability of the occurrence of disjunctive events

We saw that there are two ways of obtaining the combination 1 male and 1 female, and that the P value of this result was 1/2, instead of 1/4 as in the case of 2 males (or 2 females). The addition theorem can be applied to this result. Although the combination 1 male and 1 female can be *obtained* in 2 ways when $N_s = 2$, viz., male and female, or female and male, the combination can be obtained in only one way in a single sample result. Further-

more it makes no difference which permutation * actually occurs because the character of the sample result is the same, i.e., 1 male and 1 female, regardless of the order in which the two component members occur. Since the P value of a single permutation is 1/4, $[(1/2)(1/2) = 1/4]$, and since the permutations of a given combination of results are actually *alternative* (or disjunctive) ways in which the result can be obtained, the P value of the combination is the sum of the probabilities of all the different permutations that can yield the particular combination. Thus,

$$P_{(a+b)} = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}, \text{ or } .50$$

This is the same as the probability of *a or b*. Thus,

$$P_{a \text{ or } b} = \tfrac{1}{4} + \tfrac{1}{4} = \tfrac{1}{2}, \text{ or } .50$$

These two theorems enable us to determine the probability of the occurrence of alternative combinations when we know the probability of each. Thus, when $N_s = 2$ and the probability of $p = 1/2$, the probability of a sample result that will contain *at least* one p event (male) is

$$P_{(2 \text{ males})} + P_{(1 \text{ male})} = \tfrac{1}{4} + \tfrac{1}{2} = \tfrac{3}{4}, \text{ or } .75$$

where 1/4 is the P value for the combination of 2 males, and 1/2 is the P value of the combination of 1 male and 1 female. *Either* of these results will give a sample containing 1 male, *but* only *one* of these combinations can occur in the same sample result; hence they are exclusive alternates. The addition theorem gives 3/4 or .75 as the probability value of such a result.

From the foregoing, we see that each term of an expanded binomial represents a type (or combination) of result, and that the respective P values of each are based on both the product and the addition theorems of probability.

Binomial for Samples of $N_s = 3$

If, instead of taking random samples of 2 persons at a time, we increase the size of each sample to 3 cases, the binomial is expanded as follows:

$$(p + q)^3 = p^3 + 3p^2q + 3pq^2 + q^3 \qquad [12:7]$$

Binomial expansion when $n = 3$

There are four possible combinations $(n + 1 = 4)$ of males and females, as follows: 3 males; 2 males and 1 female; 1 male and 2 females; and 3 females. It will be observed that the composition of each possible combination is indicated by the *powers* of p and q in each term in the binomial. Thus, p^3 represents the combination of 3 males; p^2q, the combination of 2 males and 1 female, etc. As already indicated, the coefficients of each term give the number of permutations for each combination, i.e., the number of ways in which each combination may be obtained when p and q are equiprobable.

* The different ways in which a given combination can occur are known as its permutations.

For an indefinitely prolonged series of random samples, where p and q are equiprobable and $N_s = 3$, we would have

$$\begin{aligned}(\tfrac{1}{2} + \tfrac{1}{2})^3 &= (\tfrac{1}{2})^3 + 3(\tfrac{1}{2})^2(\tfrac{1}{2}) + 3(\tfrac{1}{2})(\tfrac{1}{2})^2 + (\tfrac{1}{2})^3 \\ &= \tfrac{1}{8} + 3(\tfrac{1}{8}) + 3(\tfrac{1}{8}) + \tfrac{1}{8} \\ &= \underset{\text{(3 m)}}{\tfrac{1}{8}} + \underset{\text{(2 m, 1 f)}}{\tfrac{3}{8}} + \underset{\text{(1 m, 2 f)}}{\tfrac{3}{8}} + \underset{\text{(3 f)}}{\tfrac{1}{8}}\end{aligned}$$

The probability for the first combination, 3 males, is 1/8, or .125. There are 8 possible arrangements of results for samples of 3 cases from equiprobable, mutually exclusive events in a class (here, males and females), but there is only one way of obtaining the combination of 3 males. The probability for the second combination, 2 males and 1 female, is 3/8, or .375, since this combination can be obtained in any one of 3 ways; viz., m-m-f, f-m-m, or m-f-m.

The theoretical sampling distribution for an indefinitely prolonged series of random samples of persons, where $N_s = 3$, and males and females are equiprobable, is shown in B in Fig. 12:1. The mean of the distribution is

$$N_s p = 3(\tfrac{1}{2}) = 1.5$$

and its standard deviation is

$$\sqrt{N_s pq} = \sqrt{3(\tfrac{1}{2})(\tfrac{1}{2})} = \sqrt{.75} = .87$$

Binomials for Larger Samples

The expansion of a binomial for $n = 2$ and $n = 3$ is relatively simple because $(p + q)^2$ can be obtained by multiplying $(p + q)$ by $(p + q)$. Similarly, $(p + q)^3$ can be obtained by multiplying the expansion of $(p + q)^2$ by $(p + q)$, as follows:

$$\begin{array}{r} p^2 + 2pq + q^2 \\ \hline p + q \\ \hline p^3 + 2p^2q + pq^2 \\ \quad p^2q + 2pq^2 + q^3 \\ \hline p^3 + 3p^2q + 3pq^2 + q^3 \end{array}$$

This process can of course be repeated for $(p + q)^4$, for $(p + q)^5$, etc., but it is time-consuming and arduous for larger values of n. The computation can be simplified by the following general formula for the expansion of a binomial to any power of n:

$$(p + q)^n = p^n + \frac{n}{1}p^{(n-1)}q + \frac{n(n-1)}{1 \cdot 2}p^{(n-2)}q^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}p^{(n-3)}q^3 + \cdots + q^n \qquad [12:8]$$

Binomial for any power of n

Binomial for $N_s = 6$

If N_s, the sample size, is taken as 6, we have the following:

$$\left(\frac{1}{2}+\frac{1}{2}\right)^6 = p^6 + \frac{6}{1}p^5q + \frac{6}{1\cdot 2}p^4q^2 + \frac{6\cdot 5\cdot 4}{1\cdot 2\cdot 3}p^3q^3 + \frac{6\cdot 5\cdot 4\cdot 3}{1\cdot 2\cdot 3\cdot 4}p^2q^4 + \frac{6\cdot 5\cdot 4\cdot 3\cdot 2}{1\cdot 2\cdot 3\cdot 4\cdot 5}pq^5 + \frac{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}{1\cdot 2\cdot 3\cdot 4\cdot 5\cdot 6}p^0q^6$$

The last term simplifies to q^6 since the value for the coefficient is 1.0; p^0 also equals 1.0 because any number to the zero power is equal to 1.0.

Before the probability values of males and females are substituted in the above expansion of $p + q$ for $n = 6$, it should be noted that when p and q are equiprobable, their products for any term of the expansion will be the same as q^n or p^n. Thus, when the probability of $p = 1/2$, p^6 is 1/64; p^5q^1 is 1/64; p^4q^2 is 1/64, etc. Hence, when $p = 1/2$, it is sufficient to solve first for p^n and then determine the values of the coefficients of each term in the formula. When $n = 6$, the coefficients are 1, 6, 15, 20, 15, 6, and 1, respectively. We thus have the following distribution of results:

$$\underset{}{\left(\tfrac{1}{2}+\tfrac{1}{2}\right)^6} = \underset{(6\text{ m})}{\tfrac{1}{64}} + \underset{(5\text{ m},\,1\text{ f})}{\tfrac{6}{64}} + \underset{(4\text{ m},\,2\text{ f})}{\tfrac{15}{64}} + \underset{(3\text{ m},\,3\text{ f})}{\tfrac{20}{64}} + \underset{(2\text{ m},\,4\text{ f})}{\tfrac{15}{64}} + \underset{(1\text{ m},\,5\text{ f})}{\tfrac{6}{64}} + \underset{(6\text{ f})}{\tfrac{1}{64}}$$

The first and last combination can occur in only one way, whereas the combination of 5 males and 1 female can occur in six different ways; the 4 males and 2 females, in 15 different ways; the 3 males and 3 females in 20 different ways, etc. All these different combinations can be expected to occur purely on the basis of *chance* in sampling.

The theoretical sampling distribution for $N_s = 6$ is shown in *C* in Fig. 12:1. The mean is 3.0 and the standard deviation is 1.22. The probability ratio for extreme combinations has decreased considerably. Thus, when $N_s = 6$, we would expect to obtain all males in only 1 of 64 samples. Hence, such an extreme result with only 1 or 2 samples is *unlikely*, i.e., it would be most unusual in the light of experience.

Binomial for $N_s = 12$

Let us see what the situation is for samples doubled in size, i.e., $N_s = 12$. The expansion of $(p + q)^{12}$, where $p = 1/2$, is left as an exercise for the student. However, 2^{12} is 4096; hence $(1/2)^{12}$ is 1/4096. The coefficients for each of the 13 possible combinations of results, and hence their relative frequency, are as follows: * 12 m = 1; 11 m and 1 f = 12; 10 m and 2 f = 66; 9 m and 3 f = 220; 8 m and 4 f = 495; 7 m and 5 f = 792; 6 m and 6 f = 924; 5 m and 7 f = 792; 4 m and 8 f = 495; 3 m and 9 f = 220; 2 m and 10 f = 66; 1 m and 11 f = 12; and 12 f = 1.

* Cf. M. Philip, *The Principles of Financial and Statistical Mathematics*, Prentice-Hall, New York, 1941, p. 222, for the development of the factorial formula, ${}_nC_r$, for determining the value of any coefficients of a binomial.

The theoretical sampling distribution for an indefinitely prolonged series of random samples in which the size of each sample is 12 cases is shown in *D* in Fig. 12:1. The mean frequency of males is 6.0 and the standard deviation is 1.73. Although *D* is still rather peaked, a resemblance between it and the normal bell-shaped distribution is suggested. Furthermore, the probabilities of extreme combinations of results are so small as to be *exceedingly unlikely* when only a few samples are drawn.* The probability of all males in random samples of 12 cases each is only 1/4096, a proportion whose *P* value is .0002+. In other words, in a prolonged series of random samples, we would expect such an extreme result in less than 3 out of every 10,000 samples. By any standards of what experience shows to be *likely* or *unlikely*, as applied to probable inference in sampling statistics, such a result would be *most unlikely* for only one or even several random samples. The combination of 11 males and 1 female, with a *P* value of 12/4096, or .0029+, would be expected in the long run in slightly less than 3 samples per 1000. This result would also be *most unlikely* for only one or a few samples.

It should be noted that the possibility of even the most extreme result appearing, on the basis of chance, in a single random sample is not ruled out. However, it should be emphasized that such extreme results are *unlikely*. The concept of what is *likely* or *unlikely* in sampling and measurement is thus integral to evaluating the result of a single sample, or of only a few samples. In studying universes by means of an analysis of sample data, a distinction must be made between what is *likely* on the basis of chance and what is *unlikely* on the basis of chance. Distribution D in Fig. 12:1 describes *what can happen* on the basis of chance in drawing an indefinitely prolonged series of random samples of persons, 12 at a time, from a large universe in which the proportions of men and women are assumed to be equal. If the samples are drawn randomly, the laws of chance should operate normally and give the variations in results shown in the figure. But if we drew only a single random sample of 12 persons from a universe whose male and female composition is *unknown*, we would not consider a sample yielding 12 males or 11 males and 1 female a likely result for a universe *assumed to be* evenly divided with respect to the two sexes. In fact, either we would reject the hypothesis that the sample was randomly drawn from such a universe and suspect the presence of bias in the sample itself, or if the sample were obtained by a truly random technique we would conclude that the particular universe sampled contained a greater proportion of males than of females.

The Expansion of the Binomial for the Normal Probability Curve

It may be evident from the distributions in Fig. 12:1 that if the samples are large enough, the expansion of $(p + q)^n$, where $p = 1/2$, will yield sam-

* The student can test this for himself by tossing 12 coins simultaneously until he obtains all heads or all tails on a toss.

pling distributions that should increasingly approach the continuity or smoothness characteristic of the normal probability curve. That this is the case is mathematically demonstrable.* The expansion of $(p + q)^n$, when n is very large, will yield a distribution that, fortunately, can be obtained more readily by the equation of the normal probability function:

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

where the total area of the distribution is taken as equal to unity.

The normal probability curve is a smooth, continuous distribution rather than the succession of discrete classes characteristic of the point binomial. For the binomial, however, when N_s, the size of the sample, is very large, the relative differences in the frequencies of results for each combination become very small. If N_s is taken without limit, these differences become infinitely small and hence the expansion is identical with the normal probability distribution. It should be observed, however, that samples do not have to contain many more than 25 or 30 cases to give an expanded binomial that for all practical purposes can be treated *as if* it were a perfectly smooth, continuous normal probability distribution.

The Probability of a Result Derived from the Normal Probability Distribution

How are the implications of the normal probability curve to be utilized in determining the probability of a given sample result? We have seen that by means of the binomial expansion we can calculate the probability of any combination of males and females, equally divided in a large universe, for small random samples. But what if we draw random samples of 100 cases each? What will be the probability of obtaining *at least 60 males* in random samples of a universe in which the two sexes are equally divided? The formulation of the probability estimate here is somewhat different from that for the combinations of males and females in the point binomial. When utilizing the implications of the normal probability distribution in estimating a probability value, we estimate a *range* of possibilities rather than a given value. In other words, because a normal distribution is continuous rather than discrete, we estimate the probability of obtaining *at least 60 males*, or of *60 or more males*, rather than the probability of obtaining the precise combination of 60 males and 40 females.

To make probability estimates on the basis of the normal probability curve, we need to have (1) a measure of the variability of the sampling distribution under consideration and (2) a differentiation of the area (or frequencies) of the normal probability distribution in terms of its measure of variability.

* Cf. C. C. Peters and W. R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases*. McGraw-Hill, New York, 1940, pp. 279–286.

The standard measure of variability is the standard deviation and Table I, Appendix B, gives a differentiation of the normal probability distribution in terms of x/σ, or z_x.

The standard deviation of the sampling distribution of any parameter is called its *standard error*. The standard error for the present problem is $\sigma_f = \sqrt{N_s pq}$, where N_s is the size of the sample, p the proportion of males (or p events) in the universe under consideration, and $q = 1 - p$. In this case p is thus the P value for males, viz., 1/2 or .50. Where N_s, the size of the sample, is equal to 100,

$$\sigma_f = \sqrt{100(.50)(.50)} = 5.0$$

A result of 60 or more males in a random sample of 100 cases is greater than the mean frequency of males in the universe. It will be recalled that the mean frequency for random samples of a universe is equal to $M_f = N_s p$, where N_s is the size of the sample and p is the parameter mean frequency for the universe under consideration. In this case,

$$M_f = 100(.50) = 50$$

By hypothesis, then, this value is taken as the parameter mean frequency for a sampling distribution when $N_s = 100$ and $p = .50$. The parameter mean frequency is located at the modal point of the normal sampling distribution, as indicated in Fig. 12:2. The mean of the hypothetical, normal sampling distribution therefore is taken to coincide in value with the parameter mean frequency of males in the given universe.

A Test of Significance (*T*)

A sample result of 60 or more males would lie in the area of the tail of the normal distribution that is 2.0 standard deviation units above 50, when $\sigma_f = 5.0$ and M_f is 50. This is so, because $(60 - 50)/5.0 = 2.0$. This relationship, it will be observed, is similar to that developed earlier for x/σ, or z scores. Thus,

$$z = \frac{X - M_x}{\sigma_x} \qquad [8:1]$$

z score

This z score formula is for converting the original scores of a distribution to positions on a scale in standard deviation units, x/σ. We are now concerned, however, not with the difference between a particular original score, X, and the mean of an obtained distribution, M_x, but with the difference between the statistic of a sample result (in this case, f_s, where f_s is the frequency of males obtained in the sample) and the parameter frequency of a hypothetical universe (in this case, M_f, the parameter mean frequency of males). The difference between the sample frequency of males and the parameter mean frequency is

$$f_s - M_f = 60 - 50 = 10$$

Since the parameter value of a measure is located at the mean of its sampling distribution, we shall employ the symbol f_h instead of M_f, the subscript h standing for *hypothesis*. The symbol f_h therefore complements the symbol f_s. The latter indicates the frequency of a class of events in the sample result (in this case, males), and f_h indicates the parameter frequency of that class of events for the universe under consideration (in this case, a proportion of males equal to .50, and therefore a parameter frequency of 50 when $N_s = 100$). The difference between the sample frequency and the frequency *by hypothesis* will therefore be symbolized as $f_s - f_h$.

In Formula 8:1, for z scores, the denominator represents the standard deviation of the obtained distribution of scores. In the present situation, however, σ_f represents the standard deviation of the *sampling distribution* when $f_h = 50$ and $N_s = 100$. This standard deviation is called the standard error of the statistic, i.e., the standard error of the *frequency* of males in random samples for the universe of the hypothesis. It measures the variability of the distribution of the results of random samples, and hence serves as the yardstick for measuring the variation in sample results to be expected upon the basis of *chance errors* in sampling and measurement.

We shall symbolize this new relationship by T, where T stands for the *test ratio* of a Test of Significance when N_s is large, and the normal probability distribution of large sample theory describes the form of the sampling distribution.

$$T = \frac{\text{(sample measure)} - \text{(parameter measure)}}{\text{standard error of the measure}} \qquad [12:9]$$

General form of a Test of Significance

In this case, where the measure under consideration is the frequency of a class of events,

$$T = \frac{f_s - f_h}{\sigma_f} \qquad [12:10]$$

Test of Significance for frequencies

and therefore, when $N_s = 100, f_s = 60, f_h = 50$, and $\sigma_f = 5.0$:

$$T = \frac{60 - 50}{5.0} = 2.0$$

If we now consult Table I, Appendix B, for the differentiation of the normal probability curve in terms of x/σ, we see that when $x/\sigma = 2.0$, 47.72% of the total area lies between the mean and a point two standard deviation units above it. Hence the proportion of the area above $x/\sigma = 2.0$ is 50% − 47.72% = 2.28%. This is shown in Fig. 12:2.

This value of 2.28, expressed as a proportion equal to .0228, is the P value that we set out to obtain. It is the probability value for the given T ratio of 2.0. It gives for a normal sampling distribution the *relative frequency* with which random samples drawn 100 at a time will yield results of *60 or more males*, when the universe sampled is *by hypothesis*, or by knowledge, divided

equally between males and females. On the basis of chance, which operates in all sampling, but which operates in a lawful way in random sampling, we would expect to obtain 60 or more males in approximately 23 random samples per thousand samples (or 2.3 per hundred), when $N_s = 100$.

To obviate computing a probability value for any value of T by the process of subtraction just used, we have set up a special table (Table II, Appendix B) which gives these values for T from .00 to 3.0. Thus the P value for a T ratio of 2.0 can be read directly from the table as equal to .0228.

The Evaluation of the Test of Significance

Would a P value of .023 for a single sample result be considered as indicative of a result that is *likely* to occur on the basis of chance alone? A categorical answer of yes or no cannot be given to this question when $T = 2.0$ and hence $P = .023$. Generally, however, in psychological and social science statistics a result with a P value of .023 is judged to be either (1) likely on the basis of chance, or (2) doubtful because experience has indicated that such a P value is not definitely indicative of results that "just don't happen" in the ordinary course of events. If the result is considered *doubtful*, we cannot be confident that it is either likely or unlikely on the basis of chance; therefore, additional sampling evidence will be needed. If the result is considered *likely*, this means that 60 or more males in a random sample of 100 persons is judged to be a reasonable expectancy for a universe equally divided between males and females. Such a conclusion would in effect say that the *difference* between the sample result of 60 males and the parameter mean frequency of 50 males is *not a significant difference;* rather, it is a difference likely on the basis of chance errors of sampling.

Let us consider some of the implications of the preceding with respect to polling the preferences of a large universe of voters for two political candidates. The parameter frequency of the voters' preferences for either Candidate A or Candidate B is of course unknown (if the parameter frequency were known, no poll would be necessary). If A receives 60 of the preferences in a random sample of 100 voters of the universe polled, we should either (1) doubt that the preferences of all the voters in the universe would give a majority for A, or (2) conclude that the sample result is too likely to indicate an even division of voters' preferences in the universe to warrant the inference that A will win. This would be the case because an even division of voters' preferences requires a Test of Significance in which the parameter frequency of preferences for A is taken as equal to 50%. But if a sample frequency of 60% is a likely result for the hypothesis that the preference for A is equal to 50%, we cannot be confident that in the election A will get more than 50% of the votes.

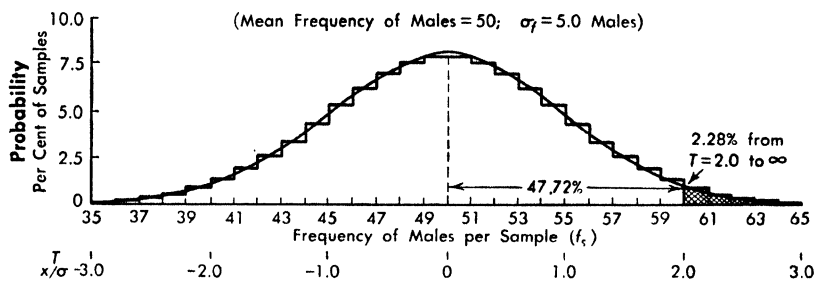
If, however, there were three candidates in the race, a plurality were sufficient to elect, and the poll results for B and C were evenly divided, 20 and 20,

the conclusion that A's election is likely would be warranted, since A would not need as much as 50% of the total vote in order to win, and a sample result of 60% for A would most likely indicate a plurality over either B or C.

The Distribution of Frequencies in the Normal Probability Distribution

In order to show the *form* of the normal probability distribution when drawn on a scale comparable to that used for the small-sample sampling distributions of the expanded binomials in Fig. 12:1, the total area of the sampling distribution in Fig. 12:2 has been made approximately the same.

Fig. 12:2. The Theoretical Sampling Distribution of the Binomial $(p + q)^{100}$, for $N_s = 100$, When p and q Are Equiprobable and the Expansion of the Binomial Is Based on the Normal Probability Function of Large Sample Theory, with the Total Area of the Distribution Taken as 100% and Scaled the Same as the Distributions in Fig. 12:1.



The distribution of the data in this figure is given in Table 12:2, and is based on the normal probability function (Formula 8:2). The proportion of the area of the normal probability distribution for any distance on the abscissa above or below the mean is given in terms of x/σ in Table I, Appendix B.

The sampling distribution in Fig. 12:2 is for random samples drawn 100 at a time; thus, $N_s = 100$. If the distribution were set up from the binomial, instead of on the basis of the normal probability function, the expansion would be equal to $(p + q)^{100}$, with p (males) and q (females) equiprobable in the universe sampled. Under these circumstances, as we have already indicated, the parameter mean frequency of males is 50, and the standard error of the sampling distribution, σ_f, is 5.0. Hence, between the mean of 50 males and a point one standard deviation above it, there will be an increase of 5 males per sample. In other words, at $T = 1.0$, $f_s = 55$ males; at $T = 2.0$, $f_s = 60$ males; at $T = -1.0$, $f_s = 45$ males, etc. An increase or decrease of 1 male in a sample is therefore equal to a change of 0.2 standard deviation unit on the abscissa scale. Thus, at $T = 0.2$, $f_s = 51$; at $T = 0.4$, $f_s = 52$, etc. Consequently, in the distribution in Fig. 12:2, the successive increases

Table 12:2. Theoretical Percentage Distributions and *P* Values of Sample Results for Binomial Based on the Normal Probability Function, When Males (*p*) and Females (*q*) Are Equiprobable and the Size of Random Samples, N_s, Is 100

Frequency of Males per Sample	Class Interval in Term Limits of σ	% Frequency of Sample Results	Probability Values *P*
65 to 100	3.0 to ∞	(0.135)	(.00135)
64	2.8 to 3.0^-	0.125	.00125
63	2.6 to 2.8^-	0.21	.0021
62	2.4 to 2.6^-	0.35	.0035
61	2.2 to 2.4^-	0.57	.0057
60	2.0 to 2.2^-	0.89	.0089
59	1.8 to 2.0^-	1.31	.0131
58	1.6 to 1.8^-	1.89	.0189
57	1.4 to 1.6^-	2.60	.0260
56	1.2 to 1.4^-	3.43	.0343
55	1.0 to 1.2^-	4.36	.0436
54	0.8 to 1.0^-	5.32	.0532
53	0.6 to 0.8^-	6.23	.0623
52	0.4 to 0.6^-	7.04	.0704
51	0.2 to 0.4^-	7.61	.0761
50	0.0 to 0.2^-	7.93	.0793
49	-0.2 to 0.0^-	7.93	.0793
48	-0.4 to -0.2^-	7.61	.0761
47	-0.6 to -0.4^-	7.04	.0704
46	-0.8 to -0.6^-	6.23	.0623
45	-1.0 to -0.8^-	5.32	.0532
44	-1.2 to -1.0^-	4.36	.0436
43	-1.4 to -1.2^-	3.43	.0343
42	-1.6 to -1.4^-	2.60	.0260
41	-1.8 to -1.6^-	1.89	.0189
40	-2.0 to -1.8^-	1.31	.0131
39	-2.2 to -2.0^-	0.89	.0089
38	-2.4 to -2.2^-	0.57	.0057
37	-2.6 to -2.4^-	0.35	.0035
36	-2.8 to -2.6^-	0.21	.0021
35	-3.0 to -2.8^-	0.125	.00125
34 to zero	-3.0 to $-\infty$	(0.135)	(.00135)
	Total	100%	1.00

in frequencies of males per sample are scaled into class intervals equal to 0.2σ.

The percentage of the total area of the normal probability distribution between the mean and any distance above or below it, in terms of x/σ (or T), can readily be obtained by referring to Table I, Appendix B. Thus, when $x/\sigma = 0.2$, 7.93% of the total area is found to lie between the mean and 0.2σ. Since the distribution is bilaterally symmetrical, the same percentage of the area lies between M and -0.2σ. The percentages of the total area for successive intervals, taken 0.2σ at a time and given in Table 12:2, are not differ-

entiated beyond $\pm 3.0\sigma$. Sample frequencies of less than 35 or more than 65 males are not indicated in Fig. 12:2. The reason for stopping at ± 3.0 is obvious: the percentage of possible sample results beyond these points is so small that they cannot be differentiated on the graph. The likelihood of such extreme results from random samples of the universe under consideration is remote. Of the total area of the normal probability curve, 99.73% lies between $M \pm 3.0\sigma$. Results beyond $M \pm 3.0\sigma$, for an indefinitely prolonged series of random samples, will occur less than 3 times in 1000, since 100.% − 99.73% = 0.27%. Half of this remainder, or 0.135%, is the percentage of the area beyond 3.0σ, and consequently $P = .00135$ is the probability value for 65 or more males in random samples when $N_s = 100$. Similarly $P = .00135$ is the probability value for less than 34 males in such samples.

From our earlier discussion of what is *likely* or *unlikely* on the basis of chance in one or only a few random samples of a defined universe, it should be evident that in the above situation samples consisting of 2/3 or more males, or of 1/3 or less males, are extremely unlikely. If such an extreme result as 70 males were obtained in a *random* sample of a universe whose parameter mean frequency was unknown but was taken at 50 *by hypothesis* for a Test of Significance, we would with confidence reject the hypothesis and conclude that the sample of 70 was drawn from a universe whose parameter mean frequency was greater than 50.

C. SMALL SAMPLE THEORY—LEPTOKURTIC SAMPLING DISTRIBUTIONS

Not all sampling distributions take the form of the bell-shaped normal probability function. That this is the case for small samples was shown in Fig. 12:1. Although these sampling distributions are uni-modal and bilaterally symmetrical, their form obviously would not correspond to that of the normal probability curve even if they were transformed from histograms into smooth continuous curves running through ordinate points at the mid-points of each interval. The normal probability function, differentiated for x/σ in Table 1, Appendix B, is thus not satisfactory for describing how the areas of small-sample sampling distributions are distributed. Hence, probability estimates for results obtained from random small samples are based on different tables of probability values from those used with large sample results whose statistics are normally distributed.

It should be emphasized that the bell-shaped probability curve, based on Formula 12:5, is not the only "normal" sampling distribution. On the contrary, there is a sampling distribution "normal" for every given size of sample and given kind of statistic. Thus, the sampling distributions in Fig. 12:1 are the "normal" probability distributions for the point binomial when $N_s = 2$, 3, 6, and 12. Unless this fact is recognized, the concept "normal probability distribution" is likely to be ambiguous. What is normal for one

situation may be non-normal for another. Generally, however, the *normal probability distribution* is employed to refer to the standard bell-shaped frequency curve of large sample theory, based on Formula 8:2. When a sampling distribution takes a different form, the difference should be reported.

Kurtosis (*Ku*)

As samples decrease in size, their respective sampling distributions are much more peaked (Fig. 12:1) than the normal probability distribution of large sample theory (Fig. 12:2). The technical term in statistics for differences in the distribution of the area (or frequencies) about the mean of a uni-modal, bilaterally symmetrical curve is *kurtosis* (from the Greek, meaning "overarching"). The normal probability curve of large sample theory is *mesokurtic* (*meso* from the Greek *mesos*, meaning "middle"), whereas the peaked distributions in Figs. 12:1–12:4 are leptokurtic (*lepto* from the Greek *leptos*, meaning "slender"). Occasionally distributions are considerably flattened throughout the middle; these are described as platykurtic (*platy* from the Greek *platos*, meaning "flat"). Tests of Significance for analyzing the kurtosis of a distribution are developed in Chapter 13, Section D. Such tests make it possible to judge whether the divergence of a given distribution from mesokurtosis is significant.

The sampling distributions of most statistics in large sample theory can be assumed to be similar in form to the normal probability curve (Fig. 12:2), provided the universe sampled is itself large. Probability estimates can therefore be made for such statistics from a differentiation of the one curve, as in Table II, Appendix B. In small sample theory, however, the form of the sampling distributions changes somewhat as N_s is increased or decreased by only one case. Hence a differentiation of the area of one curve for small samples of differing sizes is not adequate for probability estimates, and a different table of probability values is therefore required for different values of N_s.

The *t* Statistic

A satisfactory treatment of the varying forms of the distributions of small samples was first presented by William S. Gosset, an English scholar, in an article published under the pseudonym "Student" in 1908.* A table of probability values for small samples developed by R. A. Fisher from Student's distribution is presented in Table III, Appendix B. As will be indicated later, this table is employed in Tests of Significance for random samples when N_s is less than 25 or 30. When such tests are made for samples as small as these, t instead of T is used to symbolize the difference in the sampling situation.

This distinction between t and T is the basis for understanding the implica-

* Student, "The Probable Error of a Mean," *Biometrika*, 6:1–25, 1908.

tions of Fisher's *t statistic* of small sample theory. The symbol concept *t* signifies the test ratio of a Test of Significance of a statistic derived from a small sample of observations or measurements; in large sample theory, this ratio is represented by *T*.

When Is a Sample Small?

There is no sharp line of division between "small samples" of small sample theory and "large samples" of large sample theory. However, inspection of Table III, Appendix B, reveals that the probability values of *t* for sampling distributions based on 25 to 30 cases are practically identical with those based on infinitely large samples. Consequently a value of from 25 to 30 for N_s is generally taken as the basis for distinguishing between the samples of small sample theory and those of large sample theory. That this does not harmonize with the use of such terms as *small* and *large* samples in other connections may be apparent. A sample of 100 cases, for example, would be a relatively *small* sample of all the voters in the United States; a moderately *large* sample of such a universe would consist of several thousand cases. The situation here is of course different from that in small sample theory and large sample theory. Small sample theory indicates that the form of the sampling distribution is increasingly leptokurtic when the size of the sample is taken as less than 25 or 30 cases. If a sample of 100 cases is relatively *small* for a particular investigation, the implications of large sample theory, rather than small sample theory, are nevertheless used in evaluating the result.

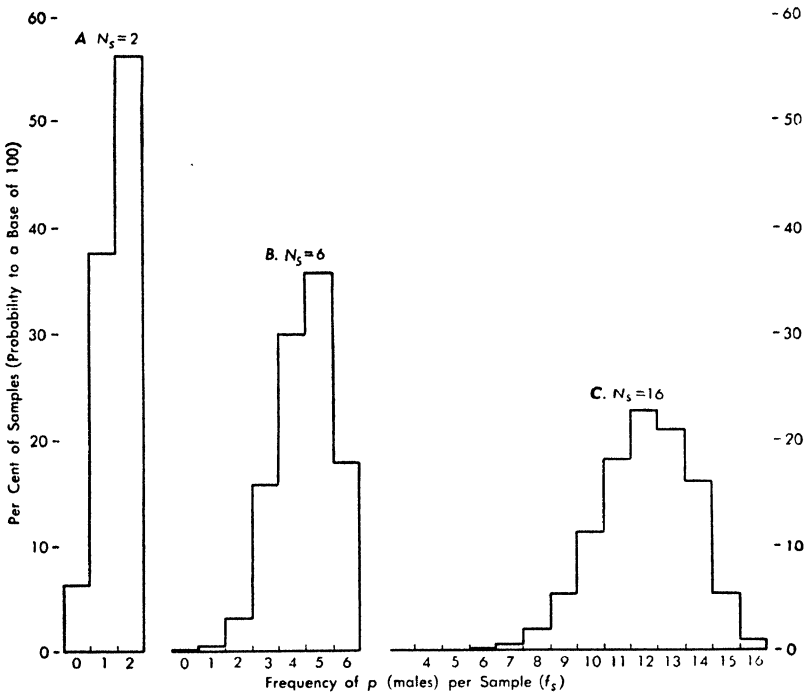
D. SKEWED SAMPLING DISTRIBUTIONS AND NORMAL PROBABILITY

The standard, normal probability distribution in Fig. 12:2 is more general than the binomial distribution. This is the case because an asymmetrical binomial distribution, and some other types of non-normal distributions, approach the symmetrical, normal probability curve when large random samples are drawn from a large universe. On the other hand, the sampling distributions of some statistics are *skewed*, regardless of whether the samples are small or large, or drawn randomly from small or large universes. This is true, for example, of the sampling distributions of correlation coefficients as their parameter values approach 1.00 or −1.00. High[11] values of *r* yield sampling distributions that are increasingly skewed and leptokurtic. However, as the parameter values of *r* approach *zero*, the sampling distributions of correlation coefficients increasingly approach the standard, normal probability form. The sampling distribution for r_h = zero is normal. This latter point is of considerable importance because, as we shall see later, one of the primary Tests of Significance for correlation coefficients is for the hypothesis that r_h is zero (the subscript *h* symbolizing the parameter value of the hypothesis tested).

The Binomial When $p \neq q$

The basis in sampling for skewed distributions can readily be seen from the expansion of the binomial when p and q are not equiprobable, i.e., $p \neq q$.

Fig. 12:3. The Theoretical Sampling Distributions of the Binomial $(p + q)^n$, When $p \neq q$ (with $p = \frac{3}{4}$ and $q = \frac{1}{4}$), with the Size of the Samples, N_s, Equal to 2, 6, and 16, and with the Total Areas of Each Distribution the Same (100%)



Furthermore, the greater the difference in the values of p and q, the greater the amount of skewness for sampling distributions based on a given size of sample. It may appear paradoxical that if samples of large universes are sufficiently large, the skewness of sampling distributions when $p \neq q$ becomes negligible and the results may for all practical purposes be treated as if the distribution were of the standard, normal form. However, this can be demonstrated mathematically; * it is graphically indicated by the distributions in Fig. 12:3.

* Cf. J. G. Smith and A. J. Duncan, *Sampling Statistics and Applications*, McGraw-Hill, New York, 1945, especially chap. 4.

If, instead of drawing random samples from a large universe equally divided between males and females, we draw them from a universe in which males outnumber females by 3 to 1, the probability of males (p) will be 3/4, and of females (q), 1/4. Let us see what happens in the expansion of the binomial for small samples when $N_s = 2$, $N_s = 6$, and $N_s = 16$. The theoretical sampling distribution will be as follows, when $N_s = 2$:

$$(p+q)^n = (\tfrac{3}{4} + \tfrac{1}{4})^2 = \tfrac{9}{16} + \tfrac{6}{16} + \tfrac{1}{16}$$

$$P = \underset{(2\text{ m})}{.5625} + \underset{(1\text{ m},\,1\text{ f})}{.3750} + \underset{(2\text{ f})}{.0625} = 1.0$$

The respective probability values of these three results range from $P = .56$ to .06, and the theoretical sampling distribution ("normal" for this situation in sampling) is graphed in A in Fig. 12:3. It is obvious that the distribution is asymmetrical, with the skewed portions (or extended tail) in the direction of fewer males in the sample results. As a matter of fact, the distribution has no *central* tendency, and its *mode* is at one end of the distribution. The parameter mean frequency is equal to $N_s p$, as for the binomial when $p = 1/2$; however, p is now equal to 3/4. Hence $M_f = 2(3/4) = 1.5$ males. This is obviously not the value of the modal interval (or point). The mode (Mo) for a binomial is equal to the following:

[12:11]

Mo = the integer value between $N_s p - q$ and $N_s p + p$ — Mode of a binomial distribution

Thus, in the preceding example,

$$N_s p - q = 2(\tfrac{3}{4}) - \tfrac{1}{4} = 1.25$$

and

$$N_s p + q = 2(\tfrac{3}{4}) + \tfrac{3}{4} = 2.25$$

hence $Mo = 2.0$, since this is the integer value between 1.25 and 2.25.

When $N_s = 6$, the expansion of $(p + q)$, for $p = 3/4$, is as follows:

$$\left(\frac{3}{4}+\frac{1}{4}\right)^6 = \left(\frac{3}{4}\right)^6 + \frac{6}{1}\left(\frac{3}{4}\right)^5\left(\frac{1}{4}\right) + \frac{6\cdot 5}{1\cdot 2}\left(\frac{3}{4}\right)^4\left(\frac{1}{4}\right)^2 + \frac{6\cdot 5\cdot 4}{1\cdot 2\cdot 3}\left(\frac{3}{4}\right)^3\left(\frac{1}{4}\right)^3$$
$$+ \frac{6\cdot 5\cdot 4\cdot 3}{1\cdot 2\cdot 3\cdot 4}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^4 + \frac{6\cdot 5\cdot 4\cdot 3\cdot 2}{1\cdot 2\cdot 3\cdot 4\cdot 5}\left(\frac{3}{4}\right)\left(\frac{1}{4}\right)^5 + \left(\frac{1}{4}\right)^6$$
$$= \frac{729}{4096} + 6\left(\frac{243}{1024}\right)\left(\frac{1}{4}\right) + 15\left(\frac{81}{256}\right)\left(\frac{1}{16}\right) + 20\left(\frac{27}{64}\right)\left(\frac{1}{64}\right) + 15\left(\frac{9}{16}\right)\left(\frac{1}{256}\right)$$
$$+ 6\left(\frac{3}{4}\right)\left(\frac{1}{1024}\right) + \frac{1}{4096}$$

$$P = \underset{(6\text{ m})}{.1780} + \underset{(5\text{ m},\,1\text{ f})}{.3560} + \underset{(4\text{ m},\,2\text{ f})}{.2966} + \underset{(3\text{ m},\,3\text{ f})}{.1318} + \underset{(2\text{ m},\,4\text{ f})}{.0330} + \underset{(1\text{ m},\,5\text{ f})}{.0044} + \underset{(6\text{ f})}{.0002} = 1.0$$

The probability of each combination ranging from all males to no males per sample varies from $P = .18$ to $P = .0002$. Graph B in Fig. 12:3 shows the form of the sampling distribution. The distribution is still skewed, but not

so much so as in A, when $N_s = 2$. The parameter mean frequency of males is

$$N_s p = 6(\tfrac{3}{4}) = 4.5$$

and the parameter modal frequency of males is equal to the following, obtained by means of Formula 12:11:

$$N_s p - q = 6(\tfrac{3}{4}) - \tfrac{1}{4} = 4.25$$
$$N_s p + p = 6(\tfrac{3}{4}) + \tfrac{3}{4} = 5.25$$

Hence, $$Mo = 5.0$$

Thus, the modal frequency of males has shifted one interval toward the center of the distribution, and although the difference between the mean and mode is still 0.5, relative to the entire dispersion or spread of the respective sampling distributions for $N_s = 2$ and $N_s = 6$, the difference is not so great as in the smaller sample.

Graph C in Fig. 12:3 shows the form of the theoretical sampling distribution when $N_s = 16$ and $p = 3/4$. The mean is

$$N_s p = 16(\tfrac{3}{4}) = 12.0$$

and the parameter modal frequency is:

$$N_s p - q = 16(\tfrac{3}{4}) - \tfrac{1}{4} = 11.75$$
$$N_s p + p = 16(\tfrac{3}{4}) + \tfrac{3}{4} = 12.75$$

Hence, $$Mo = 12.0$$

In the *point binomial* for p or $q = 3/4$, the mean and mode have the same integer value when N_s is a multiple of 4. But for a *continuous* distribution that is negatively skewed, the mean and mode are not identical in value but lie in the same integral interval, with the modal value of males slightly larger than the mean value.

When $N_s = 16$ and $p = 3/4$, the figure shows that the mean frequency of the sampling distribution is shifted four intervals from the extreme of *16 males*. Furthermore, approximately 99% of the results of random samples will be expected to lie within the limits of the mean interval (12 m) and four intervals above and four below, i.e., between 8 males and 16 males. Although the sampling distribution is negatively skewed, less than 1 sample in 100 may be expected to contain less than 8 males. The skewness is thus considerably reduced when N_s is as large as 16. The P values for p events (or males) are as follows:

16 m, P = .0100	10 m, P = .1110	4 m, P = .00003
15 m, P = .0535	9 m, P = .0524	3 m, P = .000004
14 m, P = .1336	8 m, P = .0197	2 m, P = .0000003
13 m, P = .2079	7 m, P = .0058	1 m, P = .00000001
12 m, P = .2252	6 m, P = .0014	0 m, P = .0000000002
11 m, P = .1802	5 m, P = .0003	

The P values for the last six intervals at the lower end of the distribution are too small to be shown on Graph C.

As in the case of Fig. 12:1, where $p = 1/2$, the sampling distributions in Fig. 12:3 are leptokurtic, but they become less so as N_s is increased.

If $N_s = 100$, the standard normal probability distribution is adequate for describing the form of the sampling distribution for $(p + q)^{100}$, when $p = 3/4$ and $q = 1/4$. In other words, the skewness characteristic of such a distribution is negligible, and the form of the main part of the distribution (within the limits of $M \pm 3.0\sigma$) is almost bilaterally symmetrical and mesokurtic. Thus, the normal distribution in Fig. 12:2 may for all practical purposes be taken as the form of the sampling distribution when N_s is large and $p \neq q$. It should be emphasized, however, that this treatment for sampling distributions of $p \neq q$ is warranted on the condition that N_s is large and also that p is not much greater than .95 or much less than .05. In other words, if the difference between p and q is in excess of $.95 - .05 = .90$ the samples must be unusually large if the skewness of the sampling distribution is to be negligible.

E. THE PRECISION (RELIABILITY) OF SAMPLE RESULTS AND THE SIZE OF SAMPLES

The *adequacy* of a sample result, as we saw in the preceding chapter, is dependent upon both (1) its *character* and (2) its *precision*. The character of a sample result is of prime importance. Unless we know the nature of a sample, we cannot with confidence study a given universe in the light of the sample result. From the point of view of applying the logic of probability and statistical inference to a sample result, we must draw *random* or *stratified-random* samples of the universe to be studied. Samples of this character are *adequate* provided they are sufficiently large to yield the precision needed for the particular investigation.

Precision Measured by the Standard Error

The precision of any statistic derived from a sample result is measured in terms of the standard error of the statistic, i.e., the standard deviation of the sampling distribution of the statistic. In the case of *frequencies*, we saw that the standard error of a frequency, σ_f, is equal to $\sqrt{N_s pq}$, where N_s is the size of the sample, p is the *proportion* of the events of a class in the universe under consideration, and q is equal to $1.0 - p$. As the size of random samples increases, the precision of the result also increases. Figs. 12:1–12:3 suggest, however, that the variability of the sampling distributions increases as N_s increases. That is, the abscissa scales of these various distributions are wider for the larger size of samples. If the variability of the sampling distributions of a statistic actually increases as the size of the sample increases, it does not follow that the precision of a statistic increases as N_s increases, because the precision is measured directly in terms of the variability of the sampling distribution, viz., its standard error.

Precision Generally a Function of $\sqrt{N_s}$

The contradiction suggested by Figs. 12:1–12:3 is only *apparent;* it is not real. Actually the variability of these sampling distributions will be seen to *decrease* when their respective variabilities are considered *relative* to their respective scales of measures (i.e., frequencies of males per sample result). Thus the standard error of the sampling distribution for $N_s = 2$ (Graph *A*, Fig. 12:1) is .707 males; for $N_s = 12$ (Graph *D*, in that figure), it is 1.41 males. However, the frequency of males per sample result, when $N_s = 2$, can vary only from zero to 2.0, whereas it can vary from zero to 12.0 when $N_s = 12$. Therefore, relative to the size of the samples, and hence to the different possible results per sample for these sampling distributions, a σ_f of .707 males, when $N_s = 2$, indicates a greater margin of possible error than a σ_f of 1.4 males, when $N_s = 12$. This relationship can be seen more readily if the variability of sampling distributions of frequencies is measured in terms of the size of the sample, N_s. For this, the ratio of σ_f to N_s, where $\sigma_f = \sqrt{N_s pq}$, is as follows:

$$\sigma_{\frac{f}{N_s}} = \frac{1}{N_s}\sqrt{N_s pq} = \sqrt{\frac{N_s pq}{N_s^2}} = \sqrt{\frac{pq}{N_s}}$$

Since f/N_s is the proportion, p, of the class of events under consideration (in this case *males*), the standard error is symbolized by σ_p:

$$\sigma_p = \sqrt{pq/N_s} \qquad [12:12]$$

Standard error of a proportion

When $N_s = 2$, and the parameter proportion of p events (males) for the universe under consideration is .50, q will equal

$$1.0 - .50 = .50$$

and

$$\sigma_p = \sqrt{(.50)(.50)/2} = .354$$

If the size of the sample is doubled,

$$\sigma_p = \sqrt{(.50)(.50)/4} = .250$$

(Cf. Table 12:3.) Thus, the standard error of the parameter proportion, and hence the variability of the sampling distribution, is reduced. Stated otherwise, the *precision* of the result is increased. However, the precision is not doubled; i.e., the standard error is not reduced to half its size when the size of the sample is doubled. In order to reduce the value of $\sigma_p = .354$ by half, the size of the sample must be quadrupled. If $N_s = 2$ is quadrupled, N_s will equal 8 and

$$\sigma_p = \sqrt{(.50)(.50)/8} = .177$$

which is half the value of .354, the standard error for $N_s = 2$.

The precision of a sample result thus appears to be directly proportional to the square root of the size of the sample. Or, if this relationship is stated in

terms of *error* (the opposite of *precision*), the standard error of a result is inversely proportional to the square root of the size of the sample. Whether this relationship holds strictly depends upon the probability implications of the measure of error itself (in this case, the standard error of a proportion). If the standard error of a sampling distribution of proportions, taken with respect to the mean of the sampling distribution, marks off the same percentage of probabilities (or fraction of the total area), regardless of the size of N_s, then this relationship between the precision of a result and the size of the sample will hold generally. We have seen, however, that when N_s is taken as less than 25 or 30 cases, sampling distributions do not have the standard, normal probability form. As N_s approaches 2, they become increasingly leptokurtic; and when p is not equal to q, they are skewed. Thus the percentage of probabilities between the mean ± 1 standard error for small sample distributions is *less* than it would be for large sample theory. In the latter case, as indicated in Table I, Appendix B, $M \pm 1\sigma$ includes .6826 of the total area (or about 2/3 of the whole); and within the limits of $M \pm 3\sigma$, .99730 of the total area (or nearly 100%) is included.

For small sample theory these probability values are less. This is indicated in Table 12:3 for the sampling distributions of small sample theory, when developed as continuous rather than as discrete distributions of the binomials in Figs. 12:1 and 12:3. When $N_s = 2$, for example, the limits of $M \pm 1\sigma$ include an area equal to only 50% of the total distribution; and less than 80% of the total area is included within the limits of $M \pm 3\sigma$.

There is thus a lawful relationship between the precision of a statistic and the size of the sample from which it is derived. The function describing this relation is the same for most statistics of large sample theory: Precision is directly proportional to the square root of the size of the sample. As indicated in Table 12:3, in order to double the precision of a result when $N_s = 25$ and $\sigma_p = .100$, N_s must be quadrupled; when $N_s = 100$, $\sigma_p = .05$. Or, if the parameter proportion is .75 instead of .50, $\sigma_p = .087$ when $N_s = 25$, and is half this size, viz., .043, when $N_s = 100$. On the other hand, when N_s is much less than 25, this relationship does not hold precisely. As the table shows, the greatest differences in the probability implications of small samples are for small ones, where $N_s = 2, 3, 4$, etc. The Fisher-Student t statistic has been developed to provide a basis for probability estimates in the latter case. The research worker should be acquainted with its meaning and its usefulness in evaluating the results of small samples. Generally, however, the size of samples for research investigations in psychology and related fields is taken as at least 25 or 30 cases; consequently the sampling distributions of large sample theory are ordinarily those to be used.

Precision and Reliability

The concept of reliability has been more widely used in sampling and analytical statistics than that of precision. Actually, the two terms are synony-

Table 12:3. Precision and Size of Sample

The Variability of the Sampling Distributions of Proportions as a Function of (1) the Size of the Sample and (2) the Parameter Value of the Mean Proportion *

(1) Size of Sample N_s	(2) Standard Error of a Proportion When $p_h = .50$	(3) Standard Error of a Proportion When $p_h = .75$	(4) Probabilities Proportion of Sample Results to Be Expected Within Limits of $M +$ or $-1\sigma_p$	(5) Probabilities Proportion of Sample Results to Be Expected Within Limits of $M +$ or $-3\sigma_p$	(6) Probabilities Proportion of Sample Results to Be Expected Within Limits of $M +$ and $-3\sigma_p$
Small Sample Theory †					
2	.354		.2500	.3976	.7952
3	.289		.2887	.4523	.9046
4	.250		.3045	.4712	.9424
5	.224		.3130	.4800	.9600
6	.204		.3184	.4850	.9700
8	.177		.3247	.4900	.9800
10	.158		.3283	.4925	.9850
12	.144		.3306	.4940	.9880
16	.125		.3334	.4955	.9910
20	.112		.3351	.4963	.9926
Large Sample Theory					
25	.100	.087	.3413	.49865	.9973
50	.071	.061	.3413	.49865	.9973
75	.058	.050	.3413	.49865	.9973
100	.050	.043	.3413	.49865	.9973
400	.025	.0217	.3413	.49865	.9973
1600	.0125	.0108	.3413	.49865	.9973
6400	.00625	.0054	.3413	.49865	.9973
			.3413	.49865	.9973

* The smaller σ_p, the greater the precision (or reliability) of the result.

† These probability values for small samples are from Student's Table in Peters and Van Voorhis, *op. cit.*, pp. 488–491.

mous. Both refer to the degree and nature of the variability that is characteristic of the sampling distribution of a statistic. The less the variability, the more reliable, or precise, the result. The use of the term *reliability* goes back historically to the treatment of errors of observation and measurement in physics, psychophysics, etc. The less the error of observation or of measurement, the more reliable the result; and the greater the error, the more unreliable the result. As in sampling statistics, the effect of chance factors on observation and measurement is measured in terms of the standard error of the statistic (or in terms of its *Probable Error*)‡.

From Table 12:3 we can see not only the effect of the size of random samples on the precision or reliability of a result, but also what size of such a sample

‡ For the normal distributions of large sample theory, the probable error, often referred to as P.E., is equal to .6745, or about two-thirds, of the standard error, and sets the limits of 25% of the probabilities above or below the mean.

is required to obtain a given degree of precision for a universe whose parameter proportion is taken as either .50 or .75. Furthermore, the variability of the sampling distributions of parameter proportions $\neq .50$ is less than that for $P_h = .50$. Thus, the standard errors for $P_h = .50$ of column 2 in the table represent *maximum* values of σ_p for given sizes of samples. In other words, σ_p for $p_h \neq .50$ is less than σ_p for $p_h = .50$.

The precision, or reliability, of sample results for different sizes of samples is further illustrated by Fig. 12:4. The vertical line represents the parameter value of the measure (i.e., the value of the measure *by hypothesis*, or for the hypothesis to be tested). The horizontal scales represent only the abscissas of the sampling distributions (not the areas) for $N_s = 25$, 50, 75, 100, 400, 1600, and 6400, the standard errors of which are given in Table 12:3 for a particular measure. The sampling distribution of each is assumed to be normal as in large sample theory. The length of each scale is taken as equal to the mean plus and minus 3 times the standard error of the measure. The ranges of the parameter value of the measure $\pm 1\sigma$ and $\pm 2\sigma$ are also differentiated.

Fig. 12:4 thus serves to show the *relative precision* of the sample results for any statistic whose sampling distribution has, or can be assumed to have, the form of the standard, normal probability curve of large sample theory. The particular statistic in this illustration is the *percentage* of p events in random samples. Since a percentage is a *proportion* taken to a base of 100 instead of 1.0, the standard errors of percentages are equal to the standard errors of *proportions* times 100:

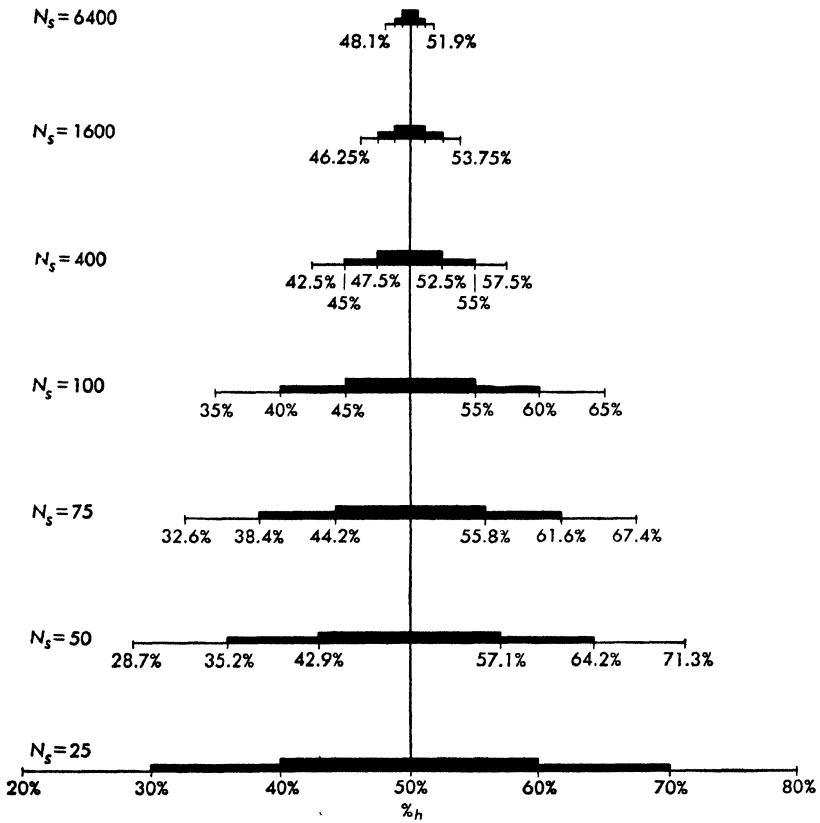
$$\sigma_{\%} = 100\sqrt{pq/N_s} \qquad [12:13]$$

Standard error of a percentage

If we draw random samples of 25 cases each from a universe whose parameter percentage is 50%, we would *in the long run* expect about 68% of the sample results to yield percentages of p events that would vary from 40% to 60%, since the range of $M_{\%_h} \pm 1\sigma_{\%} = 50\% \pm 10.0\% = 40\%$ to 60%. We would expect about 95% of the results to vary from 30% to 70%, since this is the range of $M_{\%_h} \pm 2\sigma_{\%}$ when $N_s = 25$ and $\%_h = 50\%$; and about 99.7% to vary from 20% to 80%, since this is the range of $M_{\%_h} \pm 3\sigma_{\%}$. If the size of the random samples is quadrupled to $N_s = 100$, we would expect the percentages of p events per sample to vary from 45% to 55% in about 2/3 of the samples ($P = .6826$); from 40% to 60% in about 96% of them ($P = .9544$); and from 35% to 65% in more than 99% of them ($P = .9973$). If $N_s = 1600$, the percentage of p events per sample result would be expected to vary, more than 99% of the time, from only 46.25% to 53.75%, and if $N_s = 6400$, from only 48.1% to 51.9%.

Thus the sample values of a statistic (in Fig. 12:4, a percentage of p events) become a more precise or reliable measure of the parameter as the size of samples is increased. If a sample were infinite in size, the value of a sample result would be *precisely* that of the parameter.

Fig. 12:4. The Variability in Sample Results to Be Expected for Different-Sized Samples of a Universe Whose Parameter Percentage Is Taken as 50%; Measured on the Abscissa of the Sampling Distribution in Terms of the Standard Error of the Percentage*



Expected Variation in Sample Results: $M_{\%} \pm 3\sigma_{\%}$

* The form of the sampling distribution in each case is assumed to be the standard, normal probability curve. The range of variation is shown for $M_{\%} \pm 1\sigma_{\%}$; $M_{\%} \pm 2\sigma_{\%}$, and $M_{\%} \pm 3\sigma_{\%}$.
The P value of sample results within the range of $M_{\%} \pm 1\sigma_{\%} = .6826$.
The P value of sample results within the range of $M_{\%} \pm 2\sigma_{\%} = .9544$.
The P value of sample results within the range of $M_{\%} \pm 3\sigma_{\%} = .9973$.

What is a *likely* or *unlikely* result for a single random sample of a given size? As indicated earlier, the answer to this question is based on reasonable expectancy for a given research situation. Generally, however, we would

consider a single sample result whose percentage of p events is *within* the range of $M_{\%_h} \pm 2\sigma_{\%}$ to be *likely* for the given hypothesis (in Fig. 12:4, for the hypothesis that the parameter percentage is 50%). On the other hand, we would consider a single sample result whose percentage of p events is *beyond* the range of $M_{\%_h} \pm 2.5$ or $\pm 3\sigma_{\%}$ *unlikely* for the hypothesis.

In the following chapters we shall consider further the question of what is *likely* and *unlikely* for various hypotheses on the basis of chance, and develop appropriate Tests of Significance for various types of statistics.

EXERCISES

1. Define the concept of probability as used in statistics.
2. Upon what considerations is a probability ratio based?
3. How is the fact that a single result has no probability value dealt with in statistical inference?
4. Under what circumstances is the binomial distribution similar to the normal bell-shaped probability distribution of large sample theory?
5. Define the product and the addition theorems of probability and describe how they are employed in determining the probability of events.
6. Toss ten pennies fifty times and record the number of heads on each trial. Make a histogram of the sampling distribution of the results, and compare the mean and standard deviation of the sampling distribution with the theoretical results that should be obtained in the long run, on the assumption that the coins are *fair*.
7. In Exercise 6, what are the probabilities of obtaining (in the long run) as many as 2 heads per trial? At least 7 heads per trial? No more than 3 heads per trial?
8. What is a Test of Significance? What information is needed in order to make such a test?
9. How are the implications of the normal probability distribution utilized to yield a probability estimate for T?
10. In what sense is there more than one type of "normal" sampling distribution? Cite several examples of different types and describe the circumstances in which they are used.
11. What is the difference between T and the t statistic?
12. From the point of view of sampling theory, when is a sample considered small?
13. What does a sampling distribution that is both skewed and leptokurtic look like? Under what circumstances are sampling distributions of this kind obtained?
14. In what sense does and does not the standard error of a statistic measure the adequacy of a result?
15. What is the relationship between the precision or reliability of a statistic and the size of the sample from which it is obtained?
16. How much larger does a random sample of 150 measurements need to be for the precision of the results to be tripled?

CHAPTER 13

Hypotheses and Tests of Significance

A. LIKELIHOOD AND CONFIDENCE CRITERIA

The usual research situation in the biological and social sciences requires the use of sampling and analytical statistics because the parameter values of the universes studied are ordinarily unknown and not obtainable. We saw in Chapter 11 that the initial problem is designing an investigation so that the *samples* will be *adequate*. Sample results are then analyzed so that their likely implications, i.e., what they signify, can be determined. The preceding chapter made it clear that probability theory is essential to this analysis, which culminates in what has come to be known in statistical parlance as a Test of Significance.

Postulation of Parameters

Usually there are no *empirically* determined parameter values of the universe to be studied. Furthermore, ordinarily the results of only one or, at the most, a few samples are available. Hence, we usually do not have *empirically* established sampling distributions of the statistics in which we are interested, nor do we have *empirical* measures of the standard deviations of such distributions (the standard errors of a measure). Therefore, as a rule, we have no *empirically* determined probability values for a given kind of result.

Such a research position thus requires the *postulation* of *parameter values for relevant statistical hypotheses*. The implications of such hypotheses are *tested* in the light of probability theory and of statistics derived from sample results. In order to test the implications of a statistical hypothesis, we have to make some assumption about the *form* of the sampling distribution of the parameter of the hypothesis. We can then estimate its standard error, which will serve as the basis for determining a relevant probability value. Finally, since the theory of probability describes the behavior of an indefinitely prolonged series of samples rather than a single sample, we have to evaluate a sample result in terms of whether or not it is *likely* for the hypothesis under consideration.

Hypotheses Give Direction and Meaning to Research *

The logic of sampling and analytical statistics is identical with the logic of experimental science generally, in so far as the relationship between hypotheses

* Cf. H. A. Larrabee, *Reliable Knowledge*, Houghton Mifflin, Boston, 1945.

and empirical data is concerned. That is, the logical way to *begin* any research investigation is to start with a hypothesis, and then obtain an appropriate sample of data in order to test the implications of the hypothesis. But, it may be contended, how can we begin a statistical investigation with a hypothesis if our goal is to determine the facts about a situation? Are we not likely to prejudge the empirical character of a result if we begin with a hypothesis about it? The answer of course is *no*, provided we do not permit the hypothesis to bias our observations or warp our conclusions.

We begin with a hypothesis so that a research investigation will not be an aimless collection of data. A hypothesis gives direction and meaning to research. We then try to obtain relevant facts (or sample data) and determine whether or not they are or are not likely for the particular hypothesis. If they are not likely, we reject the hypothesis and consider its logical alternatives. But if they are likely, we may accept the hypothesis as a tenable proposition about the universe studied. However, no amount of sampling and calculation will *prove* the truth of a hypothesis in the strict logical sense of *necessary* inference. Rather, a hypothesis may be found acceptable because (1) its rejection is not warranted by the evidence (sample data), and (2) alternative hypotheses are not *more* likely or *more* acceptable in view of the evidence.

In the preceding chapter some implications of sampling and probability were illustrated for a universe assumed to be equally divided between males and females. Let us now consider a more typical example of empirical research—a universe in which the division of males and females is unknown. A merchant wishes to find out what percentage of his customers are men. Let us *assume* that we can obtain a random sample consisting of 1000 of his customers, and that 52% are men. The parameter value of the percentage of men customers is unknown. It is this value which the merchant wishes us to estimate as accurately as possible from the sample result.

A *sampling distribution* of the percentages of men, where $N_s = 1000$, is not available. It is therefore impossible to compute from *empirical data* the standard error of such a distribution. Consequently, it is also impossible to calculate an *empirical* probability value for the distribution of the sexes in the universe in question. Under these circumstances, what can we do?

If we have a *random* sample of 1000 customers—and this was assumed to be the case—then we can estimate likely parameter values of the percentage of men customers. Although such an estimate cannot be exact, it will indicate a *range* of likely parameter values. The smaller this range, the *more reliable* the estimate will be.

The Probability Estimate

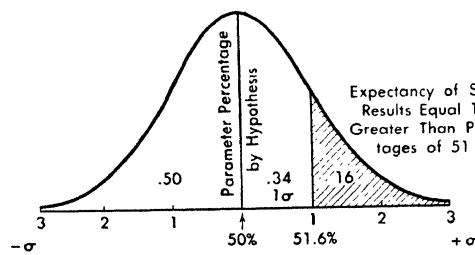
Let us proceed by postulating an even division of the sexes among the merchant's customers. This is our initial statistical hypothesis. In effect, it states that 50% of all his customers are men. The parameter percentage is taken by hypothesis as 50. We can assume that a sampling distribution con-

sisting of 1000 cases per random sample would be *normally* distributed for this particular hypothesis. The best estimate of the standard error of such a sampling distribution of percentages is equal to $100\sqrt{\frac{pq}{N_s}}$, where p is the postulated proportion of men, q is equal to $1.0 - p$, and N_s is the size of the sample. The standard error is therefore:

$$\sigma_{\%} = 100\sqrt{\frac{pq}{N_s}} = 100\sqrt{\frac{(.50)(.50)}{1000}} = 100(.0158) = 1.58\% \text{ or } 1.6\%$$

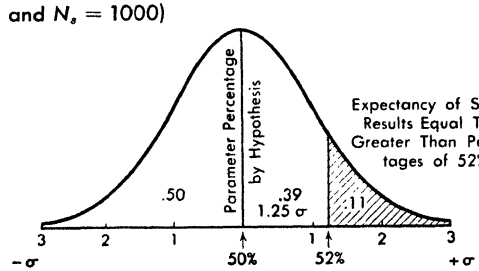
By means of this measure we can estimate the expected variability (probability) of sample results above and below the postulated parameter percentage of 50. For example, the probabilities are approximately .16 (16 in 100) that such random samples will in the long run yield a percentage of males equal to or greater than 51.6%, because the postulated parameter value (50%) plus a value one standard deviation above the mean of the sampling distribution, viz. 1.6%, is equal to 51.6%. In a normally distributed sampling distribution of large sample theory, with a total area equal to 1.00 (or unity), .34 of the results of random samples should yield proportions of males between the parameter percentage of 50 and a sample result one standard error above the mean of the sampling distribution (in this case, 51.6%). There remains .16 of such samples that will in the long run yield percentage values equal to or greater than 51.6%. This is illustrated in Fig. 13:1.

Fig. 13:1. Sampling Distribution for Parameter Percentage of 50% (Where the Total Area = 1.0 and N_s = 1000)



The actual sample result yielded 52% males. Is this a percentage which we would consider *likely* to occur in a single random sample from a universe whose parameter percentage is 50? In order to answer this question, we must first estimate the probability of a sample result

Fig. 13:2. Sampling Distribution for Parameter Percentage of 50% with the Test Ratio of the Test of Significance Equal to 1.25 (Where the Total Area = 1.0 and N_s = 1000)



as great as 52% for the hypothesis under consideration. To do this, we locate the distance (in terms of the standard deviation of the sampling distribution) of the sample percentage from the parameter percentage on the normal probability curve, and then determine the proportion of the area (probabilities) above this point. This has been done in Fig. 13:2, and from it we see that a sample result of 52% males is 1.25 standard deviations above the postulated parameter of 50% males. Referring to Table I, Appendix B, for the distribution of the normal probability function, we find that the proportion of the area between the mean and a distance 1.25 standard deviations above it is .39. The tail of the distribution above this point therefore includes the difference between .50 and .39, or .11 of the total area. Consequently, the probabilities are .11, or 11 in 100, that in the long run random samples of 1000 cases each, drawn from the universe of our statistical hypothesis, will yield percentages of males equal to or greater than 52%. This value of .11 is the probability value needed for a Test of Significance.

The Test of Significance and the Test Ratio (*T*)

Since a Test of Significance is equal to the following ratio:

$$T = \frac{s - h}{\sigma_s}$$

where s is the sample value of a statistic, h the parameter value of the statistical hypothesis to be tested, and σ the standard error of the measure (or statistic) under consideration, T for the preceding data is as follows:

$$T = \frac{52.\% - 50.\%}{1.6\%} = \frac{2.0\%}{1.6\%} = 1.25$$

As already implied in the preceding paragraph, when $T = 1.25$, $P = .11$. This means that there are approximately 11 chances in 100 of obtaining sample results equal to or greater than 52% for the hypothesis under consideration, i.e., a parameter value of 50%. (Cf. Table II, Appendix B, for probability values of T ratios of from zero to 3.0.)

We now need to evaluate the *significance* of this result, whose T ratio is 1.25 and whose P value is .11. We must decide whether the sample result is or is not likely for the hypothesis tested. This is the problem of *likelihood.* In dealing with it, we shall require *confidence criteria*, on the basis of which we may reject or not reject a statistical hypothesis in the light of the T ratio obtained from the Test of Significance.

We have already emphasized that the theory of probability describes the relative frequency of occurrence of an indefinitely prolonged series of sample results and that a single sample result has no probability value. In the long run we would expect that approximately 11 out of every 100 random samples, where $N_s = 1000$, would yield percentages of males equal to or greater than 52% for the universe of the hypothesis (parameter = 50%). What we need

to do is to decide whether our particular sample result is or is not likely, on the basis of random errors in sampling and measurement for a universe whose parameter percentage is 50. In making this decision, we have no absolute principles that are universally valid in all kinds of statistical situations to guide us. Consequently, we usually employ confidence criteria which have been generally found satisfactory in similar investigations. What should our criteria be, on the basis of this experience?

Likelihood and Confidence Criteria *

There is general agreement that when the P value of T is equal to or greater than .10, the statistic of the sample result can confidently be considered as a *likely* result for the hypothesis tested, unless on other grounds there is a strong reason to reject the hypothesis. If we apply this criterion of $p \geqq .10$ to our example, which yielded a P value of .11, we can confidently conclude that the merchant's sample result is *likely* for a universe of male and female customers equally divided with respect to each other. In other words, we cannot with confidence reject the hypothesis that 50% of his customers are males. The sample result of 52% males is too likely for the hypothesis tested to warrant rejection of the hypothesis.

Many investigators take a P value of .05 as the limiting confidence criterion in evaluating the likelihood of a sample result for a hypothesis. In other words, if a Test of Significance yields a T ratio for which the estimated probability value is equal to or greater than .05, the sample result is judged to be only a *chance* divergence from the parameter of the hypothesis being tested. The difference between s and h would be expected on the basis of random errors in sampling and measurement. On the other hand, if P is less than .05, the difference is sometimes considered *significant*. That is, in some research situations a sample result is judged to be unlikely for the hypothesis if the P value of the T ratio is less than .05 (less than 5 chances in 100). These are of course common-sense procedures to be used only when there is no strong reason to accept or reject the hypothesis.

All research investigators agree that when a Test of Significance yields a T ratio whose P value is equal to or less than .001 (1 chance in 1000), the sample value is unlikely for the statistical hypothesis tested. Some investigators, however, consider this too rigorous a criterion in many research situations. A P value of .01 is consequently taken as the limiting confidence criterion in many cases. Thus, if a Test of Significance yields a T ratio whose probability value is estimated to be equal to or *less than* .01 (1 chance in 100), the sample result is judged to be unlikely for the hypothesis. But if the P value is *greater than* .01 the sample result is judged to be likely in some cases.

If we test the hypothesis that the division of males and females among the

* Cf. R. A. Fisher, "Inverse Probability and the Use of Likelihood," *Proceedings of the Cambridge Philosophical Society*, **28**:257–261, 1932.

merchant's customers is 75% males and 25% females, we obtain a value of 16.4 for T, as follows:

$$T = \frac{52\% - 75\%}{1.4\%} = -16.4$$

where 52% is the percentage of males yielded by the sample of 1000 customers; 75% is the parameter value of the hypothesis now being tested; and 1.4% is the new estimate of the standard error of the sampling distribution of this hypothesis $\left(\sigma_{\%} = 100\sqrt{\frac{(.75)(.25)}{1000}} = 1.4\%\right)$. The minus sign with a T value, in this case -16.4, denotes the direction of the value of the statistic from the parameter for the hypothesis. Negative T ratios therefore mean that the statistic is *less* in value than the parameter.

The Test of Significance for this new hypothesis yields a T ratio greater than 16. In other words, the sample result is more than 16 standard deviation units from the parameter value of 75%, whose sampling distribution is assumed to be similar in form to the standard, normal probability distribution. The table of values for the normal probability integral (Table I, Appendix B) does not include z (or T) values greater than 5 because the area of the curve beyond 5 standard deviation units is only a very small fraction of 1%. For this Test of Significance the P value of the T ratio is considerably less than .001. Hence we can confidently reject this particular statistical hypothesis; that is, we can be confident that the sample result was not derived as a random sample from a universe of customers, 75% of whom were males.

Confidence criteria that are used in research generally are taken within the limits of the preceding P values, i.e., $P = .05$ and $P = .001$. A P value of from .05 to .02 is usually taken as the limiting criterion for results judged to be *likely* for the hypothesis tested. On the other hand, a P value of from .01 to .001 is generally taken as the limiting criterion for results judged to be *unlikely* for the hypothesis tested. A P value of .05 is characterized as the *5% confidence* level; a P value of .02 as the 2% confidence level, etc.* These criteria are sometimes referred to as Coefficients of Risk.†

In view of the foregoing confidence criteria, Tests of Significance which yield P values of between .02 and .01 may warrant only a tentative or doubtful inference. Thus, if a Test of Significance yields a T ratio whose P value is .015, we might consider the implications of the result doubtful. We might not reject the hypothesis with confidence, since P is not less than .01. Nor could we, with confidence, conclude that the hypothesis is likely since P is less than .02.

These distinctions may seem to be somewhat arbitrary, and they are, emphatically. It is also to be emphasized that the criteria for likely and

---

* *Ibid.*

† J. G. Smith and A. J. Duncan, *Sampling Statistics and Applications*, McGraw-Hill, New York, 1945, p. 164.

unlikely results should not be taken as a single point value on a probability scale. It would be unsound to take a single confidence criterion, such as $P = .01$, for all types of problems and kinds of data, and then dogmatically accept as *likely* all T ratios yielding P values greater than .01, and reject as *unlikely* all results whose P values are less than .01. *Generally it is recommended that the confidence criteria for an investigation be set up in advance of the Test of Significance, lest the* P *value of the* T *ratio bias the selection of the criteria.*

The 5% Confidence Criterion for Likely Results ($P \geqq .05$)

Bearing in mind the preceding distinctions, we can agree that, in general, T ratios whose P values are greater than .05 (5 chances in 100) signify a sample result that is too likely for the hypothesis tested to warrant its rejection with confidence. The 5% criterion is indicative of a result that is about as *likely* as getting all heads in a toss of 4 or 5 coins. Whether or not we also decide to employ the 2% confidence criterion depends on the particular research situation.

The 0.1% Confidence Criterion for Unlikely Results ($P \leqq .001$)

We can also agree that T ratios whose P values are less than .001 (1 chance in 1000) signify a result that is so unlikely for the hypothesis tested as to warrant its rejection with confidence. The 0.1% criterion is indicative of a result that is about as *unlikely* as getting all heads in a toss of 10 coins. Whether or not we also employ the 1% confidence criterion again depends on the particular research situation.

Confidence Criteria in Terms of *T* Ratios

The above confidence criteria for likely and unlikely results are expressed in terms that are general for any kind of Test of Significance, because they are in terms of the probability value of a result. Once the P value of a result is determined for a given statistic, the confidence criteria in terms of P can be employed, regardless of the form of the sampling distribution. For Tests of Significance that are based on sampling distributions assumed to have the form of the standard, normal probability curve, the evaluation of the T ratio is often simplified by stating the confidence criteria in terms of T itself. Thus, in the literature, a T ratio of 3.0 or more is frequently referred to as a *critical ratio.* Since in large sample theory a T ratio equal to 3.0 has a P value of approximately .001, this is indicative of the 0.1% confidence criterion for an unlikely result, and hence the hypothesis can be rejected with confidence. Reference to Table II, Appendix B, for the normal probability integral shows the above to be the case; i.e., when $T = 3.0$, .49865 of the total area lies between the mean and a point three standard deviation units from it; hence,

.50 − .49865 of the total area (probabilities) lies beyond 3.0σ. This difference is .00135, or approximately .001.

A T ratio of 3.0 or more can thus be taken as signifying a result that is unlikely for the hypothesis and therefore as warranting its rejection with confidence. However, a T ratio less than 3.0 does not warrant the inference that the result is likely for the hypothesis. As already stated, confidence criteria cannot logically be taken in terms of a given point P value that sharply divides likely from unlikely results.

The T ratio equivalent of the 5% confidence criterion can readily be obtained for normal sampling distributions of large sample theory by means of Table I, Appendix B. A point on the abscissa of the normal probability distribution that cuts the area into two parts—viz., 95% and 5%—will be 1.65 standard deviation units from the mean because .45 of the total area lies between the mean and a point 1.65σ from it. Hence, a T ratio of 1.65 or *less* can generally be taken to signify a result that is *likely* for the hypothesis tested.

The T ratio equivalent of the 5% confidence criterion may also be equal to approximately 2.0. This will be the case when the probability of sample results for the parameter value of a given hypothesis is considered with respect to results at both tails of the sampling distribution. Thus, .475 of the total area of the normal probability distribution lies between the mean and a point 1.96σ above or below it. Hence 5% of the probabilities lie beyond ± 1.96, or approximately $\pm 2.0\sigma$. For example, the limits of likely results for the hypothesis that the merchant's customers consist of 50% males would, by this 5% criterion, be $50\% \pm 2.0\sigma_{\%}$. When $N_s = 1000$, $\sigma_{\%}$ was found to be 1.6%. Hence $50\% \pm 2.0(1.6\%)$ gives 46.8% and 53.2% as the limits of likely sample results for the hypothesis in question. On the other hand, $50\% \pm 3.0(1.6\%)$ gives limits beyond which sample results would be unlikely for the hypothesis. These limits are 45.2% and 54.8%. A sample result of 52% males would thus be likely for the hypothesis, as earlier indicated.

The corresponding T ratio values of other percentage confidence criteria can be readily determined by means of the table of the normal probability function, provided of course that the sampling distributions for the statistics in question can be assumed to be similar in form to the standard, normal probability curve of large sample theory. The T value of the 1% confidence level is approximately 2.5.

T ratios of 2.0, 2.5, and 3.0 are thus convenient values for confidence criteria in terms of T. Although these latter T ratios do not precisely correspond to the 5%, 1%, and 0.1% confidence criteria respectively, they are widely used in sampling statistics because they are close enough for all practical purposes; furthermore, they are no more arbitrary than the percentage criteria. Since T ratios of 2.0, 2.5, and 3.0 are rounded values, they serve as convenient reference points for evaluating the results of Tests of Significance.

B. CONFIDENCE LIMITS: TESTING A CONTINUUM OF HYPOTHESES

Many Statistical Hypotheses Can Be Tested

Let us now return to the merchant who wishes to determine what percentage of his customers are men. We have already found that a random sample of 1000 customers which yielded 52% men is a likely result for the hypothesis that 50% are men. We also found that we could confidently reject the hypothesis that 75% of the customers are men. The two Tests of Significance we used led to one hypothesis which is tenable and another which is untenable. Are we warranted in concluding that 50% of the customers are men? We are not, because, as we shall shortly see, there are other tenable hypotheses, other hypotheses with parameter values which, for random samples of 1000 cases each, could on the basis of chance alone yield 52% men.

For example, the hypothesis of 51% men would be just as tenable as the hypothesis of 50% men. A Test of Significance for this new hypothesis is as follows:

$$T = \frac{52\% - 51\%}{100\sqrt{\frac{(.51)(.49)}{1000}}} = \frac{1.0\%}{1.6\%} = .625 \text{ or } .63$$

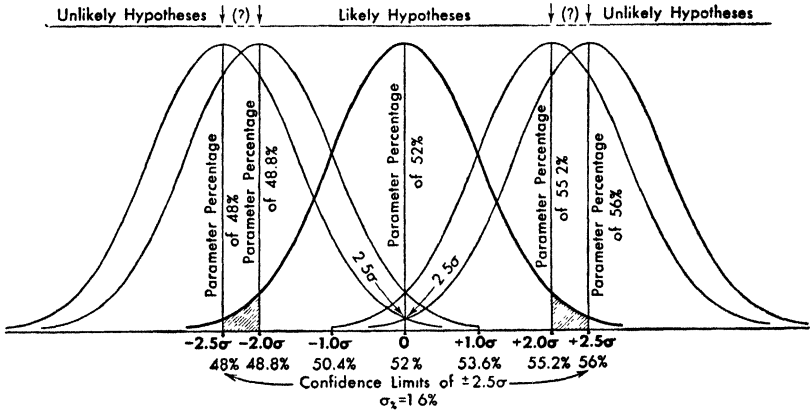
where 52% is the sample result; 51% is the parameter value of the present hypothesis; and the estimate of the standard error of the sampling distribution for this hypothesis is 1.6%.

Table I, Appendix B, reveals that approximately 23% of the area of the normal probability curve lies between the mean and a point .63 standard deviation units above it. The probabilities are therefore 27 in 100 (P equals $.50 - .23 = .27$) of obtaining, on the basis of chance alone, a sample result of at least 52% men from a universe of 51% men. Since the T ratio is less than 1.65 (and its P value is greater than .05), we cannot reject this hypothesis. Similarly, the hypothesis that 53% of the merchant's customers are men would be tenable.

We do not need to limit our hypotheses to percentages that differ by 1%. We could test the hypothesis that 52.5%, or that 52.189%, of the customers are men. Although such fractionation of parameter values is hardly relevant, the point remains that we can test practically an infinite number of different statistical hypotheses between the limits of a zero percentage and a percentage of 100, the hypotheses being different in the sense that their parameter values differ. In attempting to establish a definitive answer to the merchant's question, we shall set up the limiting values of those hypotheses which we can judge as likely or tenable, as well as limits for those which we can definitely reject as unlikely or untenable. The limits of tenable or likely hypotheses are thus the limits of a continuum of possible hypotheses; and, similarly, the limits for unlikely or untenable hypotheses mark off two continua of

hypotheses which we can reject with confidence. Between these two sets of limits, we have two continua of doubtful or tentative hypotheses. (See Fig. 13:3.)

Fig. 13:3. Sampling Distributions for a Continuum of Hypotheses with Confidence Limits for Unlikely Hypotheses Taken at $\pm 2.5\sigma$



The basic point to be remembered is this: We cannot establish the exact value of a parameter from the study of only sample results. However, we can establish a range of possible or likely parameter values, and a range of unlikely or untenable parameter values. The narrower the range of likely parameter values, the more precise (or reliable) the result will be.

Confidence criteria are used to establish the limits of likely hypotheses for a given research problem. Thus, we may take a T ratio of 2.0 (or a probability value of .05) as the criterion for the limits of tenable hypotheses, and a T ratio of 2.5 (or a probability value of .01) as the criterion for the limits of unlikely or untenable hypotheses. These limits are illustrated in Fig. 13:3. The limiting parameter values of tenable or likely hypotheses are readily obtained, since they are equal to the value of the sample percentage, 52%, plus and minus twice the standard error of the percentage. The standard error is thus equal to

$$\sigma_{\%} = 100\sqrt{\frac{(.52)(.48)}{1000}} = 100\sqrt{\frac{.2496}{1000}} = 1.6\%$$

and the sample value of $52\% \pm 2.0(\sigma_{\%})$ is equal to

$$52\% \pm 2.0(1.6\%) = 48.8\% \text{ and } 55.2\%$$

These are the limits of likely hypotheses when a T ratio criterion of 2.0 is employed. Thus the sample result of 52% would yield a T ratio of 2.0 or less for the hypothesis that the parameter percentage is 48.8% (the lower limit), as well as for the hypothesis that it is 55.2% (the upper limit).

We are now prepared to answer the merchant's question. It appears likely that between 49% and 55% of his customers are men. However, we can be much more confident in our report if we use the criteria for unlikely, rather than likely, results. This is the case because the possible parameter value will lie in the continuum between those values that can confidently be *rejected*, rather than within the narrower range of "likely" hypotheses. If we use a T ratio of 2.5 as the criterion for unlikely results, we have the following:

$$52\% \pm 2.5(1.6\%) = 48\% \text{ and } 56\%$$

These are the limiting values of hypotheses which we can reject with confidence as untenable. In other words, it is unlikely that the random sample of 1000 customers came from a universe of customers of which more than 56% or less than 48% were men. Conversely, these values may also be used in this particular problem as the limits of tenable or likely hypotheses. They are in fact limits in which we can have even greater confidence than the 49% and 55% limits already cited. This is so because we are allowing for a wider margin of error when we report 48% and 56% to the merchant as the most likely range within which his men customers are to be found. Market-wise, i.e., in so far as his buying, selling, and advertising policies are concerned, these results are sufficiently precise for the merchant to assume that 50% of his customers are men. However, it should be emphasized that any attempt on his part to "forecast" the *future* percentages of his men customers will result in headaches unless the conditions obtaining while his random sample was drawn continue to operate. Since he usually cannot be sure about this, the best thing for him to do is to sample his customers periodically. By such successive sampling and with care in differentiating *chance* differences from real differences, he can determine the *trend* in the sex of his customers.

Let us apply the criteria we have developed for evaluating the significance of the result in another example. We shall assume that a second merchant takes a random sample of 1000 of his customers and finds that 70% are women. What are the limits for the hypotheses which we can confidently reject? The standard error for the continuum of hypotheses is computed as follows:

$$\sigma_{\%} = 100\sqrt{\frac{(.70)(.30)}{1000}} = 1.4\%$$

and

$$70\% \pm 2.5(1.4\%) = 66.5\% \text{ and } 73.5\%$$

These are the limiting values for hypotheses that are definitely untenable. We can be confident that this merchant obtained his random sample from a universe of customers at least 66% but not more than 74% of whom are women. In other words, we can be confident that from about 2/3 to nearly 3/4 of his customers are women.

Fiducial Limits and Confidence Limits

R. A. Fisher has employed the concept *fiducial limits* to characterize the limits for unlikely hypotheses. However, we prefer the phrase *confidence limits* because it is more descriptive of what they represent.

We have already said that the nature of scientific method is such that hypotheses can never be completely verified in a strict logical sense. But it is of course possible to establish *likely* hypotheses through the refutation of unlikely ones. Giving the facts a chance to nullify a hypothesis is of the essence of a Test of Significance.

The Reliability of a Statistic

The confidence or fiducial limits of a statistic are often interpreted as setting the limits of the "reliability" of a sample result. A measure is the more reliable, the smaller the range of its confidence limits. Since, in random sampling, errors of sampling decrease as the size of the sample is increased, large random samples yield results whose confidence limits are indicative of a fairly precise result. It is again emphasized, however, that simply increasing the size of a sample does not necessarily increase the adequacy of the result. Only if the sample is a random or stratified-random sample of the universe studied can we have full confidence in the results of any Tests of Significance.

C. SUMMARY OF STEPS FOR THE TESTING OF HYPOTHESES

1. The formulation of a general hypothesis. (In a research investigation, a hypothesis more often takes the form of a question, or the statement of a problem. However, the formulation of a precise *statistical hypothesis* (Step 6) is eventually necessary in order that a Test of Significance may be set up and the empirical data be permitted to reveal their implications.)

2. The definition of the statistical universe to be studied. (This is determined in part from the formulation of the problem in Step 1; in the example used above, however, the investigator has to decide whether the merchant's "universe" will be drawn from customers over a period of only a few weeks or of months, etc.)

3. The designing of a research investigation in such a way that an *adequate* sample of data will be obtained for the universe to be investigated. (To be adequate, a method of sampling that will yield a random or a stratified-random sample must be used.)

4. The enumeration or measurement of the characteristics of the sample that are relevant to the investigation (in the above example, the counting of men and of all other customers in the sample).

5. The statistical organization and summarization of the sample data obtained in Step 4, by means of appropriate methods of descriptive statistics (the computation of the percentage of men customers).

6. The selection of one or more relevant *statistical* hypotheses. (This step consists in postulating the value of at least one parameter for the universe studied—for example, a proportion of .50, or a mean I.Q. of 100, or a difference of zero between two means. The choice of statistical hypotheses relevant to the general hypothesis or problem of an investigation is closely related to Step 1.)

7. The formulation of a Test of Significance, viz.,

$$T = \frac{s - h}{\sigma_s}$$

where s is the value of a statistic derived from the sample result; h is the parameter value for the hypothesis to be tested; and σ_s is the standard error of the statistic.

8. The computation of the standard error of the statistic (or statistics) obtained in Step 5, by means of the appropriate formula for the standard deviation of the hypothetical sampling distribution of the statistic. (This step, as we have seen, is necessarily based on some assumption about the *form* of the sampling distribution of the statistic.) *

9. The computation of the Test Ratio, T, from the Test of Significance.

10. The estimation, from the T ratio, of the probability of a given result. (In the case of statistics whose sampling distribution can be assumed to be similar in form to that of the standard, normal probability curve, the differentiation of the area of this curve for T, given in Table II, Appendix B, is the basis for this estimate. For small samples, the Student-Fisher table of probability values for the t ratio (Table III, Appendix B) is used. For hypotheses concerning the distribution of frequencies, the method of chi-square in Chapter 15, and the corresponding probability values in Table IV, Appendix B, may be used.)

11. An inference whether the sample result is *likely* or *unlikely* for the hypothesis tested, i.e., whether or not in the light of experience and all relevant evidence, we judge that it will or will not occur. (This inference is usually based upon *confidence criteria* which should be set up before the result is actually obtained.)

12. A conclusion, formulated in the light of Step 11, concerning the general hypothesis or problem of the investigation.

In many research problems, as in the examples used above, the procedures from Step 6 on are short-cut by establishing confidence (or fiducial) limits that indicate the precision of the result. However, it should be emphasized

* Most of the measurements of standard errors in large sample theory are developed on the assumption that the form of the theoretical sampling distribution is similar to the standard, normal probability curve. Certain exceptions to this were noted in Chapter 12; however, when the statistic under consideration is derived from a *sample* distribution of measurements whose form does not differ significantly (in kurtosis and skewness) from that of the standard, normal probability curve, we are warranted in assuming that the form of the *sampling distribution* of the statistic is "normal."

that this short-cut implies many Tests of Significance of a continuum of hypotheses.

D. TESTS OF SIGNIFICANCE FOR SOME COMMONLY USED STATISTICS

In the preceding sections we have seen that a Test of Significance requires three types of values:

1. A statistic derived from the results of a random or a stratified-random sample of a universe.
2. The parameter value of a relevant hypothesis.
3. A measure or estimate of the standard error of the sampling distribution of the statistic under consideration.

The logic underlying the interpretation of a Test of Significance is similar for all types of statistics. However, as indicated earlier, the sampling distributions may have different forms. Furthermore, in the case of chi-square, differences in sampling distributions are based on the concept of *degrees of freedom* (*d.f.*), rather than on N_s, the size of the sample (cf. Chapter 15).

In the remainder of this chapter we shall present standard error formulas, and Tests of Significance for commonly used statistics whose sampling distribution can be assumed to be similar in form to the standard, normal probability curve. Under these conditions, the probability value of a sample result will be based on the differentiation of the normal probability integral in Table I, Appendix B, and set up directly for T in Table II of that appendix.

Tests of Significance for Percentages

The example used in Sections A and B involved percentage statistics. We saw that the estimate of the standard error of the sampling distribution of percentages is:

$$\sigma_{\%} = 100\sqrt{\frac{pq}{N_s}} \qquad [13:1]$$

Standard error of a percentage

where p is the proportion of measurements or occurrences of a given class, and q is, by definition, always equal to $1.0 - p$.*

We shall present another Test of Significance for this statistic. Several years ago Dr. George Gallup published a survey result which indicated that "64% of voters called federal rationing fair despite some grumbling." † The question asked in this poll, whose object was to secure a picture of people's attitudes toward rationing, was:

DO YOU THINK THE RATIONING OF VARIOUS PRODUCTS IS BEING HANDLED FAIRLY?

* Note that the computation of this standard error is facilitated by reference to Table VII, Appendix B.

† New York *Times*, January 10, 1943.

The following results were reported:

Yes	64%
No	29%
No Opinion	7%
	100%

One question to be asked about these results is whether it is likely that *a majority* of the voters at the time thought rationing was being handled fairly. Since 64% of the sample answered "yes," it is possible that at least 51% (a simple majority) of the population would have answered likewise. A Test of Significance quickly indicates whether this supposition is correct. Since the actual size of the sample is not stated in the report, we shall assume that N_s was equal to *at least* 4000 people. The Test of Significance is, therefore, as follows:

$$T = \frac{s_{\%} - h_{\%}}{\sigma_{\%}} = \frac{64\% - 51\%}{100\sqrt{\frac{(.51)(.49)}{4000}}} = \frac{13\%}{0.79\%} = 16.5$$

where 64% is the sample result; 51% is the parameter value of the hypothesis tested; .51 is the parameter proportion for the hypothesis ($q = .49$ signifies both those who had no opinion as well as those whose answers were "no"); and 4000 is the assumed size of the sample.

The T ratio yielded by this Test of Significance is 16.5. According to the confidence criteria described in Section B, this result is most unlikely for the hypothesis tested, for the T ratio is many times greater than 2.5. Since this particular hypothesis is untenable, we can conclude that more than a majority of the universe of voters believed that the rationing of various products was being handled fairly.

It is also relevant here to determine confidence limits, either for those who thought rationing was being handled fairly, or for those who thought it was being handled unfairly. Let us consider the latter. We shall need, therefore, the limiting values of those hypotheses that can be rejected as definitely untenable for a percentage of 29% (the "no's"). We shall take the confidence limits as equal to $29\% \pm 2.5\sigma_{\%}$. Thus:

$$29\% \pm 2.5\left(100\sqrt{\frac{(.29)(.71)}{4000}}\right) = 29 \pm 2.5(.72\%) = 27.2\% \text{ and } 30.8\%$$

where $p = .29$ is the average parameter proportion for the continuum of hypotheses used to establish the confidence limits; and q is .71.

The confidence limits in rounded values are 27% and 31%. Hence we can be quite confident that at least 27% but not more than 31% of the universe of voters sampled were of the opinion that the rationing of various products was not being handled fairly (assuming a properly drawn sample).

Tests of Significance for Proportions

Tests of Significance for proportion statistics are identical with those for percentage statistics, and therefore require no further illustration. The standard error of a proportion is as follows:

$$\sigma_p = \sqrt{\frac{pq}{N_s}} \qquad [13:2]$$

Standard error of a proportion

It is relevant, however, to consider the following problem which frequently arises in both proportion and percentage statistics. *How large must a sample proportion be to indicate a result greater than would be expected on the basis of chance for a given hypothesis?* Consider, for example, the problem of determining whether there is a difference in the taste of two cola beverages. If samples of the two cola drinks, A and B, are presented in random order to a subject over a period of 25 trials under appropriately controlled experimental conditions, then solely on the basis of chance we would expect the subject's taste judgments to be correct 50% of the time (or an average of 12.5 correct trials in a series of 25 trials).

What proportion of correct judgments does a subject have to give in order to indicate that he can really taste a difference and is not just guessing? If we take the confidence criterion of 3.0 for unlikely hypotheses and restate the formula for T: $T = \frac{p_s - p_h}{\sigma_p}$ in terms of p_s, the required statistic:

then

$$p_s = \sigma_p T + p_h$$
$$= \sqrt{\frac{(.50)(.50)}{25}}\, 3.0 + .50 = .80 \qquad [13:3]$$

To determine the value of a statistic needed for the rejection of a hypothesis

where .50 is the parameter proportion of the hypothesis tested; 25 is the number of trials, N_s; and 3.0 is the confidence criterion. Thus, if a subject makes .80 or more of his judgments correctly, we can reject the chance hypothesis. In a series of 25 trials, .80 is equal to 20 correct judgments; consequently 20 or more correct judgments will warrant the rejection of the chance hypothesis and signify that a subject can taste a difference between the two beverages.

Tests of Significance for Frequencies

Sometimes it is more convenient or desirable to evaluate sample data in terms of frequencies than in proportions or percentages. We saw previously that the standard error of a frequency is equal to

$$\sigma_f = \sqrt{N_s pq} \qquad [13:4]$$

Standard error of a frequency

We could have employed this formula in the preceding example and directly determined the *frequency* of correct judgments necessary for the rejection of the chance hypothesis. Thus, where

$$T = (f_s - f_h)/\sigma_f; \quad f_s = \sigma_f T + f_h$$

Therefore,

$$f_s = \sqrt{25(.50)(.50)}3.0 + 12.5 = 20.0$$

This result is of course the same as that with Formula 13:3; that is, at least 20 correct judgments in 25 trials are necessary for us to reject the chance hypothesis with confidence (by the criterion of $T = 3.0$) and warrant the conclusion that a subject can really taste a difference.

An evaluation of sample results in terms of the frequencies of a class of events rather than of proportions or percentages is useful at times. Generally, however, it is better to convert a frequency to a proportion or percentage.

Tests of Significance for the Arithmetic Mean

The standard error of a sampling distribution of means is given by the following:

$$\sigma_M = \frac{\sigma_u}{\sqrt{N_s}} \qquad [13:5]$$

Standard error of the arithmetic mean

where σ_u is the standard deviation of the measures of the universe from which the sample was drawn and, as usual, N_s is the size of the sample. This formula ordinarily cannot be used, however, because the standard deviation of the universe being sampled is usually not known. Consequently, the estimate of the standard error of a mean must be based on the standard deviation of the distribution of a sample result. The formula for the standard error of a mean therefore becomes

$$\sigma_M = \frac{\sigma}{\sqrt{N_s - 1}} \qquad [13:5a]$$

Standard error of the arithmetic mean

where σ signifies the standard deviation of the distribution of a sample result, and N_s is the size of the sample.

We shall illustrate the use of this formula by means of a random sample of 300 Stanford-Binet I.Q. scores of high-school sophomores in a large city. The mean I.Q. of this sample was found to be 108 and the standard deviation of the distribution was found to be 12.

Inasmuch as the mean I.Q. of an unrestricted universe at this maturity level (in the United States) is assumed to be 100, we may inquire whether the result obtained for this sample of high-school sophomores is likely to hold for the more general, unrestricted universe. We can readily answer this by applying a Test of Significance in which the parameter mean is taken as 100:

$$T = \frac{M_s - M_h}{\sigma_M} = \frac{108 - 100}{\frac{12}{\sqrt{300 - 1}}} = \frac{8}{.69} = 11.6$$

where .69 of one I.Q. unit is the estimated standard error of the sampling distribution, and the difference between the sample result and the mean of the hypothesis tested is 8 I.Q. units.

Since the Test of Significance yields a T ratio of 11.6, we can be confident that the sample mean I.Q., 108, was not derived from a universe whose mean I.Q. is 100. In other words, we can be quite certain that the high-school sophomores of that particular city have a mean I.Q. higher than 100.

Confidence limits for the universe of that city's high-school sophomores would be as follows, with a T criterion of 2.5:

$$108 \pm 2.5(.69) = 106.3 \quad \text{and} \quad 109.7 \text{ I.Q. scores}$$

where 108 (the sample mean) is taken as the representative parameter value of the hypothetical means of a continuum of hypotheses, and .69 is the estimated standard error of the mean. The limits set by the criterion of $2.5T$ are I.Q.'s of 106.3 and 109.7. We can therefore be confident that the given universe of high-school sophomores has a mean I.Q. of at least 106.3 but not more than 109.7.

In using Formula 13:5a, it should be apparent that the subtraction of one case from N_s makes very little difference in the computed value of the standard error when the sample consists of a fairly large number of cases. Thus, in the preceding example, 12 divided by $\sqrt{300 - 1}$ is equal to .694, whereas 12 divided by $\sqrt{300}$ is equal to .693. Both computations give .69 as the standard error of the mean. Consequently, the result of the Test of Significance and the confidence limits would have been the same had $\sqrt{N_s}$ instead of $\sqrt{N_s - 1}$ been used. In practice $\sqrt{N_s}$ is usually used instead of $\sqrt{N_s - 1}$ for samples of 30 or more cases, i.e., samples of large sample theory.

The Reliability of a Mean

In an earlier section, we saw that the confidence limits of a statistic are often interpreted as setting the limits of the reliability of a measure. If this interpretation is applied in the preceding example, the confidence limits of 106.3 and 109.7 I.Q. units are an index of the reliability of the mean I.Q. for the universe sampled. Since a variation of less than 3 I.Q. points is very small from a psychological point of view, it follows that the mean of this sample is very reliable, because we would expect the result of the random sample to have been obtained from a universe with a mean value of between 106.3 and 109.7. The estimated range of possible parameter values for the universe is relatively small.

In psychological measurement and research in related fields, the reliability of a result must usually be judged on a relative basis. That is, there are no absolute standards of reliability. What we can do is to evaluate the reliability of the result in relation to the practical meaning or implications of the measurements analyzed.

Tests of Significance for Test Scores and Other Measures

It is possible to test the significance not only of mean results for a sample but also of different test scores or measures of a distribution. This is a particularly useful type of evaluation because of the importance attached to particular scores or measures in psychological testing. In fact, confidence limits for a test score provide the most practical and meaningful basis for evaluating the reliability of a test. If the confidence limits are relatively great, the test itself may have little or no empirical usefulness for individual diagnosis and prognosis. On the other hand, if they are relatively small, the test can have considerable usefulness.

A Test of Significance for a measure of a distribution can be well illustrated by I.Q. scores since the scale of I.Q. scores itself is familiar in psychology. First, we need the standard error of a measure (or test score). It is estimated as equal to the following:

$$\sigma_X = \sigma_x\sqrt{1 - r_{xx'}} \qquad [13:6]$$

Standard error of a measure

where the subscript X symbolizes a measure of a variable or test, x; σ_x is the standard deviation of the distribution of the variable; and $r_{xx'}$ is the reliability coefficient of the variable or test (cf. Chapter 17, Section B).

If we take the reliability of the Stanford-Binet * as $r_{xx'} = .91$, and the variability, σ_x, as equal to 15 I.Q. units, then

$$\sigma_X = 15\sqrt{1 - .91} = 15(.30) = 4.5 \text{ I.Q. units}$$

Is a Binet I.Q. score of 105 significantly greater than a mean I.Q. score of 100? A Test of Significance will quickly answer this question.

$$T = \frac{X_s - X_h}{\sigma_X} = \frac{105 - 100}{4.5} = 1.1$$

The T ratio, 1.1, signifies a result that is very likely for the hypothesis tested. In other words, an I.Q. score of 105 is not significantly greater than an I.Q. of 100 (i.e., 105 would be expected on the basis of chance errors of sampling) and consequently the psychometrician would not be warranted in concluding that an I.Q. of 105 signifies more Stanford-Binet intelligence than an average I.Q. of 100.

The Reliability of Test Scores

Confidence limits for I.Q. scores, or any other measures of a variable, can readily be established by means of Formula 13:6. However, test scores or other measures of a variable do not always have the same degree of reliability (or standard error) at all points of a scale. Formula 13:6 is most applicable to scores around the mean of a distribution. It may yield either too small or

* Cf. L. M. Terman and M. A. Merrill, *Measuring Intelligence*, Houghton Mifflin, Boston, 1937, chap. 3.

too large a measure of error for extreme values, depending upon the nature of a given variable. Thus, Terman and Merrill found that the standard error of Stanford-Binet I.Q. scores increases as the I.Q. increases; it ranges from a $\sigma_{\text{I.Q.}}$ of 2.2 for I.Q.'s below 70 to a $\sigma_{\text{I.Q.}}$ of 5.2 for I.Q.'s of 130 and over, with a $\sigma_{\text{I.Q.}}$ of 4.5 for I.Q.'s between 90 and 109. I.Q. scores on the basis of which feeble-mindedness or mental deficiency is inferred are therefore considerably more reliable than I.Q. scores that signify superior intelligence.

Confidence limits in terms of a T ratio criterion of 2.0 for likely hypotheses are as follows for a Stanford-Binet I.Q. score of 109:

$$X_{\text{I.Q.}} \pm 2.0(\sigma_X) = 109 \pm 2.0(4.5) = 100.0 \text{ and } 118.0$$

Confidence limits in terms of a T ratio criterion of 3.0 are:

$$109 \pm 3.0(4.5) = 95.5 \text{ and } 122.5$$

Thus it is likely that persons with an I.Q. score of 109 have parameter scores whose values lie between 100.0 and 118.0; and we can be quite confident that the parameter values will not be less than 95.5 or greater than 122.5.

One qualification should be made regarding this interpretation of the reliability of a test score, namely, it is made on the assumption that only *chance* errors of sampling and of measurement are responsible for the variation or difference from the parameter value, whatever it may be. The interpretation does not take into account the constant error factors that might affect the I.Q. score positively or negatively, such as coaching in the particular items (positive bias) or inadequate rapport in the test situation (negative bias).

The reliability of a test score is often expressed in relative terms, viz., the ratio of σ_X to σ, the standard deviation of the sample distribution. In the case of a Stanford-Binet I.Q. near the mean, this ratio would be 4.5/15.0, or approximately 1/3. Thus the effect of chance errors of sampling and measurement on I.Q. scores for this test is about one-third as great as the standard deviation of the total distribution.

The reliability coefficient, $r_{xx'}$, in Formula 13:6 is often obtained by correlating test results from the same sample of individuals on (1) alternate forms of a test, or (2) a second administration of it. If the variability of the test results is different, the best estimate of σ_X is obtained by taking the average of their respective standard deviations:

$$\sigma_X = \frac{\sigma_x + \sigma_{x'}}{2}\sqrt{1 - r_{xx'}} \qquad [13:6a]$$

Standard error of a measure (when two variability estimates of the test are available)

Tests of Significance for Standard Deviations

Occasionally, in research situations, the reliability of the standard deviation of a sample result must be determined. This is usually estimated in terms

of the confidence limits for untenable hypotheses. The best estimate of the standard error of a standard deviation is provided by the following formula:

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2(N_s)}} = \frac{0.707}{\sqrt{N_s}}\sigma, \text{ or } .707\ \sigma_M \qquad [13:7]$$

Standard error of the standard deviation

where σ in the numerator is the standard deviation of the distribution of the sample result, and N_s is the size of the sample.

The use of this formula will be illustrated by Brigham's Army Alpha data, published in his *Study of American Intelligence*. For a sample of 81,465 native-born whites drafted in World War I, Brigham obtained a mean mental age score, calculated from the Alpha test results, of 13.77 years. The standard deviation of this large distribution of mental age scores was 2.86 years. The size of the sample and the standard deviation of the result being known, the standard error of the group's variability in mental age scores is equal to the following:

$$\sigma_\sigma = \frac{2.86}{\sqrt{2(81{,}465)}} = .007 \text{ year of mental age}$$

The standard deviation obtained with the sample result is a precise estimate of the variability of the total universe, consisting in this case of all World War I white draftees born in the United States. The reliability or precision of the result in terms of confidence criteria equal to a T ratio of 2.5 is as follows:

$$2.86 \pm 2.5(.007) = 2.84 \text{ and } 2.88 \text{ years of mental age}$$

Such a precise result is of course to be expected from so large a sample; and, provided the sample was a random sample of the universe studied, we can have great confidence in the accuracy of the measure of variability it yielded.

The Standard Error of the Average Deviation

Although the average deviation is used less frequently than the standard deviation as a measure of variability, it is used sufficiently often to make it worth while for us to know what the standard error of such a statistic is. It is obtained by the following formula, taken in terms of the standard deviation:

$$\sigma_{AD} = \frac{0.603\sigma}{\sqrt{N_s}} \qquad [13:8]$$

Standard error of the average deviation

where as usual σ is the standard deviation of the distribution of the sample result and N_s is the size of the sample.

The standard error of the average deviation can also be estimated from the average deviation. It is equal to the following:

$$\sigma_{AD} = \frac{0.756A.D.}{\sqrt{N_s}} \qquad [13:8a]$$

In terms of *A.D.*

where *A.D.* is the average deviation of the sample distribution and N_s is the size of the sample.

Tests of Significance for Centiles

The centile method for descriptive statistics was presented in Chapter 6 for any type of distribution. When a sampling distribution for any centile measure can be assumed to have the form of the standard, normal probability curve, the measures of standard error to be discussed below can be used for Tests of Significance and confidence limits of centile statistics. The assumption of normality is warranted if the frequency distribution of the sample from which the centile measures are derived tends to be similar to the normal, bell-shaped curve.

Standard Error of Any Centile

In its implications for a Test of Significance, the standard error of any centile is analogous to the standard error of a measure or test score already described. However, the measure or test score is now stated in terms of a centile. Confidence limits for a centile have some advantage over such limits for an original measure of a distribution unless the latter is converted to a z score, or unless the value of σ_X is stated in terms of the standard deviation of the x variable (as it often is). Confidence limits for any centile, however, are themselves stated in relative terms of the centile point system.

The formula for the standard error of any centile is as follows:

$$\sigma_{C_c} = \frac{\sigma}{y}\sqrt{\frac{pq}{N_s}} \qquad [13:9]$$

Standard error of any centile

where σ is the standard deviation of the distribution from which the centile measure is derived; y is the ordinate value at the particular centile point on a normal distribution (see Table I, Appendix B); p is the proportion of the frequencies of the sample distribution which are above or below the particular centile point; q is the remaining proportion of the frequencies $(1.0 - p)$; and N_s is the size of the sample.

The standard error for either the first or second tercile points of a distribution with a standard deviation of 20 and N_s equal to 100, is as follows:

$$\left.\begin{matrix}\sigma_{C_{33}} \\ \text{or} \\ \sigma_{C_{67}}\end{matrix}\right\} = \frac{20}{.364}\sqrt{\frac{(.33)(.67)}{100}} = 54.95(.047) = 2.58$$

where .364 is obtained from Table I, Appendix B, as the value of y at a point that divides the total area of the distribution into 1/3 and 2/3, 1/6 (or .167) of the larger part being below the mean. Since $\sigma_{C_{33}} = 2.58$, confidence limits in terms of $T = 2.5$ will be:

$$C_{33} \pm 2.5\sigma_{C_{33}} = C_{33} \pm 2.5(2.58) = C_{33} \pm 6.3$$
$$= C_{27} \text{ and } C_{40}$$

We can thus be confident that the lower parameter tercile point will not be less than an original score equivalent to the 27th centile or greater than a score at the 40th centile.

With the centile method we can obtain Q, the quartile deviation, more readily than σ, the standard deviation. If the distribution of the sample result is fairly normal (as is assumed to be the case), we can express Formula 13:9 in terms of Q instead of σ, since σ is about 1.5 times larger than Q: $\sigma = 1.483Q$. But Q is a somewhat less reliable measure of variability than σ, and consequently σ is preferred. The standard error of any centile with the measure of variability in terms of Q instead of σ is given by the following:

$$\sigma_{C_c} = \frac{1.483Q}{y}\sqrt{\frac{pq}{N_s}} \qquad [13:9a]$$

Standard error of any centile, in terms of Q

Standard Error of the Median

The standard error of the median, which is the 50th centile, can be obtained from Formula 13:9; but since the values of y, p, and q are always constant, the formula can be simplified to the following:

$$\sigma_{\text{Mdn}} = \frac{\sigma}{0.399}\sqrt{\frac{(.50)(.50)}{N_s}} = \frac{1.253\sigma}{\sqrt{N_s}} \qquad [13:10]$$

Standard error of the median

Or, in terms of Q,

$$\sigma_{\text{Mdn}} = \frac{1.253(1.483Q)}{\sqrt{N_s}} = \frac{1.858Q}{\sqrt{N_s}} \qquad [13:10a]$$

In terms of Q

It will be observed in Formula 13:10 that the standard error of the median is about 25% larger than the standard error of the mean. This confirms statistically what was pointed out earlier, viz., that the mean is a more reliable measure of central tendency than the median.

Tests of Significance and confidence limits for a median are based essentially on the same logical considerations as those for means, and hence will not be discussed here.

Standard Errors of Q_1, Q_3, D_1, and D_9

The following standard error formulas for the first and third quartile points and the first and ninth decile points are commonly used:

$$\left.\begin{matrix}\sigma_{Q_1} = \sigma_{C_{25}} \\ \text{or} \\ \sigma_{Q_3} = \sigma_{C_{75}}\end{matrix}\right\} = \frac{\sigma}{0.318}\sqrt{\frac{(.25)(.75)}{N_s}} = \frac{1.362\sigma}{\sqrt{N_s}} \qquad [13:11]$$

Standard error of quartiles

$$\left.\begin{matrix}\sigma_{Q_1} \\ \text{or} \\ \sigma_{Q_3}\end{matrix}\right\} = \frac{1.362(1.483Q)}{\sqrt{N_s}} = \frac{2.020Q}{\sqrt{N_s}} \qquad [13:11a]$$

In terms of Q

The standard errors of C_{10} and C_{90} are:

$$\left.\begin{matrix}\sigma_{D_1} = \sigma_{C_{10}} \\ \text{or} \\ \sigma_{D_9} = \sigma_{C_{90}}\end{matrix}\right\} = \frac{\sigma}{0.1755}\sqrt{\frac{(.10)(.90)}{N_s}} = \frac{1.709\sigma}{\sqrt{N_s}} \qquad [13:12]$$

Standard error of D_1 and D_9

$$\left.\begin{matrix}\sigma_{D_1} \\ \text{or} \\ \sigma_{D_9}\end{matrix}\right\} = \frac{1.709(1.483_Q)}{\sqrt{N_s}} = \frac{2.534Q}{\sqrt{N_s}} \qquad [13:12a]$$

In terms of Q

Standard Errors of Centile Measures of Variability

Tests of Significance and confidence limits for the quartile deviation (Q), the tercile deviation ($T.D.$), and the C_{10} to C_{90} range (D) are based on the same logic as those for the standard error of the standard deviation. The standard error formulas for each, stated in terms of both σ and Q, are given simply for reference.

The standard error of the quartile deviation, Q, is given by the following:

$$\sigma_Q = \frac{.787\sigma}{\sqrt{N_s}} \qquad [13:13]$$

Standard error of the quartile deviation, Q

$$\sigma_Q = \frac{.787(1.483Q)}{\sqrt{N_s}} = \frac{1.167Q}{\sqrt{N_s}} \qquad [13:13a]$$

In terms of Q

The standard error of the tercile deviation, $T.D.$, is:

$$\sigma_{T.D} = \frac{648\sigma}{\sqrt{N_s}} \qquad [13:14]$$

Standard error of the tercile deviation

$$\sigma_{T.D.} = \frac{.648(1.483Q)}{\sqrt{N_s}} = \frac{.961Q}{\sqrt{N_s}} \qquad [13:14a]$$

In terms of Q

$$\sigma_{T.D.} = \frac{.648(2.317T.D.)}{\sqrt{N_s}} = \frac{1.501T.D.}{\sqrt{N_s}} \qquad [13:14b]$$

In terms of $T.D.$

The standard error of the D range is:

$$\sigma_D = \frac{2.279\sigma}{\sqrt{N_s}} \qquad [13:15]$$

Standard error of the D range

$$\sigma_D = \frac{2.279(1.483Q)}{\sqrt{N_s}} = \frac{3.380Q}{\sqrt{N_s}} \qquad [13:15a]$$

In terms of Q

$$\sigma_D = \frac{.889D}{\sqrt{N_s}} \qquad [13:15b]$$

In terms of D

* This has been derived from the general formula for the standard error of a range. Cf. C. C. Peters and W. R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases*, McGraw-Hill, New York, 1940, pp. 148–150.

Tests of Significance for Product-Moment Correlation Coefficients

An estimate of the standard error of product-moment correlation coefficients is best obtained by converting r to Fisher's **z** function, to be presented later. However, for small values of r a satisfactory estimate can be obtained with the following:

$$\sigma_r = \frac{1 - r_h^2}{\sqrt{N_s}} \qquad [13:16]$$

Standard error of product-moment correlation

where r_h is equal to the parameter value of r for the hypothesis tested, and N_s, as usual, is the size of the sample.

The Hypothesis That r *Equals Zero (The Null Hypothesis)*

Fortunately, Formula 13:16 gives a very good estimate for testing the hypothesis that the correlation coefficient for a universe is zero. The sampling distribution for this hypothesis can be assumed to be normally distributed as long as the samples consist of 25 or more cases.

The hypothesis that r_h is zero illustrates an implication of the concept of *the null hypothesis*. Some investigators classify as null hypotheses all those whose parameter values are taken as equal to zero. More generally, however, a null hypothesis is defined as one that represents what would be expected under fortuitous or chance conditions. The null hypothesis is important in correlation because a Test of Significance permits a decision as to whether or not the sample correlation coefficient could have been obtained from a universe whose parameter r is equal to zero. If this hypothesis cannot be rejected, it follows that the correlation coefficient is likely to occur on the basis of chance alone. Hence, we cannot conclude that there are any determining factors, other than chance factors, underlying the correlation between two variables. On the other hand, if the null hypothesis can be rejected, we can conclude that extra-chance factors account for at least some of the correlation in the sample result. It should be emphasized, however, that the statistical test in itself does not provide any information as to the nature of such factors.

We shall illustrate a Test of Significance for the null hypothesis with the data in Fig. 8:1, consisting of 151 pairs of height and weight measures for a sample of one-year-old girls. The product-moment correlation coefficient was found to be equal to .67. From the point of view of sampling statistics, the question arises as to whether this result can be interpreted as signifying a correlation between height and weight which can be attributable to other than purely chance factors. The Test of Significance for the hypothesis that the sample result is derived from a universe whose parameter r is equal to zero is as follows:

$$T = \frac{.67 - 0}{\frac{(1 - 0^2)}{\sqrt{151}}} = \frac{.67}{\frac{1}{\sqrt{151}}} = \frac{.67}{.08} = 8.4$$

where .67 is the sample r, 0 is the parameter r of the hypothesis, 151 is the size of the sample, and 8.4 is the test ratio, T.*

$$\sigma_{r_{h=0}} = \frac{1}{\sqrt{N_s}}$$

This Test of Significance yields a T ratio of 8.4. Consequently, in accordance with the criterion of a T ratio of 2.5 for a non-chance result, we can definitely reject the null hypothesis. It is most unlikely that the relationship between height and weight observed in this sample of 151 cases can be explained as being due to the operation of purely chance factors. It will have to be explained in terms of extra-chance factors. What these are depends on the nature of the data. In the case of these height-weight measurements, these factors are doubtless inherent in infant development, both height and weight being aspects of organic growth.

The Null Hypothesis and Significance

When a Test of Significance for the null hypothesis yields a T ratio of less than 2.5, the sample coefficient is sometimes described as "insignificant." It is important in such cases to recognize that this use of the concept "insignificant" means that the sample result is not reliably greater than zero. The implications of this conclusion may, however, have great significance with respect to a particular research problem or field of scientific inquiry. Thus, the fact that no correlation significantly greater than zero has been found between intelligence and the shape of people's heads is of considerable significance in psychological theory. The failure to find any correlation between such attributes does not in itself disprove phrenological hypotheses with absolute finality, but it does discredit them and throws the burden of proof on those who support such hypotheses.

Testing Other Hypotheses for r

The null hypothesis is by no means the only relevant hypothesis for many sample product-moment correlation coefficients. It is, however, the first to be tested because there is no point to testing further hypotheses unless the result of a Test of Significance for the null hypothesis warrants its rejection. After all, if the null hypothesis cannot be rejected, we cannot proceed on the assumption that there is any significantly greater correlation between the bi-variates than would be expected on the basis of chance alone.

When the rejection of the null hypothesis is warranted, we can then test the hypothesis that the sample result was derived from a universe with a parameter r of a definite value. A coefficient of .866 (or .87), for example, indicates a degree of correlation which in its predictive value is halfway between no predictive value (zero correlation) and perfect prediction (when r

* The standard error of r for the null hypothesis is the reciprocal of the square root of the size of the sample. The subtraction of one case from N_s is unnecessary when N_s is much greater than 30. See Table I, Appendix C, for square roots and reciprocals.

equals 1.00). Is it at all likely that the correlation coefficient for the sample of height and weight measures could have been obtained from a universe whose r is .866?

A Test of Significance for this problem cannot be based on the standard, normal probability function, because the sampling distributions of correlation coefficients of .67 or more become increasingly skewed as the size of the coefficient increases. R. A. Fisher has developed a transformation function for these high values of r which is approximately normally distributed for any value of r. This transformation function is called **z** and its formula is as follows:

$$\mathbf{z} = \tfrac{1}{2}[\log_e (1 + r) - \log_e (1 - r)] \qquad [13:17]$$

Fisher's **z** transformation function for r

where r is the parameter value of the hypothesis to be tested.* The conversion of values of r to **z** and of **z** to r is facilitated by Table 13:1.

Table 13:1. Values of Fisher's **z** Function for Given Values of r †

r	z	r	z	r	z	r	z
.00	.00	.25	.26	.50	.55	.75	.97
.01	.01	.26	.27	.51	.56	.76	1.00
.02	.02	.27	.28	.52	.58	.77	1.02
.03	.03	.28	.29	.53	.59	.78	1.05
.04	.04	.29	.30	.54	.60	.79	1.07
.05	.05	.30	.31	.55	.62	.80	1.10
.06	.06	.31	.32	.56	.63	.81	1.13
.07	.07	.32	.33	.57	.65	.82	1.16
.08	.08	.33	.34	.58	.66	.83	1.19
.09	.09	.34	.35	.59	.68	.84	1.22
.10	.10	.35	.37	.60	.69	.85	1.26
.11	.11	.36	.38	.61	.71	.86	1.29
.12	.12	.37	.39	.62	.73	.87	1.33
.13	.13	.38	.40	.63	.74	.88	1.38
.14	.14	.39	.41	.64	.76	.89	1.42
.15	.15	.40	.42	.65	.78	.90	1.47
.16	.16	.41	.44	.66	.79	.91	1.53
.17	.17	.42	.45	.67	.81	.92	1.59
.18	.18	.43	.46	.68	.83	.93	1.66
.19	.19	.44	.47	.69	.85	.94	1.74
.20	.20	.45	.48	.70	.87	.95	1.83
.21	.21	.46	.50	.71	.89	.96	1.95
.22	.22	.47	.51	.72	.91	.97	2.09
.23	.23	.48	.52	.73	.93	.98	2.30
.24	.24	.49	.54	.74	.95	.99	2.65

* R. A. Fisher, *Statistical Methods for Research Workers*, Oliver & Boyd, London, 7th ed., 1938, pp. 202–206. (Note that bold-face type is used to distinguish this function from *z* scores.)

† Table 13:1 is adapted from Table VII of Fisher: *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver & Boyd, Ltd., Edinburgh, by permission of the Author and Publishers.

The standard error of z is estimated by the following formula:

$$\sigma_z = \frac{1}{\sqrt{N_s - 3}} \qquad [13:18]$$

Standard error of Fisher's z function

where N_s is the size of the sample. The standard error of the z function is taken independent of the parameter value of r in this formula, which is an approximation formula.

To test the hypothesis that a sample r may have been obtained from a bi-variate universe with a correlation coefficient of .87, we first transform this parameter value of r and the sample result for r (.67 for the height-weight correlation) to Fisher's z function. According to Table 13:1, when r equals .87, z equals 1.33; when r equals .67, z equals .81. Hence the Test of Significance for this hypothesis is as follows:

$$T = \frac{z_s - z_h}{\sigma_z} = \frac{.81 - 1.33}{\frac{1}{\sqrt{151 - 3}}} = \frac{-.52}{.08} = -6.5$$

Since the T ratio is considerably larger than a criterion of 2.5 or 3.0, we can with confidence reject the hypothesis that the correlation between the heights and weights of the 151 infants could have arisen as a random sample of a universe whose correlation is as large as .87.

Although any number of hypotheses for values of r can be tested with Fisher's z transformation function, we shall proceed to determine confidence limits and thereby find the limiting r values of the hypotheses which are likely (or unlikely) for the sample correlation of the height-weight measures.

Confidence Limits for the Reliability of the Sample r

We have seen that confidence limits for a statistic give a measure of its reliability. In this case, the sample r is .67. When r equals .67, z equals .81. Since the standard error of z is independent of the parameter value of r, σ_z is .08, as computed above. The confidence limits in terms of z are therefore as follows:

$$z \pm 2.5(\sigma_z) = .81 \pm 2.5(.08) = .61 \text{ and } 1.01$$

In order for these confidence limits to be meaningful, they must be converted back to their respective r values. As indicated in Table 13:1, when z equals .61, r equals .54; and when z equals 1.01, r equals .76. Consequently, the confidence limits for r in terms of a T criterion of 2.5 are equal to correlation coefficients of .54 and .76. We can therefore be confident that the sample result, .67, was derived from a universe whose product-moment correlation coefficient was at least .54 but not greater than .76. Hence these values of r are the estimated limits of reliability for the sample result. Since a coefficient between .54 and .76 has some predictive value, we can be confident that the

height of infants can be estimated from their weight, or that their weight can be estimated from their height, with a fair degree of accuracy on the average.

Tests of Significance for Other Correlation Coefficients

Estimates of the standard errors for correlation coefficients other than product-moment r are presented here for reference. Their use is analogous to that of product-moment r in testing the null hypothesis, i.e., that the parameter correlation is zero. Although they do not provide very satisfactory estimates of confidence limits, they are satisfactory for determining whether a sample result can be interpreted as significantly greater than zero, i.e., whether there is any correlation between two attributes that cannot be explained on the basis of chance.

Standard Error of Spearman's Rank-Difference Correlation Coefficient (Rho):

The following formula is satisfactory for testing the null hypothesis that a parameter *rho* is equal to zero:

$$\sigma_\rho = \frac{(1 - \rho_h^2)}{\sqrt{N_s}} \qquad [13:19]$$

Standard error of *rho*

When testing the null hypothesis, the numerator is equal to 1, and this formula therefore becomes:

$$\sigma_\rho = \frac{1}{\sqrt{N_s}} \qquad [13:19a]$$

Standard error of *rho*, for the null hypothesis

This estimate of the standard error of *rho* is the same as that for the standard error of r and likewise is unsatisfactory for the determination of confidence limits. A better procedure is to treat *rho* as r and then use Fisher's z transformation function. Tables which are sometimes used for converting *rho* to r are a mathematical over-refinement because at no point is the correction greater than 0.018. This value is usually considerably less than the variation in *rho* or r that can be expected to result from chance errors in sampling and measurement.

Standard Error of a Biserial Correlation Coefficient

The standard error of biserial r is estimated by the following formula when both p and q are greater than .05: *

$$\sigma_{r_{bi}} = \frac{\frac{\sqrt{pq}}{y} - r_{bi}^2}{\sqrt{N_s}} \qquad [13:20]$$

Standard error of biserial r

* No satisfactory formulas are available for use when p or q is less than .05.

where p is the proportion of the total group which is in the higher part of the dichotomized variable; q is equal to $1.0 - p$; y is the value of the ordinate for the normal curve at a point which divides the distribution into two parts, with the proportion of the area above the point equal to p (see Table I, Appendix B); and N_s is the size of the sample. The computation of $\sqrt{pq}$ is facilitated by Table VII, Appendix B.

In testing the null hypothesis for a biserial correlation coefficient, the r term in the numerator becomes zero, and hence the preceding formula becomes:

$$\sigma_{r_{bi}} = \frac{\frac{\sqrt{pq}}{y}}{\sqrt{N_s}} = \frac{\sqrt{pq}}{y\sqrt{N_s}} \qquad [13:20a]$$

Standard error of biserial r, for the null hypothesis

The Standard Error of a Tetrachoric Correlation Coefficient

The formula for the best estimate of the standard error of tetrachoric r is rather complex, but the following formula is satisfactory for testing the null hypothesis:

$$\sigma_{r_t} = \frac{\sqrt{pp'qq'}}{yy'\sqrt{N_s}} \qquad [13:21]$$

Standard error of r_t, the tetrachoric coefficient for the null hypothesis

where p is the proportion of occurrences of a given class for the first variable; p' is the proportion of occurrences of the given class for the second variable; q is equal to $1.0 - p$; q' is equal to $1.0 - p'$; y is the ordinate value of the normal distribution for the point of division between p and q; and y' is the ordinate value for the second variable (see Table I, Appendix B).

Standard Error of the Coefficient of Association

The standard error of the Coefficient of Association in estimating the correlation between two sets of dichotomized non-variable attributes is given by the following:

$$\sigma_A = \frac{1 - A^2}{2}\sqrt{\frac{1}{a}\cdot\frac{1}{b}\cdot\frac{1}{c}\cdot\frac{1}{d}} \qquad [13:22]$$

Standard error of the Coefficient of Association

where a, b, c, and d signify the number of frequencies in the respective cells of the two-by-two cross-tabulation of the two attributes correlated.

In testing the null hypothesis, i.e., that A_h equals zero, the above formula simplifies to one-half the square root of the product of the reciprocals of the frequencies of each of the four cells:

$$\sigma_A = \frac{\frac{1}{a}\cdot\frac{1}{b}\cdot\frac{1}{c}\cdot\frac{1}{d}}{2} \qquad [13:22a]$$

Standard error of A for the null hypothesis

Tests of Significance for the Skewness and Kurtosis of Distributions

Two important properties of uni-modal distributions are (1) skewness and (2) kurtosis. We saw earlier that the skewness of the normal bell-shaped distribution is zero; i.e., the distribution is bilaterally symmetrical above and below the mean. Furthermore, the area of a normal distribution is distributed between successive intervals above and below the mean in the proportions given in Table I, Appendix B. If the results from a sample distribution are not exactly normal but can be treated as if they were derived from a normally distributed universe, most statistics from the sample can be assumed to have normal sampling distributions. Or z scores can be developed for such distributions and be interpreted on the basis of the normal curve.

For these reasons it is often relevant to ascertain whether or not a sample distribution diverges significantly from the normal, probability type. The most exact Test of Significance for this purpose is made in terms of chi-square (see Chapter 15, Section A). However, rough tests can be made for the two properties of skewness and kurtosis. The tests presented are based on the centile method and were developed by T. L. Kelley.*

Test of Significance for Skewness

The skewness of a distribution can be measured in terms of the relationship of the range to the median. If the median is exactly midway between the limits of a distribution, it is more likely to be bilaterally symmetrical than when it is closer to one limit than the other. The limits of a sample distribution, however, are in themselves relatively unstable; consequently the median and the D range, i.e., C_{10} to C_{90}, can be compared since the latter represents (approximately) the most stable limits for measuring a broad range of a sample distribution.

The Centile Measure of Skewness. A measure of skewness therefore can be formulated as follows:

$$Sk = \frac{C_{10} + C_{90}}{2} - mdn \qquad [13:23]$$

Centile measure of skewness

This formula will give zero for skewness if the median is exactly midway between C_{10} and C_{90}. If the skewness is negative, the tail of the distribution is extended toward the lower end of the scale. If the skewness is positive, the tail is extended toward the upper end of the scale.

The value of a measure of skewness is relative to the range of scores as well as to their size. Hence, in order to evaluate the significance of a given measure of skewness for a sample result, we need to test a relevant hypothesis. The most relevant hypothesis is that Sk is zero, since this will be its value if it is derived from a normal distribution. If this hypothesis cannot be rejected with confidence, we can conclude that the skewness in a sample result can

* T. L. Kelley, *Statistical Method*, Macmillan, New York, 1924, p. 77.

be attributed to chance errors of sampling—we can consider that the sample was drawn from a bilaterally symmetrical universe. Such a conclusion signifies not that the sample result was necessarily drawn from a normally distributed universe, but, rather, that it could have been. On the other hand, if the hypothesis that *Sk* is zero can be rejected with confidence, it is unlikely that the sample result was drawn from a normally distributed universe.

The Standard Error of Sk. The standard error of *Sk*, for the hypothesis that *Sk* is zero, is computed as follows:

$$\sigma_{Sk} = \frac{.5185(C_{90} - C_{10})}{\sqrt{N_s}} = \frac{.5185D}{\sqrt{N_s}} \qquad [13:24]$$

Standard error of skewness

Test of Significance for the Skewness of Test Scores. Table 13:2 shows the distribution of the scores of 250 college students on a true-false information

Table 13:2. Distribution of the Test Scores of 250 College Students

Test Scores	f	c.f.
100–101	11	250
98– 99	18	239
96– 97	33	221
94– 95	20	188
92– 93	25	168
90– 91	43	143
88– 89	36	100
86– 87	22	64
84– 85	11	42
82– 83	10	31
80– 81	14	21
78– 79	2	7
76– 77	4	5
74– 75	1	1

C_{90}: $\frac{9}{10}(250) = 225.$ $C_{90} = 97.5 + \frac{4}{18}(2) = 97.94$

C_{75}: $\frac{3}{4}(250) = 187.5.$ $C_{75} = 93.5 + \frac{19.5}{20}(2) = 95.45$

C_{50}: $\frac{1}{2}(250) = 125.$ $C_{50} = 89.5 + \frac{25}{43}(2) = 90.66$

C_{25}: $\frac{1}{4}(250) = 62.5.$ $C_{25} = 85.5 + \frac{20.5}{22}(2) = 87.36$

C_{10}: $\frac{1}{10}(250) = 25.$ $C_{10} = 81.5 + \frac{4}{10}(2) = 82.30$

test on a course in general psychology. The frequencies are cumulated in the right-hand column and the centile values required for measuring the skewness of this distribution are given at the bottom of the table. *Sk* is as follows:

$$Sk = \frac{C_{10} + C_{90}}{2} - mdn = \frac{82.30 + 97.94}{2} - 90.66$$
$$= 90.12 - 90.66 = -.54$$

The distribution is therefore negatively skewed. But is the skewness greater than would be expected on the basis of chance for random samples drawn from a bilaterally symmetrical normal distribution? The Test of Significance that will answer this question is as follows:

$$T = \frac{Sk_s - Sk_h}{\sigma_{Sk}} = \frac{-.54 - 0}{.51} = -1.1$$

In this equation the value of σ_{Sk}, .51, is computed as follows:

$$\sigma_{Sk} = \frac{.5185(D)}{\sqrt{N_s}} = \frac{.5185(97.94 - 82.30)}{\sqrt{250}}$$
$$= \frac{8.11}{15.81} = .51$$

Since T is only 1.1, we cannot reject with confidence the hypothesis that *Sk* is zero; hence the skewness of the sample result is not significantly greater than zero. In other words, so far as skewness is concerned, this sample result could have been drawn from a normally distributed universe.

Test of Significance for Kurtosis

A Centile Measure of Kurtosis In a normal distribution, the quartile deviation (*Q.D.*) is approximately one-fourth as large as the *D* range. This relationship was taken by Kelley as the basis for the following measure of kurtosis:

$$Ku = \frac{(C_{75} - C_{25})/2}{C_{90} - C_{10}} = \frac{Q.D.}{D} \qquad [13:25]$$

Centile measure of kurtosis

In a normal distribution the ratio of *Q.D.* to *D* is computed as .2632, rather than .25. Earlier (page 348) we saw that a normal distribution is characterized as *mesokurtic;* a flattened distribution as *platykurtic;* and a peaked distribution as *leptokurtic.* In terms of the preceding centile measure of kurtosis, a distribution is mesokurtic when $Ku = .2632$; it is platykurtic if $Ku > .2632$; and it is leptokurtic if $Ku < .2632$.

The Standard Error of Ku. The standard error of the centile measure of kurtosis is given by the following:

$$\sigma_{Ku} = \frac{.27779}{\sqrt{N_s}} \qquad [13:26]$$

Standard error of kurtosis

As indicated at the bottom of Table 13:2, C_{75} is 95.45 and C_{25} is 87.36. The values of C_{10} and C_{90} have already been given. Hence,

$$Ku = \frac{(C_{75} - C_{25})/2}{C_{90} - C_{10}} = \frac{(95.45 - 87.36)/2}{97.94 - 82.30}$$
$$= \frac{4.045}{15.64} = .2586$$

And

$$\sigma_{Ku} = \frac{.27779}{\sqrt{N_s}} = \frac{.27779}{\sqrt{250}} = .0176$$

The distribution is therefore not mesokurtic but leptokurtic. But is this divergence from mesokurtosis significantly greater than would be expected on the basis of chance errors in sampling and measurement? The following Test of Significance will answer this question:

$$T = \frac{Ku_s - Ku_h}{\sigma_{Ku}} = \frac{.2586 - .2632}{.0176}$$
$$= \frac{.0046}{.0176} = 0.3$$

where .2632 is the kurtosis of the hypothesis to be tested, i.e., mesokurtosis characteristic of a normal distribution. Since T is only 0.3, we cannot reject this hypothesis. Hence, we conclude that the leptokurtosis of the distribution in Table 13:2 can be attributable to chance errors of sampling and measurement; in other words, the sample could have been drawn randomly from a universe that is normally distributed as far as mesokurtosis is concerned.

E. THE PROBABLE ERROR AND TESTS OF SIGNIFICANCE

Until recent years, Tests of Significance and confidence limits were taken in terms of the probable error (*P.E.*) of a statistic more often than in terms of the standard error, and some statisticians still do so. However, no Tests of Significance have been developed here in terms of *P.E.* for the preceding statistics because the probable error is derived from the standard error and therefore an unnecessary computation is involved in the result. In all sampling distributions that can be assumed to have the form of the standard, normal probability curve, the *P.E.* of any statistic is given by the following:

[13 : 27]

$$P.E._s = .6745\sigma_s$$

Probable error of any statistic whose sampling distribution is normal

where the subscript s represents any statistic. *P.E.* is a measure of variability for a normal sampling distribution that marks off the range of 25% of the sample results above or below the mean. This range is given by $.6745\sigma$ (cf. Table I, Appendix B).

The *P.E.*'s for the statistics in this chapter of the preceding sections are therefore equal to $.6745\sigma$, and are given below for reference.

$$P.E._{\%} = .6745(100\sqrt{pq/N_s}) = 67.45\sqrt{pq/N_s} \qquad [13:28]$$
P.E. of a percentage

$$P.E._p = .6745\sqrt{pq/N_s} \qquad [13:29]$$
P.E. of a proportion

$$P.E._f = .6745\sqrt{N_s pq} \qquad [13:30]$$
P.E. of a frequency

$$P.E._M = .6745\frac{\sigma}{\sqrt{N_s}} = \frac{.6745\sigma}{\sqrt{N_s}} \qquad [13:31]$$
P.E. of arithmetic mean

$$P.E._X = .6745\sigma_x\sqrt{1 - r_{xx'}} \qquad [13:32]$$
P.E. of a measure

$$P.E._\sigma = \frac{.6745(.707\sigma)}{\sqrt{N_s}} = \frac{.477\sigma}{\sqrt{N_s}} = .477\sigma_M \qquad [13:33]$$
P.E. of the standard deviation

$$P.E._{AD} = \frac{.6745(.603\sigma)}{\sqrt{N_s}} = \frac{.407\sigma}{\sqrt{N_s}} \qquad [13:34]$$
P.E. of the average deviation

$$P.E._{C_c} = .6745\frac{\sigma}{y}\sqrt{\frac{pq}{N_s}} \qquad [13:35]$$
P.E. of any centile in terms of σ

$$P.E._{C_c} = \frac{.6745(1.483Q)}{y}\sqrt{\frac{pq}{N_s}} = \frac{Q}{y}\sqrt{\frac{pq}{N_s}} \qquad [13:35a]$$
P.E. of any centile in terms of Q

$$P.E._{Mdn} = \frac{.6745(1.253\sigma)}{\sqrt{N_s}} = \frac{.845\sigma}{\sqrt{N_s}} \qquad [13:36]$$
P.E. of the median in terms of σ

$$P.E._{Mdn} = \frac{.6745(1.858Q)}{\sqrt{N_s}} = \frac{1.253Q}{\sqrt{N_s}} \qquad [13:36a]$$
P.E. of the median in terms of Q

$$P.E._Q = \frac{.6745(.787\sigma)}{\sqrt{N_s}} = \frac{.531\sigma}{\sqrt{N_s}} \qquad [13:37]$$
P.E. of the quartile deviation in terms of σ

$$P.E._Q = \frac{.6745(1.167Q)}{\sqrt{N_s}} = \frac{.787Q}{\sqrt{N_s}} \qquad [13:37a]$$
P.E. of the quartile deviation in terms of Q

$$P.E._{T.D.} = \frac{.6745(.648\sigma)}{\sqrt{N_s}} = \frac{.437\sigma}{\sqrt{N_s}} \qquad [13:38]$$
P.E. of the tercile deviation in terms of σ

$$P.E._{T.D.} = \frac{.6745(.961Q)}{\sqrt{N_s}} = \frac{.648Q}{\sqrt{N_s}} \qquad [13:38a]$$
P.E. of the tercile deviation in terms of Q

$$P.E._{T.D.} = \frac{.6745(1.501T.D.)}{\sqrt{N_s}} = \frac{1.012T.D.}{\sqrt{N_s}}$$

[13 : 38b] *P.E.* of the tercile deviation in terms of *T.D.*

$$P.E._D = \frac{.6745(2.279\sigma)}{\sqrt{N_s}} = \frac{1.537\sigma}{\sqrt{N_s}}$$

[13 : 39] *P.E.* of the *D* range (C_{90} to C_{10}) in terms of σ

$$P.E._D = \frac{.6745(3.380Q)}{\sqrt{N_s}} = \frac{2.280Q}{\sqrt{N_s}}$$

[13 : 39a] *P.E.* of the *D* range in terms of *Q*

$$P.E._D = \frac{.6745(.889D)}{\sqrt{N_s}} = \frac{.600D}{\sqrt{N_s}}$$

[13 : 39b] *P.E.* of the *D* range in terms of *D*

$$P.E._r = \frac{.6745(1 - r_h^2)}{\sqrt{N_s}}$$

[13 : 40] *P.E.* of product-moment *r*

$$P.E._r = \frac{.6745}{\sqrt{N_s}}$$

[13 : 40a] $P.E._r$ for the null hypothesis

$$P.E._z = \frac{.6745}{\sqrt{N_s - 3}}$$

[13 : 41] *P.E.* of Fisher's z function (approximate)

$$P.E._\rho = \frac{.6745(1 - \rho^2)}{\sqrt{N_s}}$$

[13 : 42] $P.E._\rho$ of rank-difference correlation, rho

$$P.E._\rho = \frac{.6745}{\sqrt{N_s}}$$

[13 : 42a] $P.E._\rho$ for the null hypothesis

$$P.E._{r_{bi}} = \frac{.6745\dfrac{\sqrt{pq}}{y} - r_{bi}^2}{\sqrt{N_s}}$$

[13 : 43] *P.E.* of biserial *r*

$$P.E._{r_{bi}} = \frac{.6745\sqrt{pq}}{y\sqrt{N_s}}$$

[13 : 43a] $P.E._{r_{bi}}$ for the null hypothesis

$$P.E._{r_t} = \frac{.6745\sqrt{pp'qq'}}{yy'\sqrt{N_s}}$$

[13 : 44] $P.E._{r_t}$ for the null hypothesis

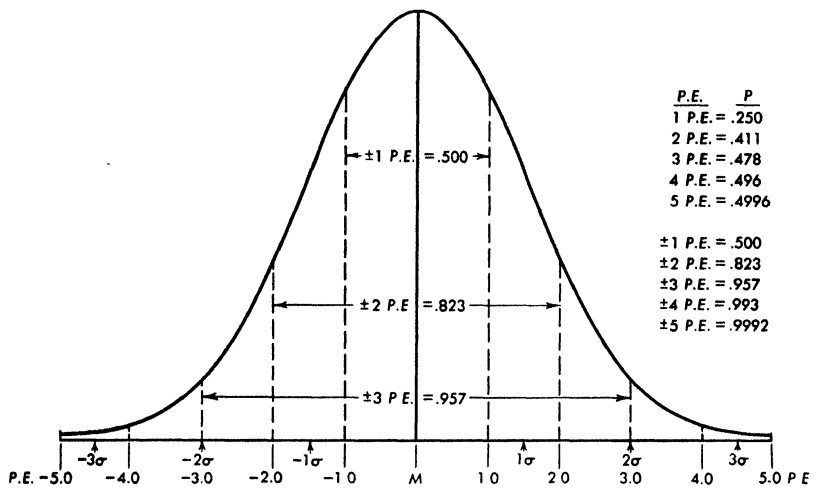
$$P.E._A = \frac{.6745\sqrt{\frac{1}{a}\cdot\frac{1}{b}\cdot\frac{1}{c}\cdot\frac{1}{d}}}{2} = .3373\sqrt{\frac{1}{a}\cdot\frac{1}{b}\cdot\frac{1}{c}\cdot\frac{1}{d}}$$

[13 : 45] $P.E._A$ for the null hypothesis

The Probability Implications of P.E.

Since the *P.E.* of a statistic whose sampling distribution is normally distributed is only about two-thirds (.6745) as large as the standard error, the probability values are somewhat different. The *logic* of any Test of Significance in terms of *P.E.* is the same, however, as of one in terms of σ.

Fig. 13:4. Relation of the Probable Error (*P.E.*) to the Standard Error (σ) for the Normal Probability Distribution



The probability implications of *P.E.* are shown in Fig. 13:4. Since the range of 3.0 standard errors is equal to about 4.6 *P.E.*'s, the practical limits of a normal sampling distribution are often taken as either ± 4 or 5 *P.E.* units.

Percentage confidence levels in terms of *P.E.* are usually rounded as follows, when the probabilities include both tails of the distribution:

$$5\% \text{ level} = \pm 3.0 \; P.E.$$
$$P = .043$$

$$1\% \text{ level} = \pm 4.0 \; P.E.$$
$$P = .007$$

$$0.1\% \text{ level} = \pm 5.0 \; P.E.$$
$$P = .0008$$

If the probabilities of only one tail are concerned, the 1% and 0.1% confidence levels will still be approximately 4.0 and 5.0 *P.E.*, respectively. The 2% confidence level will be 3 *P.E.*

T ratios in terms of *P.E.* may also be used. Thus a *T* ratio of less than 3.0 *P.E.* units provides a satisfactory criterion for likely hypotheses.

$T < 3.0$ *P.E.* = likely hypotheses
$T > 3.0 < 4.00$ *P.E.* = tentative or unlikely hypotheses
$T > 4.0$ *P.E.* = unlikely hypotheses

The P values for ± 5 units of *P.E.* are summarized in Table 13:3. However, the P value of any fraction of *P.E.* can be obtained from Table I, Appendix B, by converting the *P.E.* value into x/σ; this is done by multiplying the *P.E.* value by .6745. Thus, 2.2 *P.E.* units equal 2.2(.6745) standard error units, or 1.48σ.

Table 13:3. Probability Implications in Normal Sampling Distributions of 1 to 5 Probable Error Units

(Total Area = 1.0)

Range	Probability	Range	Probability
Within M + or − 1 *P.E.*	$P = .250$	Beyond 1 *P.E.*	$P = .250$
" M + or − 2 *P.E.*	$P = .411$	" 2 *P.E.*	$P = .089$
" M + or − 3 *P.E.*	$P = .478$	" 3 *P.E.*	$P = .022$
" M + or − 4 *P.E.*	$P = .496$	" 4 *P.E.*	$P = .004$
" M + or − 5 *P.E.*	$P = .4996$	" 5 *P.E.*	$P = .0004$
Within M + and − 1 *P.E.*	$P = .500$	Beyond ±1 *P.E.*	$P = .500$
" M + and − 2 *P.E.*	$P = .823$	" ±2 *P.E.*	$P = .177$
" M + and − 3 *P.E.*	$P = .957$	" ±3 *P.E.*	$P = .043$
" M + and − 4 *P.E.*	$P = .993$	" ±4 *P.E.*	$P = .007$
" M + and − 5 *P.E.*	$P = .9992$	" ±5 *P.E.*	$P = .0008$

F. TESTS OF SIGNIFICANCE FOR SMALL SAMPLES

Fisher's *t* Statistic *

We pointed out earlier that Tests of Significance for large sample and small sample theory are based upon the same logic and that their form is the same in both cases, i.e.:

$$T = \frac{s - h}{\sigma_s}$$ Test of Significance (large sample theory)

$$t = \frac{s - h}{\sigma_s}$$ Test of Significance (small sample theory)

When there are less than 25 or 30 cases in a sample, the probability implications of small rather than large sample theory are utilized. The essence of the distinction between T and t is the fact that the forms and variability of the sampling distributions are different. Consequently, the probability values for a given t ratio will be different from those of a T ratio. (Cf. the probability values of small sample and large sample theory in Table 12:3.)

The t ratio of small sample theory is especially valuable in agricultural economics and biometrics (the fields of research in which Fisher's work has

* R. A. Fisher, *Statistical Methods for Research Workers*, pp. 126 ff.

centered) since often only a few cases are necessary for an adequate result. In general, this is not true in psychology and the social sciences, although there are some notable exceptions, as when the sampling unit itself represents a collective entity. Thus 10 school systems may be an adequate sample for research on school budgets.

The following section on the t ratio for small samples has been included for convenience of reference.

Table 13:4. Distribution of t for Tests of Significance of Small Samples *

$(N_s - 1)$	Probability: P					
	.5	.1	.05	.02	.01	.001
1	1.000	6.314	12.706	31.821	63.657	636.619
2	.816	2.920	4.303	6.965	9.925	31.598
3	.765	2.353	3.182	4.541	5.841	12.941
4	.741	2.132	2.776	3.747	4.604	8.610
5	.727	2.015	2.571	3.365	4.032	6.859
6	.718	1.943	2.447	3.143	3.707	5.959
7	.711	1.895	2.365	2.998	3.499	5.405
8	.706	1.860	2.306	2.896	3.355	5.041
9	.703	1.833	2.262	2.821	3.250	4.781
10	.700	1.812	2.228	2.764	3.169	4.587
11	.697	1.796	2.201	2.718	3.106	4.437
12	.695	1.782	2.179	2.681	3.055	4.318
13	.694	1.771	2.160	2.650	3.012	4.221
14	.692	1.761	2.145	2.624	2.977	4.140
15	.691	1.753	2.131	2.602	2.947	4.073
16	.690	1.746	2.120	2.583	2.921	4.015
17	.689	1.740	2.110	2.567	2.898	3.965
18	.688	1.734	2.101	2.552	2.878	3.922
19	.688	1.729	2.093	2.539	2.861	3.883
20	.687	1.725	2.086	2.528	2.845	3.850
21	.686	1.721	2.080	2.518	2.831	3.819
22	.686	1.717	2.074	2.508	2.819	3.792
23	.685	1.714	2.069	2.500	2.807	3.767
24	.685	1.711	2.064	2.492	2.797	3.745
25	.684	1.708	2.060	2.485	2.787	3.725
26	.684	1.706	2.056	2.479	2.779	3.707
27	.684	1.703	2.052	2.473	2.771	3.690
28	.683	1.701	2.048	2.467	2.763	3.674
29	.683	1.699	2.045	2.462	2.756	3.659
30	.683	1.697	2.042	2.457	2.750	3.646
40	.681	1.684	2.021	2.423	2.704	3.551
60	.679	1.671	2.000	2.390	2.660	3.460
120	.677	1.658	1.980	2.358	2.617	3.373
∞	.674	1.645	1.960	2.326	2.576	3.291

* Table 13:4 is abridged from Table III of Fisher: *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver & Boyd, Ltd., Edinburgh, by permission of the Author and Publishers.

Probability Values for t

Since the logic and procedure for Tests of Significance with small and large samples are similar, the only difference that remains to be considered concerns the determination of the probability of a t result. Such estimates are to be obtained from Table 13:4, the R. A. Fisher table of probability values for t of small samples. The t values are given in the body of Table 13:4. P values are given at the top of each column, and the size of the sample is given in terms of $N_s - 1$ in the left-hand column.*

The P values in the table include the probabilities for both tails of the sampling distribution. Thus, when N_s is 15 and when t is 3, the $N_s - 1$ row, i.e., 14, shows that the probabilities are about 1 in 100 that the result might diverge 3.0σ or more above or below the parameter mean of the sampling distribution.

Table 13:4 can be used with Tests of Significance for the various statistics discussed in the preceding sections, and it should be used when N_s is less than 25 or 30.

EXERCISES

1. In what way does a hypothesis give meaning and direction to a research investigation?
2. For any Test of Significance, distinguish between making a probability estimate and evaluating the likelihood or unlikelihood of a result.
3. What is the meaning of a confidence criterion? Distinguish between the different kinds of confidence criteria that are employed in statistical work, and indicate the relation between percentage confidence levels and T ratio criteria.
4. Give an example of a Test of Significance in which the probability of only one tail of the sampling distribution must be considered in evaluating the significance of the T ratio.
5. Give an example of a Test of Significance in which the probabilities of both tails of the sampling distribution are relevant to an evaluation of the significance of the T ratio.
6. What is the difference between confidence criteria and confidence limits? For what particular purpose are the latter employed?
7. What is the null hypothesis? Give several examples of statistical hypotheses that are of this form.

Set up relevant Tests of Significance for the data in the following nine problems, determine the value of T for each test, and interpret your results:

8. Of a random sample of 500 adults, 45% say they will vote for Mr. A; 35% say they will vote for Mr. B, and the remainder say they will vote for other candidates.

* This table was originally developed in terms of *degrees of freedom*, which can usually be treated as equal to $N_s - 1$ for the Tests of Significance of the statistics considered in this chapter (cf. chap. 15).

9. Use the results of the Gallup Poll on a loan to England presented on page 317.
   a. For the NATIONAL SAMPLE RESULT, assume a total sample of 4000 cases.
   b. For OCCUPATIONS, assume that the sample consisted of 500 business and professional people, 1000 white-collar workers, 1000 farmers, and 1500 manual workers.
   c. For EDUCATION, assume that the college group consisted of 300 people; that the high-school group consisted of 2700, and that the grammar or no-school group consisted of 1000.
   d. For POLITICAL PARTY, assume that the Republican voters totaled 1900 and the Democratic voters totaled 2100.

10. The mean percentage grade score received by a random sample of 75 college students is 80%; the standard deviation of the sample result is 7%.

11. A subject is given 50 trials to judge whether pairs of weights are similar or different; the order of like and unlike pairs is randomized. The subject's judgments are correct in 30 of the 50 trials.

12. A subject is presented a series of 20 pairs of tones of the same pitch, each pair differing in intensity; he is to judge whether the second tone of each pair is louder or softer than the first. His judgment is correct in 11 cases out of the 20.

13. A person receives a score of 124 on a test whose standard deviation is 25 test score units, whose mean is 110, and whose reliability coefficient is estimated to be .88.

14. The product-moment correlation between the intelligence test scores and the heights of a group of 50 high-school students is .05.

15. The product-moment correlation between the mechanical aptitude test scores and the ages of a group of 75 high-school seniors is $-.11$.

16. With the variables in Table 6:7 (page 148), determine:
   a. the kurtosis of each distribution
   b. the skewness of each distribution

Set up confidence limits and interpret the reliability of the statistics in the following four problems:

17. The mean and standard deviation for the 1838 policemen and firemen on the Bennett Mechanical Comprehension Test (page 187) were 35.6 and 10.1 respectively.

18. The median score made on an intelligence test by a random sample of 150 people is 83. The quartile deviation of the sample result is 13.

19. Use either variable given in Table 6:7 (page 148) and determine:
   a. the tercile deviation
   b. the D range
   c. the 12th vigintile
   d. the quartile deviation
   e. the 4th quintile

20. A product-moment correlation coefficient of .80 between an intelligence test and an achievement test is obtained from a random sample of 90 high-school seniors. Set up confidence limits both in terms of r derived from the standard error of r, and in terms of r derived from the standard error of Fisher's $\mathbf{z}$ function.

CHAPTER 14

Tests of Significance for Differences Between Statistics

Any Test of Significance for a difference between two statistics is a test of the hypothesis that the parameter difference between two statistics is equal to zero. We shall characterize all such hypotheses as null hypotheses.* The importance of Tests of Significance for differences should be apparent. If a difference between two statistics is not significantly greater than zero, we cannot infer that the difference obtained from the sample result is more than a chance difference.

Research problems often require Tests of Significance for the hypothesis of "no difference." For example, in studying psychological differences between two groups, a comparison of sample results may show a difference, but a Test of Significance may not indicate a difference significantly greater than zero. In the latter case the result is based on random samples of the two groups and if the measures of psychological characteristics are determined by satisfactory methods, we can be confident that no real differences exist in the characteristics compared. Again, in any controlled, experimental investigation, a Test of Significance for any sample difference is crucial in determining whether the result yielded by the experimental variable is actually any different from what would be expected on the basis of chance alone. This point is illustrated in some of the examples in this chapter.

A. THE STANDARD ERROR OF A DIFFERENCE BETWEEN ANY TWO STATISTICS

The general formula for the standard error of a difference between any two statistics is

$$\sigma_{(s_x - s_y)} = \sqrt{\sigma_{s_x}^2 + \sigma_{s_y}^2 - 2r_{s_x s_y}\sigma_{s_x}\sigma_{s_y}} \qquad [14:1]$$

Standard error of a difference between two statistics

* Some authors, following R. A. Fisher, restrict the use of the concept *null hypothesis* to those situations in which the statistics under consideration are, by hypothesis, taken as random samples from the same universe. In such cases the variance of the universe whose parameter is assumed to be zero is estimated from the combined results of the samples. In this chapter, however, we shall use the variance of each statistic, following the procedures of classical statistics. (Cf. C. C. Peters and W. R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases*, McGraw-Hill, New York, 1940, pp. 177 ff.)

where s_x represents any statistic (proportion, mean, median, standard deviation, correlation coefficient, etc.) derived from sample x, and s_y represents any other statistic derived from sample y which is compared with the first; $\sigma_{s_x}^2$ is the variance of the sampling distribution of the first statistic, and $\sigma_{s_y}^2$ is the variance of the sampling distribution of the second statistic; $r_{s_x s_y}$ represents the correlation between the two statistics compared, and σ_{s_x} and σ_{s_y} are the standard errors of each statistic. The standard error formulas for statistics whose sampling distributions can be assumed to be normal have already been given in the preceding chapter; they are used in Formula 14:1. The only additional value required is the correlation coefficient, $r_{s_x s_y}$, between the two statistics whose difference is being compared.

Formula 14:1 is general and can be used in analyzing differences between statistics when the standard error of each statistic, as well as any correlation between them, can be satisfactorily estimated. Most Tests of Significance for a difference between two statistics are set up for two statistics of the same class. That is, differences between proportions, or between means or between correlation coefficients are usually analyzed rather than a difference between a proportion and a mean, for example.

Standard Error of a Difference for Independent Samples

Whenever a difference between two statistics is obtained from independent samples, the correlation between them is zero. Consequently, the third term of Formula 14:1 is equal to zero, and the general formula for the standard error of a difference simplifies to the following:

[14 : 2]

$$\sigma_{(s_x - s_y)} = \sqrt{\sigma_{s_x}^2 + \sigma_{s_y}^2}$$

Standard error of a difference between two statistics derived from non-correlated samples

Thus the estimate of the standard error is the square root of the sum of the variances of the sampling distributions of each statistic. Even when the statistics compared are drawn from *dependent* samples, and hence there is likely to be some correlation between them, it may be unnecessary to compute the correlation coefficient for the third term of the general formula. If the standard error of the difference is based on the abbreviated formula and yields a T ratio equal to or greater than 2.5 or 3.0, the rejection of the null hypothesis is usually warranted, despite the possibility of a positive correlation between the two statistics. The third term will be negative if there is any *positive* correlation. A negative term will decrease the estimate and hence increase the value of the T ratio. If, however, there is a negative correlation between the two statistics, the standard error of the difference will be increased and the T ratio will be smaller.

On the other hand, when a result derived from dependent samples yields a T ratio of less than 2.5, the use of the third term may make a real difference

in the conclusion drawn from the Test of Significance. Hence, under such circumstances, the correlation coefficient should be computed.

B. TESTS OF SIGNIFICANCE FOR A DIFFERENCE BETWEEN ANY TWO STATISTICS

The logic of a Test of Significance for a difference between any two statistics is the same as that described in the preceding chapter for Tests of Significance of single statistics. Thus the Test of Significance yields a test ratio, T, as follows:

$$T = \frac{\text{sample difference} - \text{parameter difference of zero}}{\text{standard error of difference}}$$

$$= \frac{(s_x - s_y) - 0_h}{\sigma_{(s_x - s_y)}} \qquad [14:3]$$

T ratio for Test of Significance for a difference between two statistics

where $(s_x - s_y)$ is the difference between the two statistics; zero is the parameter value of the difference for the hypothesis tested; and $\sigma_{(s_x - s_y)}$ is the standard error of the difference between the two statistics.

This Test of Significance yields a T ratio similar in its implications to the T ratios already considered. Whenever there is justification for assuming that the sampling distribution of differences is normal, the distribution for T in Table II, Appendix B, can be used to obtain a probability estimate for interpreting the result. When the samples are random and consist of 25 or more cases, the assumption regarding the sampling distribution of differences is usually warranted for the statistics considered in the preceding chapter.

Confidence Criteria for the Significance of a Difference

To determine whether the probability value of a test ratio warrants the rejection of the null hypothesis, we shall employ the same criteria as were used earlier. In other words, if the T ratio is equal to or greater than 2.5 (approximately the 1% confidence level), we shall ordinarily conclude that the difference for the sample result is unlikely on the basis of chance alone. Hence we can reject the null hypothesis and infer that the difference is attributable (at least in part) to extra-chance factors.

On the other hand, if the T ratio is less than 2.0 the rejection of the null hypothesis is not warranted. Hence in such cases we shall conclude that the difference is likely on the basis of chance alone and is therefore "insignificant." As was emphasized regarding a Test of Significance for a correlation coefficient (Chapter 13), a difference may not be statistically "significant," but it may have real significance from the psychological or research point of view.

If the T ratio is equal to or greater than 2.0 but less than 2.5, we shall ordinarily conclude that the difference is of doubtful significance. Such a

ratio warrants only a tentative or doubtful inference with respect to the rejection or non-rejection of the null hypothesis.

C. TESTS OF SIGNIFICANCE FOR A DIFFERENCE BETWEEN PERCENTAGES (OR PROPORTIONS) DERIVED FROM NON-CORRELATED SAMPLES

The standard error of a difference between percentage or proportion statistics is obtained by substituting the standard error of each statistic in Formula 14:1. If the statistics are derived from independent (non-correlated) samples, the third term in the formula is zero and the standard error of the difference becomes a special case of Formula 14:2. Thus, for proportions:

$$\sigma_{(p_x - p_y)} = \sqrt{\sigma_{p_x}^2 + \sigma_{p_y}^2} = \sqrt{\frac{p_x q_x}{N_x} + \frac{p_y q_y}{N_y}} \qquad [14:4]$$

Standard error of a difference between proportions derived from non-correlated samples

If percentage statistics rather than proportions are used, the standard error obtained with Formula 14:4 is multiplied by 100:

$$\sigma_{(\%_x - \%_y)} = 100 \sqrt{\frac{p_x q_x}{N_x} + \frac{p_y q_y}{N_y}} \qquad [14:5]$$

Standard error of a difference between percentages derived from non-correlated samples

A Test of Significance for the difference between two percentages is as follows:

$$T = \frac{(\%_x - \%_y) - 0}{\sigma_{(\%_x - \%_y)}} \qquad [14:6]$$

T ratio for a Test of Significance of the difference between two percentages

where the parameter difference of the hypothesis tested is taken as zero. The usefulness of this Test of Significance will be illustrated in connection with the evaluation of differences among groups in an opinion poll.

Elmo Roper, director of the *Fortune* magazine polls, in 1945 conducted an opinion poll * of a national stratified sample of adults on the question:

> WE WANT TO KNOW HOW THE PUBLIC RATES PRESIDENT TRUMAN ON SEVERAL SPECIFIC THINGS, FROM WHAT THEY HAVE SEEN OF HIM UP TO NOW. SO FAR AS HIS HANDLING OF OUR RELATIONS WITH FOREIGN COUNTRIES, CONGRESS, HOME PROBLEMS GOES, WOULD YOU SAY PRESIDENT TRUMAN IS DOING AN EXCELLENT, GOOD, ONLY FAIR OR POOR JOB?

* New York *Herald Tribune*, October 13, 1945.

The opinions of the respondents who supported Roosevelt and Dewey in 1944 were analyzed for each aspect of the question. For home problems, the breakdown was as follows:

President Truman's Handling of Home Problems	Candidate Voted for or Favored in 1944 (R) Roosevelt	(D) Dewey
Excellent	13.6%	11.3%
Good	48.8	44.8
Only fair	19.3	26.1
Poor	3.4	5.0
Don't know	14.9	12.8
	100.0%	100.0%

These two groups of results represent independent sub-samples of the total sample and are therefore uncorrelated. Let us assume that each sub-sample consists of 1000 cases, so that $N_R = 1000$ and $N_D = 1000$. (Roper did not report the actual size of his sample.) For convenience the Roosevelt and Dewey groups have been taken as equal, although in a stratified-random sample a somewhat greater proportion of Roosevelt supporters would be expected in view of the actual 1944 election returns. The results when combined are as follows:

President Truman's Handling of Home Problems	Roosevelt Supporters	Dewey Supporters
Excellent or good	624	561
Only fair or poor	227	311
	851	872
Don't know	149	128
Total (N assumed)	1000	1000

We shall consider only the results for the respondents who had an opinion (eliminating the *D.K.*'s). The percentages, based on those who held an opinion, now become:

	$\%_R$	$\%_D$
Good or excellent	(624/851) = 73.3%	(561/872) = 64.3%
Only fair or poor	(227/851) = 26.7%	(311/872) = 35.7%
	100.0%	100.0%

Is the difference between the opinions of the R and D samples significantly greater than zero? The Test of Significance used here is as follows (the comparison being made in terms of the percentage of *favorable* opinions; the result would be the same, however, if the percentage of unfavorable opinions were used):

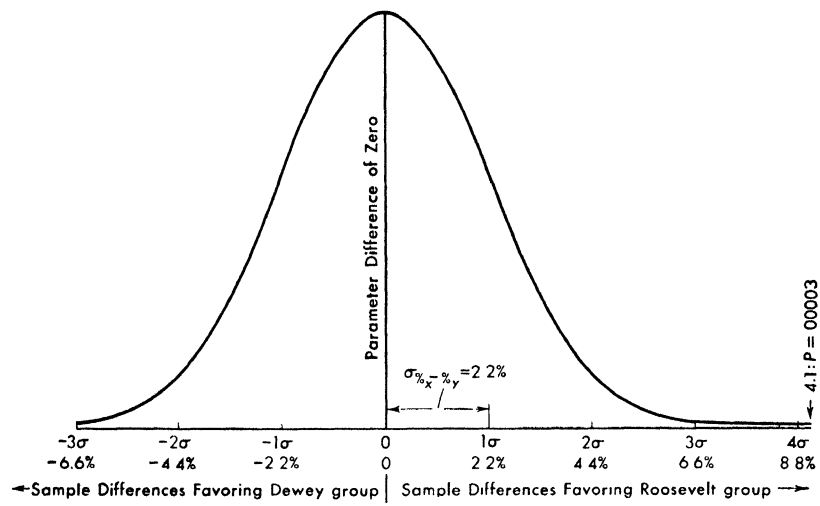
$$T = \frac{(\%_R - \%_D) - 0}{\sigma_{(\%_R - \%_D)}} = \frac{(73.3\% - 64.3\%) - 0}{100\sqrt{\frac{(.733)(.267)}{851} + \frac{(.643)(.357)}{872}}} = \frac{9.0\%}{2.2\%} = 4.1$$

Since the T ratio is greater than 2.5 or 3.0, we can reject with confidence the null hypothesis that the difference is zero. In other words, the difference is

too great to be likely on the basis of chance for the hypothesis of *no* difference.

It follows, then, that at the time of this poll more Roosevelt supporters than Dewey supporters thought that Truman was doing a good or excellent job in handling home problems. (However, a considerable majority of both groups held this opinion.)

Fig. 14:1. Sampling Distribution of a Difference Between Two Percentages for the Hypothesis That the Difference Is Zero. (Roper Data)



Let us now consider the statistical implications of this result in relation to the sampling distribution of differences whose standard error was found to be 2.2%. We assume that the hypothetical sampling distribution is normal, as shown in Fig. 14:1. The mean of this distribution is the parameter percentage for the difference of the hypothesis tested, or zero. All sample differences above this mean indicate that more Roosevelt supporters had favorable opinions; all sample differences below this mean indicate that more Dewey supporters had favorable opinions. The statistical problem is to determine whether the difference obtained from a single sample result diverges so far above the parameter percentage of zero as to indicate that a zero difference or a difference in the other direction is unlikely.

As indicated in Fig. 14:1, when the T ratio is 4.1, the difference is 4.1 times the standard error. The probabilities are only .00003 (3 in 100,000) of obtaining on the basis of chance a sample difference as large as the one obtained, viz., 9.0%.

D. TESTS OF SIGNIFICANCE FOR A DIFFERENCE BETWEEN PERCENTAGES DERIVED FROM CORRELATED SAMPLES

The preceding Test of Significance involved two sub-samples which were uncorrelated. If, however, two sub-samples are matched by individual pairs, as in the experimental method of matched groups, the standard error of the difference will undoubtedly be somewhat reduced because the technique of matching pairs usually restricts the fluctuations (or variability) in sampling. The correlation between the two statistics in the third term of Formula 14:1 takes this restriction into account.

For percentages, Formula 14:1 becomes:

$$\sigma_{(\%_x - \%_y)} = 100\sqrt{\sigma_{p_x}^2 + \sigma_{p_y}^2 - 2r_{p_x p_y}\sigma_{p_x}\sigma_{p_y}}$$

$$= 100\sqrt{\frac{p_x q_x}{N_x} + \frac{p_y q_y}{N_y} - 2r_{xy}\sqrt{\frac{p_x q_x p_y q_y}{N_x N_y}}}$$

[14:7]

Standard error of the difference between two percentages derived from matched or correlated samples

with the correlation between two proportions, $r_{p_x p_y}$, equal to r_{xy}, the correlation between the paired results or variates of the matched samples. Very often r_{xy} is not obtainable for proportions or percentages. If the standard error estimated with Formula 14:5 shows that the difference is significant, there is ordinarily no need to be concerned with the r term in Formula 14:7. But if Formula 14:5 gives a T ratio between 2.0 and 2.5, or less than 2.0, Formula 14:7 may yield a T ratio greater than 2.5, and thus warrant the inference that the difference is significant.

This type of problem with matched samples is illustrated by the following example, in which an analysis is made of the effect, on the attitudes of listeners, of introducing a commercial announcement in the middle of a radio program.*

Two groups of 100 listeners each were matched pair by pair with respect to age, sex, education, and general attitude toward the type of program to be used in the experiment. The "control" group, C, heard the program without the commercial. The experimental group, E, heard the program with the commercial. The over-all attitudes of Groups C and E toward the program are as follows:

Group C: Like = 60%; Dislike = 40%
Group E: Like = 50%; Dislike = 50%

Thus, three-fifths of the control group liked the program when it did not include the commercial, whereas the experimental group, which heard the com-

* The methodological technique of the Program Analyzer developed by Drs. Paul Lazarsfeld and Frank Stanton is employed in such analyses. Cf. J. G. Peatman and T. Hallonquist, *The Patterning of Listeners' Attitudes Toward Radio Broadcasts: Methods and Results*, Stanford Univ. Press, Stanford University, 1945, especially chap. 1.

mercial, was evenly divided between "like" and "dislike." Is the difference of 10 percentage points in the "likes" of the two groups significantly greater than zero? The Test of Significance which must be set up in order to answer this question is as follows:

$$T = \frac{(60\% - 50\%) - 0}{\sigma_{(\%_C - \%_E)}}$$

If the hypothesis that the difference is zero can be rejected, then, in view of the matching controls introduced, the more favorable attitudes of Group C should be attributable, at least in part, to the absence of the middle commercial from the program they heard.

The T ratio obtained in terms of the standard error of the difference, the latter estimated by Formula 14:5, is as follows:

$$T = \frac{(60\% - 50\%) - 0}{100\sqrt{\frac{(.60)(.40)}{100} + \frac{(.50)(.50)}{100}}} = \frac{10\%}{100\sqrt{.0049}} = \frac{10\%}{7\%} = 1.4$$

Since the T ratio is 1.4, the null hypothesis cannot be rejected.

A more accurate estimate of the standard error of the difference can be obtained with the complete formula (14:7), which takes into account any correlation between the attitudes of these matched samples. In order to do this, the attitudes of the two groups toward the program must be correlated. Since their attitudes are dichotomized, a tetrachoric coefficient can be used for r_{xy} in Formula 14:7. The cross-tabulation is as follows:

		Group C		
		(q_C) Dislike	(p_C) Like	N_E
Group E	(p_E) Like	8 (.08)	42 (.42)	50
	(q_E) Dislike	32 (.32)	18 (.18)	50
	N_C	40	60	100 (N_s)

$r_t = .71$

There is obviously a fairly high degree of correlation between the attitudes of the two groups. Reference to Thurstone's *Computing Diagrams* * shows that the correlation is .71. The standard error is now as follows:

* L. Chesire, M. Saffir, and L. L. Thurstone, *Computing Diagrams for the Tetrachoric Correlation Coefficient*, Univ. of Chicago Bookstore, Chicago, 1933.

$$\sigma_{(\%C-\%E)} = 100\sqrt{\frac{(.60)(.40)}{100} + \frac{(.50)(.50)}{100} - 2(.71)\sqrt{\frac{(.60)(.40)(.50)(.50)}{10{,}000}}}$$

$$= 100\sqrt{.0049 - .0034} = 100\sqrt{.0015} = 3.9\%$$

The Test of Significance is therefore:

$$T = \frac{(60\% - 50\%) - 0}{3.9\%} = \frac{10\%}{3.9\%} = 2.6$$

The T ratio is greater than 2.5, which indicates a probability of less than 1 in 100 of a difference as great as 10% occurring on the basis of chance errors in sampling and measurement in a universe with a parameter difference of zero. We are warranted in rejecting this hypothesis with confidence and concluding that the introduction of a commercial announcement in the middle of the program had a negative effect (increased the dislikes) on the attitudes of the listeners.

The importance of the more precise estimate made possible by Formula 14:7 is obvious from this example. The result completely reverses the inference that would have been made had the correlation between the two matched groups been unknown or been assumed to be zero.

E. TESTS OF SIGNIFICANCE FOR A DIFFERENCE BETWEEN ARITHMETIC MEANS DERIVED FROM NON-CORRELATED SAMPLES

Tests of Significance for mean differences play an essential role in evaluating results in many research studies in psychology and related fields. A Test of Significance for a difference between the means of independent samples is as follows, σ_{M_x} being obtained by Formula 13:5a.

[14 : 8]

$$T = \frac{(M_x - M_y) - 0}{\sigma_{(M_x - M_y)}} = \frac{(M_x - M_y)}{\sqrt{\sigma_{M_x}^2 + \sigma_{M_y}^2}}$$

Test of Significance with the standard error of the difference between arithmetic means derived from uncorrelated samples

The following example, with data from Klineberg,* illustrates a Test of Significance for differences between means obtained from *independent* samples. Hence, the third term of the general formula (14:1) is zero. Klineberg used a non-language intelligence test in order to study the intelligence of various European groups. All his subjects were boys ranging in age from 10 to 12 years. He obtained 10 samples of 100 boys in three European cities and seven rural areas, making a total of 1000 subjects tested. In all cases, the mean intelligence scores of the urban groups were greater than those of the rural groups.

* O. Klineberg, "A Study of Psychological Differences Between Racial and National Groups in Europe," *Archives of Psychology*, **20**:1–58, 1931.

His results for the combined three urban groups and the combined seven rural groups were as follows:

Groups	N_s	Mean Intelligence Test Score	Standard Deviation
City	300	215.7	45.1
Rural	700	187.1	50.9

The difference between the mean intelligence test scores is 28.6. The question of whether this difference is likely on the basis of chance or whether it may be ascribed, at least in part, to non-chance factors can be answered by a Test of Significance, as follows:

$$T = \frac{(215.7 - 187.1) - 0}{\sqrt{\left(\frac{45.1}{\sqrt{300}}\right)^2 + \left(\frac{50.9}{\sqrt{700}}\right)^2}} = \frac{28.6}{\sqrt{(2.60)^2 + (1.92)^2}} = \frac{28.6}{3.23} = 8.9$$

Since the T ratio shows that the mean difference is 8.9 times the standard error of the difference, the hypothesis that the parameter difference is zero can be rejected with confidence. That is, the difference between the mean intelligence test scores of these rural and urban groups cannot be attributed merely to chance factors of sampling and measurement. Extra-chance factors operated to produce at least some of it. What these factors were is not a statistical but a research problem which Klineberg considers in his analysis of the results.

Fisher's Null Hypothesis for Differences *

It was indicated earlier that Fisher's development and use of the concept *null hypothesis* are somewhat more specific than our more generalized use of the concept. We shall illustrate his method for the hypothesis of a zero difference between means of samples.

Fisher interprets the null hypothesis as asserting that the statistics of the samples under consideration are derived from the same universe, and hence the parameter difference will be zero. The Test of Significance is therefore made in order to determine whether the data support or nullify the hypothesis. However, in our earlier examples we based the standard errors of differences on the variances of each statistic, as in classical statistics, whereas Fisher bases his on an estimate of the variance of the universe, derived from the average of the standard deviations of the sample results.

When there are two or more samples, the average of their respective standard deviations taken in reference to their own means is as follows:

* R. A. Fisher, *Statistical Methods for Research Workers*, Oliver & Boyd, London, 7th ed., 1938, pp. 128 ff.

[14 : 9]

$$\sigma = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2 + \cdots + \Sigma x_n^2}{N_1 + N_2 + \cdots + N_n}}$$

Average of standard deviations for two or more samples, with deviations of each taken from their respective means

And the standard error of the difference between the means of any two samples in the same universe is (when the size of each sample is greater than 30, and therefore N instead of $N - 1$ is used):

[14 : 10]

$$\sigma_{(M_1 - M_2)} = \sqrt{\left(\frac{\sigma}{\sqrt{N_1}}\right)^2 + \left(\frac{\sigma}{\sqrt{N_2}}\right)^2}$$

Standard error of the difference between means, when the variance of the sampling distribution is based on the average of both samples

But since σ is the estimate of σ_u (the standard deviation of the sampling distribution in the universe of the hypothesis), and therefore the variance of both samples is the same, σ can be removed from the radical:

[14 : 10a]

$$\sigma_{(M_1 - M_2)} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

Standard error of the difference between two means when the variance of both samples is the same

In using Klineberg's sample data to test Fisher's null hypothesis we assume that both his samples were drawn randomly from the same universe (rural-urban). The Test of Significance is made to determine whether or not this is the case. Σx^2 is not given in Klineberg's result, and we shall therefore have to determine it from σ and N. Since

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}}, \quad \Sigma x^2 = N\sigma^2$$

For the city sample,

$$\Sigma x_1^2 = 300(45.1)^2$$

and for the rural sample,

$$\Sigma x_2^2 = 700(50.9)^2$$

Therefore

$$\sigma_{s(1,2)} = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2}{N_1 + N_2}} = \sqrt{\frac{300(45.1)^2 + 700(50.9)^2}{300 + 700}} = \sqrt{\frac{2423770}{1000}} = 49.23$$

Hence the Test of Significance for Fisher's null hypothesis is:

$$T = \frac{(M_1 - M_2) - 0}{\sigma_{(M_1 - M_2)}} = \frac{(215.7 - 187.1) - 0}{49.23\sqrt{\frac{1}{300} + \frac{1}{700}}} = \frac{28.6}{3.40} = 8.4$$

With a T ratio of 8.4, the mean difference is of course considerably greater than would be expected on the basis of chance. The null hypothesis is therefore rejected with confidence, which means that the rural and urban samples are samples of two different universes with different parameter means, rather than samples from the same universe.

This Test of Fisher's null hypothesis gives no information beyond that derived from the preceding test, in which the variance of each sample was estimated separately. The above T ratio of 8.4 is slightly less because the average of the standard deviations of the two samples, $\sigma_{s_{(1,2)}} = 49.23$, is somewhat larger than the σ of the urban groups, which constituted only 3/10 of the total sample of 1000 cases.

F. TESTS OF SIGNIFICANCE FOR A MEAN DIFFERENCE BETWEEN CORRELATED SAMPLES

The following example illustrates a Test of Significance for mean differences when the means are obtained from samples which are not independent of each other. It also represents the type of analysis often required in determining whether the experimental variable in a psychological experiment has made any difference in the result.

An investigator wishes to determine whether systematic coaching affects intelligence test performance. He gives an intelligence test to a sample of 200 twelve-year-old boys drawn randomly from the city's school population. He then divides the total group into two groups of 100 each, and matches them pair by pair on the basis of their intelligence test scores. He uses one group of 100 as a control group (C), and the other group as an experimental group (E); the latter is coached systematically over a period of several weeks. At the end of this period he administers an alternative form of the intelligence test to both groups. The results for his experiment are as follows:

Matched Groups		Intelligence Test Score Results			
	N_s	At Beginning of Experiment		At End of Experiment	
		Mean	σ	Mean	σ
Control group (C)	100	95	12	97	13
Experimental group (E)	100	95	12	105	15

The correlation between the intelligence test scores of the two groups, matched pair by pair, is .60.

Since the two groups were matched, there should be little or no difference in their initial mean scores or in the variability of their initial performance. That this was actually the case is indicated by the mean of 95 and the standard deviation of 12 for each group at the beginning of the experiment.

That coaching may have had a real effect on intelligence test performance is suggested by the final results. The mean score for the experimental group is now 105, as against 97 for the control group, a difference of 8 points. The question is whether such a difference is likely to occur by chance or whether the null hypothesis can definitely be rejected. If it can be, then in view of the experimental design of the investigation, we are warranted in concluding that coaching has a definite positive effect on the intelligence test performance of 12-year-old boys.

The standard error of the difference between the means of matched groups is:

[14 : 11]

$$\sigma_{(M_C - M_E)} = \sqrt{\sigma_{M_C}^2 + \sigma_{M_E}^2 - 2r_{M_C M_E}\sigma_{M_C}\sigma_{M_E}}$$

Standard error of the difference between means derived from correlated samples

The correlation between the means of bi-variates is the same as the correlation between the variates themselves. Hence, the above formula may be restated as follows:

$$\sigma_{(M_C - M_E)} = \sqrt{\sigma_{M_C}^2 + \sigma_{M_E}^2 - 2r_{CE}\sigma_{M_C}\sigma_{M_E}} \qquad [14 : 11a]$$

The Test of Significance which will enable us to answer the research question is therefore as follows:

$$T = \frac{(M_E - M_C) - 0}{\sqrt{\sigma_{M_C}^2 + \sigma_{M_E}^2 - 2r_{CE}\sigma_{M_C}\sigma_{M_E}}}$$

$$= \frac{(105 - 97) - 0}{\sqrt{\left(\frac{13}{\sqrt{100}}\right)^2 + \left(\frac{15}{\sqrt{100}}\right)^2 - 2(.60)\frac{13}{\sqrt{100}}\frac{15}{\sqrt{100}}}} = \frac{8}{\sqrt{1.69 + 2.25 - 2.34}} = \frac{8}{1.26} = 6.3$$

where 1.26 is the standard error of the difference between the means of the two groups.

Since the T ratio is 6.3, we can be confident that the mean difference between the intelligence test performance of the control and the coached groups at the end of the experiment is unlikely on the basis of chance. Hence, we can reject the null hypothesis with confidence, and we are warranted in concluding that coaching has a real effect upon the intelligence test performance of boys at this age level.

It should be observed that the positive correlation of .60 between the intelligence test performance of the matched pairs in the two groups served to reduce the estimate of the standard error of the difference. Had this correlation been zero, the standard error of the difference would have been:

$$\sqrt{1.69 + 2.25} = 1.98$$

Although this standard error is considerably larger than 1.26, the T ratio (8/1.98 = 4.0) given by it would still be large enough to warrant the rejec-

tion of the null hypothesis. Therefore, if it had not been convenient to compute the correlation coefficient, the Test of Significance without the third term of the standard error formula would still have given a T ratio large enough to warrant the rejection of the null hypothesis. However, if there is any likelihood that the correlation between matched samples will be negative rather than positive, it should be computed and the third term of Formula 14:11a should be used, since a negative correlation serves to increase rather than decrease the size of the standard error.

Effect of Heterogeneity of "Matched Samples"

In the preceding experiment the subjects were matched pair by pair on the basis of initial intelligence test performance. Two additional factors were also *controlled*, viz., age and sex, by virtue of the restriction of the samples to 12-year-old boys. Had the samples been heterogeneous in age and sex, variability in these factors might in themselves have accounted in part for the experimental result. That is, uncontrolled differences in age and sex might have been partly responsible for the higher performance of the coached group.

Therefore such factors should be controlled either (1) by setting up control and experimental groups that are relatively homogeneous, or (2) by matching the two groups on such factors as well as on the behavior to be studied (in this case, intelligence test performance). If neither of these procedures is used, a measure of control can be introduced by analyzing statistically the possible role of variability or heterogeneity within the groups as well as between them.*

G. TESTS OF SIGNIFICANCE FOR A DIFFERENCE BETWEEN STANDARD DEVIATIONS

The comparison of differences in variability between two or more groups is often involved in a research problem. For example, in the study of sex differences, tradition held that there is greater variability among women than among men. No broad generalization, however, is warranted prior to obtaining empirical evidence on this problem. Furthermore, it is necessary to obtain empirical data for particular traits and attributes and to bring together the results of many investigations before broad generalizations about sex differences in variability are justified.

The analysis of differences in variability is also important in other types of research investigations, as in experiments in which the experimental variable may definitely affect the variability of the sample result. It is often relevant to determine whether a difference in variability between experi-

* Cf., in this regard, Eugene Shen, "The Place of Individual Differences in Experimentation," chap. 14 in Quinn McNemar and Maud A. Merrill (eds.), *Studies in Personality*, McGraw-Hill, New York, 1942.

mental and control groups can be attributed merely to chance or whether the difference is significantly greater than would be expected on the basis of chance.

The formula for the standard error of a difference between standard deviations is as follows:

[14 : 12]

$$\sigma_{(\sigma_x-\sigma_y)} = \sqrt{\sigma_{\sigma_x}^2 + \sigma_{\sigma_y}^2 - 2r_{\sigma_x\sigma_y}\sigma_x\sigma_y}$$

Standard error of the difference between standard deviations

where $\sigma_{\sigma_x}^2$ is the estimated variance of the sampling distribution of the standard deviation of the x variable; $\sigma_{\sigma_y}^2$ is a similar measure for the y variable; $r_{\sigma_x\sigma_y}$ is the correlation between the standard deviations of the two variables; σ_{σ_x} is the estimated standard error of the standard deviation of the x variable; and σ_{σ_y} is the estimated standard error of the standard deviation of the y variable. (See Formula 13:7.)

The correlation coefficient for the standard deviations of two variables can be directly estimated from the correlation of the bi-variates. It is equal to the square of the correlation coefficient, r_{xy}. The preceding formula thus becomes:

$$\sigma_{(\sigma_x-\sigma_y)} = \sqrt{\sigma_{\sigma_x}^2 + \sigma_{\sigma_y}^2 - 2r_{xy}^2\sigma_{\sigma_x}\sigma_{\sigma_y}} \qquad [14 : 12a]$$

If the standard deviations are derived from independent samples, the correlation is zero and the third term of the above formula becomes zero.

A Test of Significance for the difference in the variability of two groups will be illustrated with David Wechsler's data on the Information subtest of the Wechsler-Bellevue Scale for measuring intelligence.* Wechsler gives the means and standard deviations on each subtest for successive age groups from 7 to 60 years of age. The variability of performance on the Information subtest *tends* to increase with age.

We shall compare the variability of two age groups, viz., from 25 through 29 years and from 40 through 44 years. The mean score of both these groups is the same, 10.1 points. But the standard deviation of the younger group is 2.98, whereas for the older group it is 3.70. Is this difference of 0.72 points reliable, i.e., significantly greater than zero? The following Test of Significance answers this question, the two age groups constituting independent (non-correlated) samples:

$$\frac{(\sigma_x - \sigma_y) - 0}{\sigma_{(\sigma_x-\sigma_y)}} = \frac{(\sigma_x - \sigma_y) - 0}{\sqrt{\left(\frac{\sigma_x}{\sqrt{2N_x}}\right)^2 + \left(\frac{\sigma_y}{\sqrt{2N_y}}\right)^2}} = \frac{(3.70 - 2.98) - 0}{\sqrt{\left(\frac{3.70}{\sqrt{2(75)}}\right)^2 + \left(\frac{2.98}{\sqrt{2(125)}}\right)^2}}$$

* David Wechsler, *The Measurement of Adult Intelligence*, Williams & Wilkins, Baltimore, 3rd ed., 1944. (Data from Table 39, p. 222.)

where N_x, the size of the older sample, is 75, and N_y, the size of the younger sample, is 125. T is equal to 2.0:

$$T = \frac{0.72}{\sqrt{(.3021)^2 + (.1885)^2}} = \frac{0.72}{\sqrt{.1268}} = \frac{0.72}{.36} = 2.0$$

The difference between the variabilities of the two groups is twice as large as the standard error of the difference. The P value for a T ratio of 2.0 or more, where only one tail of the sampling distribution of differences is concerned (as in Fig. 14:1), is approximately .02 (equivalent to the 2% confidence level; cf. Table II, Appendix B). Thus, for the hypothesis that the difference is zero, the probabilities are 2 in 100 that a difference as large as 0.72 in a sample result may be due to chance factors of sampling and measurement.

If we are satisfied with these odds as a basis for rejecting the null hypothesis, we will conclude that there is a real difference in variability in the younger and older groups on the Information test of the Bellevue-Wechsler Scale. Certainly, when $T = 2.0$, we cannot accept the null hypothesis as likely. But if we wish to be cautious, we shall tentatively reject the null hypothesis and conclude that there is a real difference only if further samples of test scores from these two age groups support this generalization.

On the other hand, Wechsler's data for all the age groups above 25 to 29 years indicate that the T ratio of 2.0 is sufficiently large to warrant rejection of the null hypothesis. The variability of their scores on the Information test is shown in Table 14:1. These data give a compelling reason for accepting a T ratio of 2.0 as a satisfactory criterion because additional independent samples of older age groups all yield measures of variability larger than that of the 25 to 29 age group.

Table 14:1. Differences Between Successive Age Groups in Results on the Information Test of the Bellevue-Wechsler Scale of Intelligence *

Age Group	N	M	σ
25–29	125	10.1	2.98
30–34	110	9.8	3.12
35–39	100	9.8	3.37
40–44	75	10.1	3.70
45–49	60	9.5	3.21
50–54	45	9.6	4.08
55–59	36	9.5	3.86

* David Wechsler, *The Measurement of Adult Intelligence*, Williams & Wilkins. Baltimore, 3rd ed., 1944. (Data from Table 39, p. 222.)

Combining the Results of Several Groups for a Test of Significance

To determine the T ratio for the difference in variability in several groups, we shall combine the results for the three age groups from 40 to 54.

The standard deviation of a single group result is $\sqrt{\Sigma x^2/N}$, where the devia-

tions, x, are taken from the mean of the sample. If the several groups to be combined have different means, the deviations of each sample must be taken from the weighted mean of the combined result. We shall compute this mean first for the three age groups, 40 to 44, 45 to 49, and 50 to 54. The weighted mean for two or more samples is given by the following formula:

$$M_c = \frac{N_1M_1 + N_2M_2 + \cdots + N_nM_n}{N_1 + N_2 + \cdots + N_n} \qquad [14:13]$$

Weighted mean of two or more groups combined

where the subscripts 1, 2, . . . n identify the several groups to be combined (c). Therefore, for the three age groups whose N's and M's are given in Table 14:1, we have:

$$M_c = \frac{75(10.1) + 60(9.5) + 45(9.6)}{75 + 60 + 45} = \frac{1759.5}{180} = 9.8$$

The standard deviation of combined groups, determined from the respective standard deviations of each, is as follows:

$$\sigma_c = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + \cdots + N_n\sigma_n^2 + N_1(M_1 - M_c)^2 + N_2(M_2 - M_c)^2 + \cdots + N_n(M_n - M_c)^2}{N_1 + N_2 + \cdots + N_n}} \qquad [14:14]$$

Standard deviation of two or more combined groups, with deviations taken from the weighted mean of the combination

where the subscripts 1, 2, . . . n identify the groups whose measures are to be combined; and M_c is the weighted mean of the combined samples (from Formula 14:13). For the three age groups, the standard deviation of the combined result is 3.66:

$$\sigma_c = \sqrt{\frac{75(3.70^2) + 60(3.21^2) + 45(4.08^2) + 75[(10.1 - 9.8)^2] + 60[(9.5 - 9.8)^2] + 45[(9.6 - 9.8)^2]}{75 + 60 + 45}}$$

$$= \sqrt{\frac{2407.95}{180}} = \sqrt{13.3775} = 3.66$$

We now have the standard deviation of the three age groups combined and can test the significance of the difference in variability in these three age groups and in the younger age group (25 to 29). The Test of Significance is as follows:

$$T = \frac{(\sigma_c - \sigma_y) - 0}{\sigma_{(\sigma_c - \sigma_y)}} = \frac{(3.66 - 2.98) - 0}{\sqrt{\left(\frac{3.66}{\sqrt{2(180)}}\right)^2 + \left(\frac{2.98}{\sqrt{2(125)}}\right)^2}}$$

$$= \frac{0.68}{\sqrt{(.1929)^2 + (.1885)^2}} = \frac{.68}{.27} = 2.5+$$

The T ratio, 2.5+, confirms what was suggested by the data in Table 14:1, viz., that the variability in the scores on the Information test is significantly greater for the older age groups than for the younger one. There is less than 1 chance in 100 that the difference in variability of 0.68 would occur in a universe whose parameter difference is zero. Hence, we can reject the null hypothesis with confidence and conclude that the variability of the test results of the older age groups is somewhat greater than that of the younger group.

H. TESTS OF SIGNIFICANCE FOR A DIFFERENCE BETWEEN COEFFICIENTS OF RELATIVE VARIATION

We saw earlier, in presenting measures of variability for descriptive statistics (Chapter 7), that a comparison of the variability of two or more groups is sometimes misleading if it is made directly in terms of their standard deviations, especially if their means differ considerably. We saw further that when the variability of two or more distributions derived from the same scale or type of measure is compared, Pearson's Coefficient of *Relative* Variation (V) can be used to avoid any misleading implications that might arise from the direct comparison of the standard deviations themselves. V expressed as a percentage is $(100)\sigma_x/M_x$.

Since occasions do arise in which comparisons of *relative variability* are required, we shall give the standard error of V and present a Test of Significance for the difference between two Coefficients of Relative Variation. The standard error of V is given by the following:

$$\sigma_V = \frac{V}{\sqrt{2N_s}}\sqrt{1 + 2\left(\frac{V}{100}\right)^2} \qquad [14:15]$$

Standard error of the Coefficient of Relative Variation

where V is the Coefficient of Relative Variation and N_s is the size of the sample.

By Formula 14:1, the formula for the standard error of a difference between two Coefficients of Relative Variation is as follows:

$$\sigma_{(V_x - V_y)} = \sqrt{\sigma_{V_x}^2 + \sigma_{V_y}^2 - 2r_{V_xV_y}\sigma_{V_x}\sigma_{V_y}} \qquad [14:16]$$

Standard error of the difference between Coefficients of Relative Variation

For independent (non-correlated) samples, the third term of this formula becomes zero. In the following problem, based on a further comparison of some of Wechsler's Information test data, $\sigma_{(V_x - V_y)}$ will therefore be equal to:

$$\sigma_{(V_x - V_y)} = \sqrt{\sigma_{V_x}^2 + \sigma_{V_y}^2}$$
$$= \sqrt{\left[\frac{V_x}{\sqrt{2N_x}}\sqrt{1+2\left(\frac{V_x}{100}\right)^2}\right]^2 + \left[\frac{V_y}{\sqrt{2N_y}}\sqrt{1+2\left(\frac{V_y}{100}\right)^2}\right]^2} \qquad [14:17]$$

Standard error of the difference between Coefficients of Relative Variation derived from non-correlated samples

Wechsler reports the standard deviation variability of 7-year-olds on the Information subtest as equal to 1.11. This σ is less than half that of the 25 to 29 age group, which was given in Table 14:1 as 2.98. However, it would be misleading to work directly with these standard deviations because the mean score of the 7-year-olds is only one-fourth as large as that of the older group: $M_x = 2.5$ and $M_y = 10.1$. The Test of Significance must therefore be made in terms of their respective Coefficients of Variation, V, which will take into account these differences between means. Thus:

$$V_x = \frac{1.11}{2.5}(100) = 44.4\% \qquad \text{(7-year-olds)}$$

$$V_y = \frac{2.98}{10.1}(100) = 29.5\% \qquad \text{(25 to 29 age group)}$$

The standard deviation of the 7-year-olds is nearly 45% as large as its mean, whereas for the older age group it is only 30% as large as its mean. The difference in relative variability is 44.4% − 29.5% = 14.9. Is this difference significantly greater than zero?

The Test of Significance is as follows, N_x being 50 and N_y being 125:

$$T = \frac{(V_x - V_y) - 0}{\sqrt{\sigma_{V_x}^2 + \sigma_{V_y}^2}} = \frac{(44.4 - 29.5) - 0}{\sqrt{\left[\frac{44.4}{\sqrt{2(50)}}\sqrt{1 + 2\left(\frac{44.4}{100}\right)^2}\right]^2 + \left[\frac{29.5}{\sqrt{2(125)}}\sqrt{1 + 2\left(\frac{29.5}{100}\right)^2}\right]^2}}$$

$$= \frac{14.9}{\sqrt{(5.244)^2 + (2.023)^2}} = \frac{14.9}{5.6} = 2.7$$

The T ratio is 2.7 and consequently the difference in the relative variability of Information test scores of the 7-year-olds and the 25 to 29 age group is significantly greater than zero. We can reject the null hypothesis with confidence and conclude that the relative variability of the older age group is smaller than that of the younger. Taken in conjunction with the Tests of Significance in Section G, the results do not support the hypothesis that variability on the Information test of the Wechsler-Bellevue Scale increases with age.

I. TESTS OF SIGNIFICANCE FOR A DIFFERENCE BETWEEN PRODUCT-MOMENT COEFFICIENTS OF CORRELATION

The standard error of a difference between correlation coefficients obtained from independent samples is relatively simple to compute because the third term of Formula 14:1 will be zero and the formula simplifies to the following:

[14:18]

$$\sigma_{(r_{12}-r_{34})} = \sqrt{\sigma_{r_{12}}^2 + \sigma_{r_{34}}^2} = \sqrt{\left(\frac{1 - r_{12}^2}{\sqrt{N_{12}}}\right)^2 + \left(\frac{1 - r_{34}^2}{\sqrt{N_{34}}}\right)^2}$$

Standard error of the difference between product-moment correlation coefficients, derived from non-correlated samples

where the subscript 12 designates the first two variables correlated and the subscript 34 the second two variables correlated. (If the second pair of variables includes either of the first two, the subscript will be 13 or 23, and there is a possible correlation between the two sample results.)

Formula 14:18 has the same limitations as the standard error of a correlation coefficient (Formula 13:16). As the value of r increases, the estimates of the standard error are increasingly unsatisfactory. We saw that Fisher's **z** function can be used to test the significance of and to establish confidence limits for high values of r. Similarly, the significance of differences involving high values of r can be tested in terms of **z**. The standard error for differences in **z** is equal to:

[14:19]

$$\sigma_{(z_{12}-z_{34})} = \sqrt{\sigma_{z_{12}}^2 + \sigma_{z_{34}}^2} = \sqrt{\frac{1}{N_{12}-3} + \frac{1}{N_{34}-3}}$$

Standard error of the difference between **z**'s derived from uncorrelated samples

where the first term of the formula is the variance of the **z** function for the first two variables correlated, and the second term is the variance of this function for the second two variables correlated. The distribution of differences between **z** is normal, and hence a Test of Significance based upon this formula can be interpreted for samples of over 25 cases by means of the table of the normal probability integral (Table I, Appendix B).

Research that calls for a comparison of correlation coefficients is often based upon *dependent* rather than independent samples, and consequently the third term of the general formula for the standard error of a difference may be required. The exception to this for dependent samples arises when there is a negative correlation between the samples, and the T ratio is equal to or greater than 2.5. As pointed out previously, in such cases the third term of the general formula serves to increase the estimated standard error of the difference, and hence reduces the size of the test ratio.

The standard error of the difference between correlation coefficients derived from dependent samples differs,* depending on whether one array is or is not common to the two sets of bi-variates. In the first case:

[14:20]

$$\sigma_{(r_{12}-r_{13})} = \sqrt{\sigma_{r_{12}}^2 + \sigma_{r_{13}}^2 - 2r_{r_{12}r_{13}}\sigma_{r_{12}}\sigma_{r_{13}}}$$

Standard error of the difference between correlation coefficients derived from bi-variate samples with one array in common

If there is no common array:

[14:21]

$$\sigma_{(r_{12}-r_{34})} = \sqrt{\sigma_{r_{12}}^2 + \sigma_{r_{34}}^2 - 2r_{r_{12}r_{34}}\sigma_{r_{12}}\sigma_{r_{34}}}$$

Standard error of the difference between correlation coefficients derived from dependent samples but with no array in common

* Cf. C. C. Peters and W. R. Van Voorhis, *op. cit.*, pp. 195–199.

The correlation between the correlation coefficients for the third term of each of these formulas differs and is laborious to compute. Nor will a conversion to z help in this particular situation because no formula is available for estimating the correlation between z functions.

The attitudes of a group of 79 listeners toward a radio program were obtained in the Program Analyzer Laboratories of the Columbia Broadcasting System.* The two major parts of the program consisted of comedy dialogue and a group of songs. The listeners' attitudes, expressed during the program, were correlated with their responses to an opinion questionnaire administered at the end of the broadcast, in which the subjects were asked whether they would have turned the program off if they had been at home (unfavorable response) or would have listened to the end (favorable response). The following results were obtained: The correlation between questionnaire responses (variable 1) and attitudes toward the comedy dialogue (variable 2) was .74 (r_{12}), whereas the correlation between questionnaire responses (variable 1) and attitudes toward the songs (variable 3) was only .45 (r_{13}). These results suggest that the group's over-all opinion of the program was affected more by the comedy dialogue than by the songs. However, it is relevant to ascertain, by means of a Test of Significance, whether the difference between the two correlation coefficients is significant, or whether it might be expected on the basis of chance.

We shall first present a Test of Significance using the formula (14:20) which requires the correlation between attitudes toward the comedy dialogue (variable 2) and the songs (variable 3). This correlation, r_{23}, is .52. As indicated in Formula 14:20, for the Test of significance we shall need the correlation between the correlation coefficients, viz., $r_{(r_{12}r_{13})}$. This is equal to .50:

$$r_{(r_{12}r_{13})} = r_{23} - \frac{r_{12}r_{13}(1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{23}r_{12}r_{13})}{2(1 - r_{12}^2)(1 - r_{13}^2)}$$

$$= .52 - \frac{(.74)(.45)[1 - .52^2 - .74^2 - .45^2 + 2(.52)(.74)(.45)]}{2(1 - .74^2)(1 - .45^2)}$$

$$= .50$$

[14 : 22]
Correlation between correlation coefficients whose bi-variates have one array in common

The standard error of the difference between r_{12} and r_{13}, is therefore:

$$\sigma_{(r_{12}-r_{13})} = \sqrt{\left(\frac{1 - .74^2}{\sqrt{79}}\right)^2 + \left(\frac{1 - .45^2}{\sqrt{79}}\right)^2 - 2(.50)\frac{1 - .74^2}{\sqrt{79}}\frac{1 - .45^2}{\sqrt{79}}} = .077$$

And the Test of Significance for the null hypothesis, i.e., that the parameter difference is zero, is:

$$T = \frac{(r_{12} - r_{13}) - 0}{\sigma_{(r_{12}-r_{13})}} = \frac{(.74 - .45)}{.077} = \frac{.29}{.077} = 3.8$$

where .29 is the difference between the two correlation coefficients; zero is the parameter value of the hypothesis tested; .077 is the estimated value of the

* Cf. J. G. Peatman and T. Hallonquist, *op. cit.*

standard error of the difference between the two correlation coefficients; and 3.8 is the test ratio.

The Test of Significance for the difference between these two correlation coefficients yields a T ratio greater than 2.5 or 3.0. Hence, we can be confident that, since the difference is significantly greater than zero, it is not merely a chance difference. However, the statistical results of a Test of Significance do not in themselves indicate the nature of or the reasons for the relationship. When we can definitely reject the null hypothesis, as in this case, we examine the character of possible extra-chance factors responsible for the difference. Any inferences of causal relations must be based on an analysis of the character of the data correlated. In this case, knowledge of the nature and sequence of events, together with interviews with the subjects, warrants the inference that the subjects' opinions of the program as a whole were determined more by their attitudes toward the comedy dialogue than by their attitudes toward the songs.

We shall now present a Test of Significance for these data and use only the variances of each coefficient (Formula 14:18), in order to see whether the labor involved in computing the third term of Formula 14:20 was necessary in this particular case. The estimate of the standard error of the difference between the two correlation coefficients is now as follows:

$$\sigma_{(r_{12}-r_{13})} = \sqrt{\left[\frac{(1-.74^2)}{\sqrt{79}}\right]^2 + \left[\frac{(1-.45^2)}{\sqrt{79}}\right]^2} = .103$$

The standard error is increased somewhat, from .077 to .103. Hence, the T ratio below differs from the T of 3.8 obtained with the complete standard error formula:

$$T = \frac{(r_{12} - r_{13}) - 0}{\sigma_{(r_{12}-r_{13})}} = \frac{.29}{.103} = 2.8$$

Thus, in this particular case the third term of Formula 14:20 yields a T ratio greater than 3.0, whereas the abbreviated formula gives a T ratio less than 3.0 but greater than 2.5. The null hypothesis can of course be more confidently rejected with a T ratio of 3.8 than with one of 2.8.

EXERCISES

1. What hypothesis is usually considered in a Test of Significance for a difference between two statistics? Why?
2. Under what circumstances is it appropriate to omit the third term of Formula 14:1?
3. Compare the logic underlying a Test of Significance for a statistic with that for a difference between two statistics.
4. Using the data in Table 5:14, determine the percentage of college freshmen with an intelligence test score above 75, the percentage of their best friends with an intelligence test score above 75, and the significance of the difference between these two percentages.

5. Using the same data, determine whether the difference in the means, variability, and relative variability of the freshmen and their best friends is significant for each of the following variables:
   a. average grades
   b. intelligence test scores
   c. ages

   (Note that the variables to be compared are not derived from independent samples.)
6. Using the data in Table 14:1, determine whether the difference between the means and between the standard deviations on the Bellevue-Wechsler scale is significant for the following age groups:
   a. 25 to 29 and the 35 to 39 age group
   b. 25 to 29 and the 50 to 59 age group
7. Using the same data, combine the results for the first three age groups (25 to 39) and determine whether the difference between the mean and standard deviation of this combined group is significantly different from the mean and standard deviation of the combined 40 to 59 age group.
8. Using the correlation coefficients obtained in Exercise 12, Chapter 9, for the data in Table 5:14, determine whether the difference in the correlations between average grades and intelligence test scores is significant.

CHAPTER 15

Chi-Square and Tests of Significance

Karl Pearson's chi-square technique is a statistical method for the testing of hypotheses concerning distributions of frequencies. Since *categorical data* consist basically of the data of frequencies, chi-square is especially useful in testing hypotheses about such data. However, it can be used generally to include classes of frequencies derived from variables.

The statistical hypotheses which can be tested by chi-square are many, the restrictions being mainly of two kinds. First, as already indicated, the hypotheses must concern *statistical frequencies* of categories or classes. Chi-square is not directly applicable to hypotheses involving other kinds of data or statistical measures, but it can be adapted to proportions or percentages. Second, it is usually not a reliable technique if the number of hypothetical frequencies for any class or category is less than 10.* It should also be emphasized that N_s, the size of the sample, from which the frequencies per class are derived, should be fairly large.† Finally, the technique is based on the assumption that the frequencies of each class are independent of each other.

Except for these limitations, chi-square provides a *general* technique of analysis. The number of statistical hypotheses that can be tested is limited only by the total number of frequencies in a sample. Thus, a group of empirical data can be tested for their possible divergence (on the basis of chance) from any hypothetical grouping of N frequencies, as long as there are no less than 10 hypothetical frequencies per class. If, for example, a random sample of 100 people consists of 60 men and 40 women, we can test the hypothesis that the division of frequencies in this sample is only a chance deviation from a hypothetical universe in which the sexes are evenly divided, i.e., 50% men and 50% women. With the chi-square technique we can obtain an estimate of the probability of a sample result of 40 or less women and 60 or more men for a universe in which the sexes are equally divided. We can also test the hypothesis that the universe is divided in the proportion of .75 men and .25

* G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, Charles Griffin & Co., London, 12th ed., 1940. According to these two authors (p. 422), "No theoretical cell frequency should be small . . . 5 should be regarded as the very minimum and 10 is better." R. A. Fisher likewise agrees that 5 is the minimum. (*Statistical Methods for Research Workers*, Oliver & Boyd, London, 7th ed., 1938, p. 87.)

† "N should be at least 50, however few the number of cells." (Yule and Kendall, *op. cit.*, p. 422.)

women, .65 women and .35 men, etc. In such cases, the proportionate values of each class are converted into frequencies on the basis of N_s, the size of the sample.

Regardless of the particular hypothesis tested, the chi-square technique consists in a comparison of the *chance* implications of the hypothesis with the sample result. If the sample result could be expected to occur in random samples of the hypothetical universe on the basis of chance alone, it is a chance implication of the hypothesis and cannot be rejected. On the other hand, if the sample result cannot be interpreted as a chance implication of the hypothesis, the latter can be rejected. As R. A. Fisher puts it, the facts are given the opportunity to disprove a hypothesis. If a hypothesis is disproved, the possible implications of this fact are considered.

A. CHI-SQUARE FOR THE DISTRIBUTION OF NON-VARIABLE AND VARIABLE ATTRIBUTES

Calculation of Chi-Square

The first step in computing chi-square is to establish for each sample category or class the number of frequencies which would be expected on the basis of the hypothesis to be tested. In other words, a relevant hypothesis stated in terms of frequencies per category must first be set up and then tested. Some hypotheses are, of course, more relevant than others. Very often the most relevant hypothesis is the *null* hypothesis of a purely *chance* distribution of frequencies into two or more classes. Thus, in the case of a coin assumed to be fair, a hypothetical distribution of frequencies for heads and tails is equally divided into two independent categories. In the case of a die, there would be six independent categories of events, and a chance distribution of the frequencies of each would be $1/6(N_s)$, where N_s is the size of the sample. Similarly, people's attitudes toward an event might be assumed, *on the basis of chance alone*, to be dichotomized into two categories, each equal in size.

What the chi-square technique does is to compare the divergence or deviation of the sample frequencies per category from the hypothetical frequencies for each category studied. The *greater* the difference between sample frequencies and hypothetical frequencies per category, the less the probability that the differences are attributable only to chance errors of sampling and measurement. Chi-square itself is a *measure* that expresses the extent of the differences between hypothetical and sample results. The value of chi-square for a given hypothesis having been computed, the probability of differences as great as those between the sample frequencies per category or class and the hypothetical frequencies per category or class occurring on the basis of chance alone can then be estimated. In the light of this probability value, the hypothesis can then be rejected or not rejected, depending on the confidence

criteria used in judging the character of the result, i.e., whether or not it is a *likely* or *unlikely* result for the hypothesis.

Once the hypothetical frequencies for a given problem are set up, there remain the following relatively simple steps in computing chi-square:

1. The *difference* between the hypothetical and the sample frequencies is determined for each category or class.
2. Each of the differences per category or class is squared.
3. The *ratio* of the resulting squares to the hypothetical frequency per category or class is obtained.
4. These ratios are added to give the value of chi-square for the hypothesis tested. Thus, by formula:

$$\chi^2 = \sum \left[\frac{(f_s - f_h)^2}{f_h} \right] \qquad \begin{matrix}[15:1] \\ \text{Chi-square}\end{matrix}$$

where f_s is the number of sample frequencies per category or class; f_h is the hypothetical frequencies for corresponding categories or classes; and Σ symbolizes the process of summing all the ratios for the categories or classes under consideration.

A Chi-Square Test of Significance of Consumers' Brand Preferences (a Dichotomy)

A random sample of interviews with 1000 housewives gives the following results:

Housewives' preferences for Brand A = 600 (f_{s_A})
Housewives' preferences for Brand B = 400 (f_{s_B})

A relevant hypothesis here is a chance distribution of housewives' preferences into two categories, viz., 50% for Brand A and 50% for Brand B. Since a hypothesis for a chi-square test is stated directly in terms of frequencies, the division of hypothetical frequencies would be as follows:

Brand A, 500 preferences (f_{h_A})
Brand B, 500 preferences (f_{h_B})

If this hypothesis can be rejected with confidence, the conclusion follows that the 600 preferences for Brand A are not merely a chance result but rather indicate that a majority (more than 50%) of all housewives of the universe sampled prefer that brand.

Chi-square for this hypothesis is computed as shown in Table 15:1. The chi-square value is found to be 40. However, is a chi-square value as great as 40 likely for the hypothesis tested, on the basis of chance alone? The greater the differences between sample and hypothetical frequencies, the greater the value of χ^2 and the less the likelihood of their chance occurrence.

Table 15:1. Computation of Chi-Square for a Test of Significance of Consumer's Brand Preferences

	Sample of Consumers' Brand Preferences (f_s)	Frequencies by Hypothesis (f_h)	Differences * $f_s - f_h$	Differences Squared $(f_s - f_h)^2$	Chi-Square Ratio $\frac{(f_s - f_h)^2}{f_h}$
Brand A	600	500	100	10,000	20
Brand B	400	500	100	10,000	20
	$N_s = 1000$	1000			$\chi^2 = 40$

From Table 15:2:
For 1 *d.f.*, $P = .001$, for $\chi^2 \geqq 10.83$

* The signs of differences, $(f_s - f_h)$, can be neglected because the differences are squared.

The Probability of Chi-Square

Sampling distributions of the chi-square statistic are not of the normal, bell-shaped type unless there are around 30 classes or categories. For most Tests of Significance in terms of chi-square, the number of classes or categories is considerably less—often only 2, as in Table 15:1. The form of the sampling distribution varies considerably for 2, 3, 4, etc., classes up to 30. For 2 degrees of freedom, the sampling distribution is a curve like that shown in Fig. 15:1, in which the ordinate represents the frequency of the sample results and the abscissa represents the value of x^2. When there are 2 degrees of freedom, the form of the sampling distribution is like that in Fig. 15:2, which is similar to a dichotomized *half* of the standard, normal probability curve.

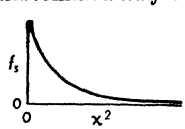


Fig. 15:1.

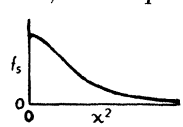


Fig. 15:2.

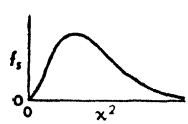


Fig. 15:3.

The mode of both these sampling distributions is at a χ^2 value of zero. But when there are 3 degrees of freedom, the mode shifts from zero and the curve is extremely skewed, as in Fig. 15:3. As the number of classes or categories increases, the form of the sampling distribution gradually approaches the normal, bell-shaped curve.

What we need for chi-square, therefore, is probability values for categories or classes ranging from 2 to 30. Beyond 30, the implications of the normal probability curve can be utilized. The probability values required are presented in Table 15:2.† This table is set up differently from Table II, Ap-

† R. A. Fisher and F. Yates, *Statistical Tables for Biological, Agricultural, and Medical Research*, Oliver & Boyd, London, 1938, Table IV, p. 27.

pendix B, for T of large sample theory, in that the body of the table consists of chi-square rather than P values, the latter being given for 11 values of P, as indicated at the head of the columns. Furthermore, these probability values are developed in terms of *degrees of freedom*,* rather than of the total sample frequencies, N_s.

Degrees of Freedom (*d.f.*)

The concept of *degrees of freedom* is analogous in its implications to N_s. However, in the case of chi-square the P values in Table 15:2 were originally developed not for N_s but in terms of the *number of classes or categories* for which frequency values based on the *hypothesis* could be *freely assigned.*†

The number of *degrees of freedom* (*d.f.*) for any hypothesis is equal to the number of categories or classes for which hypothetical frequency values can be freely assigned. This means that the number of degrees of freedom is equal to the total number of categories or classes *minus* the number of constraints imposed upon the data in establishing the hypothetical frequencies. In Table 15:1 there was one constraint, viz., that the sum of the hypothetical frequencies be equal to N_s, the size of the sample. Thus, as soon as a hypothetical value of 500 was set up for Brand A, the number of frequencies for Brand B had to be taken as 500, because the total number of frequencies (500 and 500) must equal 1000, the number of the observations in the sample. Since there were 2 categories, the number of *degrees of freedom* is therefore 1.

Since this type of constraint always enters into the determination of the hypothetical frequencies for one class or category in any problem, *d.f.* is always at least *one* less than the total number of classes or categories. We shall see later that additional constraints are sometimes imposed in setting up hypothetical frequencies. For each additional limitation, an additional degree of freedom is lost. Therefore,

$d.f.$ = n classes or categories minus n constraints imposed by the hypothesis

Table 15:2 gives the distribution of chi-square values for degrees of freedom from 1 to 30, and for 11 P values. In the first row of the body of the table are the chi-square values to be expected for sampling distributions with 1 degree of freedom. When $d.f. = 1$, the probabilities are at least 99 in 100 ($P = .99^+$) of obtaining, on the basis of random sampling, a chi-square value equal to or greater than .00. According to the last column of the first row, when $d.f. = 1$ the probabilities are only 1 in 1000 ($P = .001$) of obtaining a sample result in which chi-square is equal to or greater than 10.83.

---

* H. M. Walker, "Degrees of Freedom," *Journal of Educational Psychology*, 31:253–269, 1940.

† Cf. Karl Pearson, *Tables for Statisticians and Biometricians*, Cambridge University Press, Cambridge, 1914, pp. xxxi–xxxiii, 26–28.

Table 15:2. Distribution of Chi-Square *

Probability Values for Chi-Square with Degrees of Freedom from 1 to 30

d.f.	Probability: P										
	.99	.95	.90	.50	.30	.20	.10	.05	.02	.01	.001
1	.00	.00	.02	.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	.02	.10	.21	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	.12	.35	.58	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	.30	.71	1.06	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	.55	1.14	1.61	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	.87	1.64	2.20	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	2.17	2.83	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.73	3.49	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	3.32	4.17	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.94	4.86	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	4.58	5.58	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	5.23	6.30	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	5.89	7.04	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	6.57	7.79	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	7.26	8.55	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	7.96	9.31	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.25
17	6.41	8.67	10.08	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.79
18	7.02	9.39	10.86	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	10.12	11.65	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	10.85	12.44	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	11.59	13.24	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	9.54	12.34	14.04	21.34	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	13.09	14.85	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	13.85	15.66	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	14.61	16.47	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	15.38	17.29	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	16.15	18.11	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	16.93	18.94	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	17.72	19.77	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	18.49	20.60	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

* Table 15:2 is abridged from Table IV of Fisher: *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver & Boyd, Ltd., Edinburgh, by permission of the author and Publishers.

The value of chi-square in Table 15:1 was 40. If Table 15:2 were extended sufficiently, we would see that the probabilities for a chi-square value as great as 40 are very small. However, as indicated in the table, there is only 1 chance in 1000 of obtaining chi-squares equal to or greater than 10.82 when *d.f.* = 1. But this is the 0.1% *confidence level* developed in Chapter 12. Hence we can reject the hypothesis and conclude that it is *likely* that a majority of the universe of housewives prefer Brand A.

Chi-Square as a Test of Significance

We have seen that a Test of Significance yields a test ratio, T, as follows:

$$T = \frac{s - h}{\sigma_s}$$

where s is the sample value of a statistic; h is the parameter value by hypothesis; and σ_s is the standard deviation of the sampling distribution of the statistic. When the sampling distribution can be assumed to have the form of the standard, bell-shaped normal probability curve, the P value of the T ratio can be obtained directly from the table of normal probability values (Table II, Appendix B).

The value of χ^2 is analogous in its logical implications to a test ratio; that is, the chi-square statistic is in itself a Test of Significance. Thus

$$\chi^2 = \text{sum}\left(\frac{\text{square of difference of sample and theoretical frequencies}}{\text{theoretical frequencies}}\right)$$

Instead of the sampling distribution of χ^2 being measured in terms of σ_s, as in the case of T, it is set up in terms of frequencies. Just as a T ratio of 2.0 indicates a sample result 2 standard deviations above the parameter mean of the sampling distribution, so χ^2 is a measure of the distance on the abscissa of the sampling distribution. Chi-square itself measures the difference between a χ^2 of zero (which signifies no difference between the sample result and the hypothesis) and the χ^2 value of the sample result. Because the form of the sampling distribution of chi-square varies with different degrees of freedom from 1 to 30, the probability values of a given value of χ^2 are set up for various P values and confidence levels, as shown in Table 15:2.

Chi-Square for d.f. > 30

We have said that when there are more than 30 degrees of freedom, the sampling distribution of chi-square is similar in form to the standard, normal probability curve. Fisher * has indicated that in such cases the expression, $\sqrt{2\chi^2} - \sqrt{2(d.f.) - 1}$, is distributed normally with a standard error of 1.0. Thus for a problem with 35 *d.f.* and χ^2 equal to 65, the Test of Significance is set up in terms of T for large sample theory, as follows:

[15 : 2]

Chi-square Test of Significance when the number of categories or classes is more than 30

$$T = \frac{s - h}{\sigma_s} = \frac{\left(\sqrt{2\chi^2} - \sqrt{2(d.f.) - 1}\right) - 0}{1.0}$$

$$= \sqrt{2(65)} - \sqrt{70 - 1} = 11.4 - 8.3 = 3.1$$

A T ratio of 3.1 (Table II, Appendix B) signifies a sample result that will occur less than 1 time in 1000 in random sampling. Hence such a result is extremely unlikely and the hypothesis can be rejected with confidence.

* R. A. Fisher, *op. cit.*, p. 85.

A Chi-Square Test of Significance for a Trichotomy

Chi-square is particularly useful for testing hypotheses concerning trichotomies for which the sample data cannot well be dichotomized and analyzed by a Test of Significance for a percentage or a proportion. Amen's data on pre-school children's responses to pictures in Table 2:1 are a case in point. If we assume that the data of her 4-year-old group consisted of a sample of 99 responses, we can employ chi-square to determine whether or not their distribution differed significantly from a purely chance division (1/3 of N_s for each category). The number of d.f.'s will be 2.

The sample and hypothetical frequencies per category are presented in Table 15:3, and the value of χ^2 is found to be 14.79. According to Table 15:2, a difference as great as this will occur in random sampling less than 1 time in 1000. Hence, the null (chance) hypothesis can be rejected, and it can be concluded that a plurality of the responses are of the outer activity type.

Table 15:3. Chi-Square for the Hypothesis of a Chance Division of Frequencies in a Trichotomy

Category of Response, Amen's Data	Sample Result f_s	Chance Hypothesis f_h	Differences $(f_s - f_h)$	Differences Squared $(f_s - f_h)^2$	Chi-Square Ratio $(f_s - f_h)^2/f_h$
Static form	23	33	10	100	3.03
Outer activity	51	33	18	324	9.82
Inner activity	25	33	8	64	1.94
	$N_s = 99$	99			$\chi^2 = 14.79$

From Table 15:2:
For 2 d.f., $P = .001$, for $\chi^2 \geqq 13.82$

A Chi-Square Test of Significance for the Distribution of a Variate

Tests of Significance for the skewness and kurtosis of uni-modal types of distributions were presented in Chapter 13. In effect, two of the fundamental properties of the normal, bell-shaped curve, rather than the distribution as a whole, were analyzed separately by the centile method. The parameter skewness of such a distribution is taken as zero, and the parameter kurtosis (when measured in terms of Q/D) as .263. Both of these tests, however, are approximations and do not take into account the differences from one class interval to the next between a sample distribution and the normal, bell-shaped distribution. It is possible by means of the chi-square statistic, however, to make a single Test of Significance for the *form* or character of a variate distribution as a whole. A distribution of any type can be set up by hypothesis, and the differences between sample frequencies and hypothetical

frequencies per class interval can be evaluated by a chi-square Test of Significance.

To illustrate this procedure, we shall use the distribution of test scores on page 435, for which measures of skewness and kurtosis were obtained by the centile method and then evaluated in terms of appropriate Tests of Significance. It will be recalled that the T ratio for skewness was 1.1 and for kurtosis 0.6. In the light of these two results the hypothesis that the test might have yielded a more normally distributed variate with larger samples of students was not rejected. By means of an analysis of the sample distribution with a chi-square Test of Significance, we shall analyze the divergence of the distribution itself (not simply its properties of kurtosis and skewness) from the normal, bell-shaped distribution.

The Calculation of Hypothetical Frequencies per Class Interval for a Normal Distribution

The first step in analyzing a frequency distribution with chi-square is to determine the number of frequencies for each class interval on the basis of the hypothesis to be tested. In other words, the hypothetical frequencies (f_h) for each category or class (class interval in this case) must be determined, as was done in Tables 15:1 and 15:3.

It is relatively easy to set up a normal distribution for a given number of hypothetical frequencies equal to N_s, the size of a sample, if the intervals are taken in z score units and only the frequencies at the mean and at the midpoints of successive z score intervals are to be obtained. Table I, Appendix B, which gives the ordinate values of a normal distribution whose total area is taken as unity, can be used in computing the number of frequencies at any x/σ point on the abscissa, once the number of frequencies at the mean is determined. The latter is equal to the following:

[15 : 3]

$$f_{v_M} = \frac{N_s}{\sigma'\sqrt{2\pi}} = \frac{N_s}{2.51\sigma'}$$

Number of frequencies at the mean of a finite distribution taken by hypothesis as normal

where N_s is the size of the sample, σ' is the standard deviation of the distribution *in unit step-intervals*, the constant π equals 3.1416, and $\sqrt{2\pi}$ is 2.51.

The mean of the distribution of 250 test scores is 90.66 (or 90.7) and its standard deviation is 5.70. Since the size of the original class intervals of the data is 2.0 units (the distribution is given in Table 15:4):

$$\sigma' = \sigma/i = 5.70/2.0 = 2.85$$

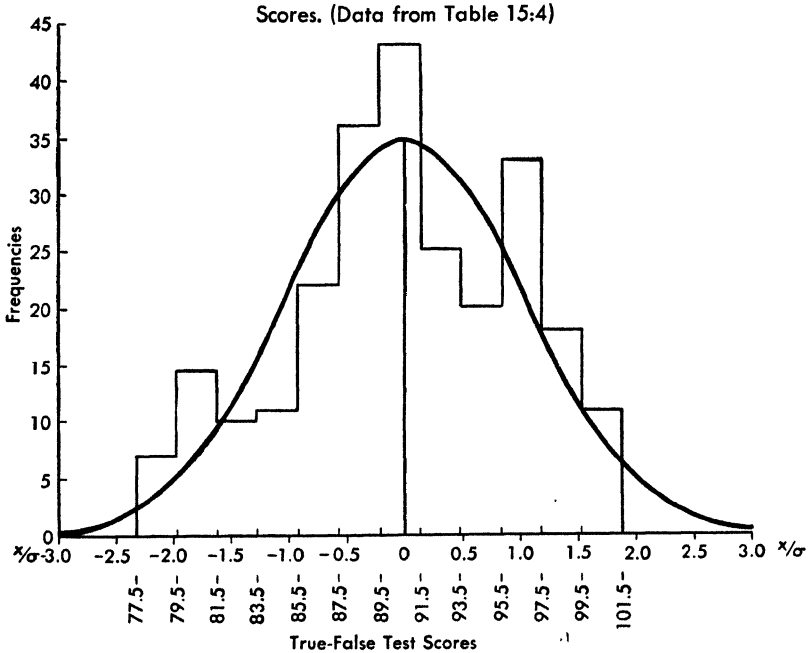
Hence the hypothetical number of frequencies at the mean is:

$$f_{v_M} = \frac{250}{2.85(2.51)} = 34.9$$

The number of frequencies at successive z score intervals above and below the mean can now be readily determined, since the fractional height of an ordinate at any point to the height at the mean is a fixed proportion (*see* Table IA, Appendix B, page 511). By means of this table, the frequencies at other points are found to be equal to the following:

y at mean = 34.9 frequencies
y at $\pm 0.5\sigma$ = .88(34.9) = 30.7 "
y at $\pm 1.0\sigma$ = .61(34.9) = 21.3 "
y at $\pm 1.5\sigma$ = .32(34.9) = 11.2 "
y at $\pm 2.0\sigma$ = .14(34.9) = 4.9 "
y at $\pm 2.5\sigma$ = .04(34.9) = 1.4 "
y at $\pm 3.0\sigma$ = .01(34.9) = 0.3 "

Fig. 15:4. The Normal Probability Curve Fitted to a Sample Distribution of Test Scores. (Data from Table 15:4)



These values were used to plot the normal curve in Fig. 15:4, which also portrays the distribution of the actual sample result.

Although the hypothetical, normal distribution itself can be readily obtained, we still do not have the hypothetical frequencies of the normal curve for the class intervals of the sample distribution. But it is these frequencies that we need for a chi-square Test of Significance. In order to obtain them we must (1) lay off the original score limits of the class intervals in terms of their

z score (x/σ) equivalents, (2) determine the proportion of the area in the normal distribution that lies within each interval, and (3) take these proportionate areas to a base of N_s so as to obtain the hypothetical frequencies of each class interval. The results are presented in Table 15:4.

Column 1 of this table lists the 14 classes (class intervals) of the test score variable. However, the last two are combined with the 12th because of the few hypothetical frequencies in these intervals. Columns 2 and 3 give the frequency distribution of the sample. The upper mathematical limits of each class interval are given in Column 4; these are the points in the distribution to be converted to z score equivalents.

Column 5 gives the deviation value, x, of the upper limit of each class interval. Thus for Class 1, the upper limit is 101.5 and its x value is $101.5 - 90.7 = 10.8$. Column 6 gives the x/σ or z score equivalents of the upper limits of each interval. The proportion of the area in each class interval can be obtained from Table I, Appendix B, which differentiates the area of the normal curve above and below the mean in x/σ or z score units. The values in Column 7 are read directly from that table. Thus, .4706 of the area lies between the mean and $x/\sigma = 1.89$. This is the upper limit of the sample distribution. The normal distribution, however, theoretically extends to infinity, and hence the total area of the upper half of the distribution, .5000, is also given.

The proportions of the total area within each class interval in Column 8 are obtained from Column 7. Thus, the area of the tail at the upper end of the distribution is equal to .0294. This is found by taking the difference between .5000 and .4706, the proportion of the area between the mean and 101.5. Similarly, the proportion of the area in Class Interval 1 is $.4706 - .4383 = .0323$; in Class Interval 2, $.4383 - .3830 = .0835$, etc. The area values in Column 7 must be added at the class interval in which the mean is located, because .0557 is the area between 91.5 and the mean of 90.7, and .0832 is the area between 89.5 and the mean. This sum is .1389, as shown in Column 8. The proportion of the area between the upper limit of Class Interval 12 and the mean is given in Column 7 as .4750. Hence .0250 of the total area lies below this point, i.e., between 79.5 and ∞.

The proportion of the area within each class interval in Column 8 is converted to frequencies by multiplying each proportion by N_s, the size of the sample. These hypothetical frequencies for the chi-square Test of Significance are given in Column 9. At the upper tail of the distribution, the frequencies (7.35) which lie beyond the limits of the sample distribution are combined with those of the highest class interval, and hence 15.43 is the number of hypothetical frequencies for Class Interval 1.

The Computation of Chi-Square

It is now possible to determine by chi-square whether the difference between the sample distribution of frequencies and a normal distribution of frequencies is or is not attributable to random errors of sampling and measurement.

Table 15:4. Distribution of Sample Frequencies of Test Scores, and Determination per Class Interval of the Hypothetical Frequencies for a Normal, Bell-Shaped Distribution

(1) Classes	(2) Mean = 90.7, σ = 5.70 Test Scores	(3) Frequencies per Class Interval f_s	(4) Upper Limit of Intervals	(5) Upper Limits Minus M x	(6) z Score Values x/σ	(7) Proportion of Area from M * (a)	(8) Proportion of Area Within Each Class	(9) Hypothetical Frequencies per Class f_h
					(To ∞)	.5000		
1	100–101	11	101.5	10.8	1.89	.4708	.0294	7.35 } 15.45
2	98–99	18	99.5	8.8	1.54	.4382	.0324	8.10 }
3	96–97	33	97.5	6.8	1.19	.3830	.0552	13.80
4	94–95	20	95.5	4.8	.84	.2996	.0834	20.85
5	92–93	25	93.5	2.8	.49	.1879	.1117	27.92
6	90–91	43	91.5	.8	.14	.0557	.1322	33.05
7	88–89	36	89.5	−1.2	−.21	.0832	.1389	34.72
8	86–87	22	87.5	−3.2	−.56	.2123	.1291	32.28
9	84–85	11	85.5	−5.2	−.91	.3186	.1063	26.58
10	82–83	10	83.5	−7.2	−1.26	.3962	.0776	19.40
11	80–81	14	81.5	−9.2	−1.61	.4463	.0501	12.52
12	78–79	2	79.5	−11.2	−1.96	.4750	.0287	7.18
13	76–77	4		−13.2	−2.31		.0250	6.25
14	74–75	1			(To ∞)	.5000		
		N = 250					Σ = 1.0000	250.00

* From Table I, Appendix B.

The chi-square analysis is given in Table 15:5, the computations being made in exactly the same way as in Tables 15:1 and 15:3. Chi-square is found to be 27.75. What is the P value for this result? To answer this question, the number of degrees of freedom must first be determined. This is either 9 or 11, depending upon the way in which the sampling of the universe is interpreted.

Table 15:5. Chi-Square Test of Significance for a Normal Distribution (Data from Table 15:4)

(1) Class	(2) Sample Frequencies f_s	(3) Hypothetical Frequencies f_h	(4) Differences $(f_s - f_h)$	(5) Differences Squared $(f_s - f_h)^2$	(6) Chi-Square Ratio $(f_s - f_h)^2/f_h$
(1)	11	15.45	4.45	19.80	1.28
(2)	18	13.80	4.20	17.64	1.28
(3)	33	20.85	12.15	147.62	7.08
(4)	20	27.90	7.90	62.41	2.24
(5)	25	33.05	8.05	64.80	1.96
(6)	43	34.72	8.28	68.56	1.97
(7)	36	32.28	3.72	13.83	.43
(8)	22	26.58	4.58	20.98	.79
(9)	11	19.40	8.40	70.56	3.64
(10)	10	12.52	2.52	6.35	.51
(11)	14	7.18	6.82	46.51	6.48
(12)	7	6.25	.75	.56	.09
	$N_s = 250$	$N_h = 250$			$\chi^2 = 27.75$

From Table 15:2:
For 11 *d.f.*, $P = .01$ for $\chi^2 \geqq 24.72$
For 9 *d.f.*, $P = .001$ for $\chi^2 \geqq 27.88$

If, on the one hand, the sample is considered as being drawn from a universe of test scores made by college students in an introductory psychology course, *d.f.* is equal to the number of class intervals (12) less 1; for at least one degree of freedom is lost by the constraint of N_s, the size of the sample. If, on the other hand, the sample is considered as being drawn from a restricted universe with the mean and standard deviation equal to those of the sample result, two additional constraints are imposed by these parameters, and *d.f.* = 9. In the present problem, either interpretation leads to the same conclusion, even though the former interpretation is generally the one intended.

Table 15:2 shows that the probabilities for 11 *d.f.* are only .01 (1 in 100) that χ^2 will be equal to or greater than 24.72, and only .001 (1 in 1000) that it will be equal to or greater than 31.26. According to the 1% confidence criterion, the hypothesis of a normal variate can be rejected; that is, the distribution of these test results cannot be considered as purely a random divergence from a normally distributed universe. This result, together with an inspection of Fig. 15:1, suggests either that the test itself was not properly designed to yield

an adequate differentiation among both the brighter and the less informed students, or that the variation of students' abilities may have been atypical; both factors can of course be present. However, the causes underlying the result cannot be answered on the basis of these statistical results alone.

It should be noted that the chi-square Test of Significance for the distribution in Fig. 15:1 yields a different result from the centile analysis in terms of skewness and kurtosis developed in Chapter 13 at the end of Section D. Although these two properties are often a sufficient expression of the normality or non-normality of a distribution, they cannot always be relied upon to give adequate Tests of Significance for a distribution both as a whole and in detail. The divergence between the frequencies of Class Intervals 3 and 11 (Table 15:5) and the hypothetical frequencies was not taken into account sufficiently by the centile measures of skewness and kurtosis. Hence, whenever the distribution of a large number of cases is erratic at points that do not materially affect C_{90}, C_{75}, C_{50}, C_{25}, and C_{10} (the centile values employed for measuring skewness and kurtosis), a Test of Significance for the hypothesis of a normally distributed variate can be more accurately set up in terms of chi-square.

B. CHI-SQUARE TESTS OF SIGNIFICANCE FOR THE INDEPENDENCE OF TWO ATTRIBUTES

Chi-square was employed in Tables 15:1, 15:3, and 15:5 to test hypotheses concerning the distribution of a single attribute or variable. The technique is also useful in testing hypotheses about *co-relationships* between two attributes or variables. These chi-square tests, which are usually referred to as *tests of independence* between two sets of sample data, are in effect Tests of Significance for the null hypothesis of *no correlation* between cross-tabulated attributes or bi-variates. A chi-square analysis will indicate whether the correlation is any greater than would be expected on the basis of chance.

Chi-Square Test of Significance for Correlation Between Dichotomized Attributes

Random samples of 100 men and 100 women are interviewed concerning their habits of listening to a radio program. Sixty-five of the men say they are *non-listeners* and 35 say they are *listeners* to the program. Among the women, 90 are non-listeners and 10 are listeners. The results are cross-tabulated in Table 15:6.

This table indicates that there is a tendency for a greater proportion of listeners to be found among men than among women. Is this a chance relationship, or is there a real tendency toward correlation of sex and listening habits? In other words, is the correlation in this sample result significantly greater than zero?

Table 15:6. Cross-Tabulation of Listening Habits Toward a Radio Program with the Sex of the Respondents

	Sex Groups: Men	Sex Groups: Women	n_r
Non-Listeners	a 65	b 90	155 $(a + b)$
Listeners	c 35	d 10	45 $(c + d)$
n_c	100 $(a + c)$	100 $(b + d)$	$N_s = 200$ $(a + b + c + d)$

The first step in a chi-square Test of Significance for independence between attributes consists in determining the hypothetical distribution of frequencies to be expected on the basis of chance alone. For this, the marginal totals of the dichotomies of each attribute, as well as N_s, must be taken into account. Table 15:6 indicates that 155 members of the total sample were non-listeners and only 45 were listeners. If this division of listeners and non-listeners is taken into account, as well as the equal division of the whole group on the basis of sex, the following three constraints are imposed in setting up the theoretical frequencies of the hypothesis to be tested: (1) N_s, the size of the total sample; (2) the equal division on the basis of sex; and (3) the proportionate division of non-listeners and listeners. These three constraints on the hypothetical frequencies mean the loss of three degrees of freedom. Since there are only four categories of frequencies in the table, only one degree of freedom remains.

In order to establish a hypothetical distribution of frequencies for the four cells which will be divided on the basis of chance expectancy, the following computations are necessary for any one cell of the two by two cross-tabulation (although this discussion is based on cell a, any other cell could have been used):

1. The probability of non-listeners is determined. This is based upon the data of the sample result and is equal to the ratio of non-listeners to the total sample, viz., 155/200, or 31/40.

2. The probability of men is determined. This is the ratio of men to the total sample, viz., 100/200, or 1/2.

3. The probability of subjects who are both men and non-listeners is then determined. This is equal to the *product* of the probability of men and the probability of non-listeners, or

$$(155/200)(100/200) = .3875$$

because the probability of the joint occurrence of two independent events (assumed for the hypothesis) is equal to the product of their respective probabilities. This, then, is the probability, under the conditions of the sample result, of obtaining male non-listeners on the basis of chance. The probability value of .3875 provides an estimate of the result to be expected for cell *a* if the attributes are in fact independent.

Since chi-square is a technique for testing hypotheses about *frequencies*, the probability value of .3875 must be converted to the *number* of frequencies to be expected for a sampling distribution in which $N_s = 200$. Hence the hypothetical frequencies for any cell are equal to the product of the probability value for the cell and N_s (the size of the sample). For cell *a* this is 77.5 (the product of .3875 and 200). This value is therefore the hypothetical frequencies of male non-listeners to be expected on the basis of chance alone for samples of 200 cases, 155 of whom are non-listeners and 100 of whom are men.

The calculation of the hypothetical frequencies for any cell of a cross-tabulation of the data of two attributes may be summarized as follows:

[15 : 4]

$$f_h = \left(\frac{n_r}{N_s}\right)\left(\frac{n_c}{N_s}\right) N_s = \frac{n_r n_c}{N_s}$$

Hypothetical frequencies for any cell of cross-tabulated attributes (based on product theorem of the probability of the joint occurrence of independent events)

where n_r is the number of cases in the row that intersects the cell; n_c is the number of cases in the column that intersects the cell; and N_s is the size of the sample. Thus for cell *a* in Table 15:6:

$$f_{h(a)} = \left(\frac{155}{200}\right)\left(\frac{100}{200}\right) 200 = \frac{(155)(100)}{200} = 77.5$$

Once the number of hypothetical frequencies for any cell of a fourfold table is determined, the frequencies for the remaining three cells are strictly determined, for there is only one degree of freedom. The hypothetical frequencies of cell *c* are equal to $(a + c) - a$, i.e., $100 - 77.5 = 22.5$; those of cell *b* are equal to $(a + b) - a$, i.e., $155 - 77.5 = 77.5$ and those of cell *d* are equal to $(c + d) - d$, or $(b + d) - d$. The hypothetical distribution of frequencies to be expected on the basis of chance for the data in Table 15:6 are summarized in Table 15:7.

Having established the hypothetical frequencies to be expected for each cell on the basis of chance, we can now proceed with a chi-square Test of Significance. Chi-square is calculated exactly as before. The computations are given in Table 15:8, where the value of chi-square is seen to be 17.92. Table 15:2 indicates that when there is one degree of freedom the probabilities are only 1 in 1000 of obtaining, on the basis of chance alone, a chi-square

Table 15:7. The Hypothetical Distribution of Frequencies for a Chi-Square Test of Independence of the Cross-Tabulated Data in Table 15:6

	Sex Groups: Men	Sex Groups: Women	n_r
Non-Listeners	a: 77.5	b: 77.5	155
Listeners	c: 22.5	d: 22.5	45
n_c	100	100	$N = 200$

Table 15:8. Computation of Chi-Square for the Test of Independence (Hypothetical Data from Table 15:7; Sample Data from Table 15:6)

Cell	Frequencies from Sample f_s	Hypothetical Frequencies for the Test of Independence f_h	Differences $(f_s - f_h)$	Differences Squared $(f_s - f_h)^2$	Chi-Square Ratio $(f_s - f_h)^2/f_h$
a	65	77.5	12.5	156.25	156.25/77.5 = 2.016
b	90	77.5	12.5	156.25	156.25/77.5 = 2.016
c	35	22.5	12.5	156.25	156.25/22.5 = 6.944
d	10	22.5	12.5	156.25	156.25/22.5 = 6.944
					$\chi^2 = \Sigma = 17.92$

For 1 *d.f.*: $P = .001$ for $\chi^2 \geqq 10.83$

value equal to or greater than 10.82. Since the chi-square value obtained is considerably larger than 10.82, we can reject the null hypothesis. In other words, we can be quite confident that at least some of the correlation in the sample result is not fortuitous but is based on other than chance factors. The correlation, although apparently not large, is nevertheless significantly greater than zero.

Pearson's Short-Cut Computation of χ^2 for 2 by 2 Cross-Tabulations *

The preceding Test of Significance for the cross-tabulated data of dichotomized attributes can be quickly computed by a short-cut method that is algebraically equivalent but eliminates the separate computation of the hypothetical frequencies. This was developed by Karl Pearson and is obtained by the following:

* K. Pearson, *op. cit.*, p. xxxiv.

$$\chi^2 = \frac{N_s(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)} \quad [15:5]$$

χ^2 for Test of Independence of 2 by 2 cross-tabulations

For the data in Table 15:8, χ^2 by this short-cut method is as follows:

$$\chi^2 = \frac{200[(65)(10) - (90)(35)]^2}{(65 + 35)(90 + 10)(65 + 90)(35 + 10)} = \frac{1{,}250{,}000{,}000}{69{,}750{,}000} = 17.92$$

Chi-Square Test of Significance for Correlation Between Attributes with More Than Two Categories

Table 4:12 presented a 2 by 4 cross-tabulation between income status of 2026 respondents and their opinion on private vs. government management of business. These data are given in Table 15:9, together with the hypothetical frequencies in brackets.

Table 15:9. Sample and Hypothetical Frequencies for a Chi-Square Test of Independence of Respondents' Income Status and Their Opinions on Private vs. Government Management of Business

	Income Status: Low	Lower Middle	Upper Middle	High	n_r
Private management	*a* 230 [291]	*b* 660 [665]	*c* 570 [531]	*d* 225 [198]	1685 $(a + b + c + d)$
Government management	*e* 120 [59]	*f* 140 [135]	*g* 68 [107]	*h* 13 [40]	341 $(e + f + g + h)$
n_c	350 $(a + e)$	800 $(b + f)$	638 $(c + g)$	238 $(d + h)$	$N_s = 2026$ $(a + b + \cdots + g + h)$

Again as in the preceding problem, the marginal totals for each row and column are taken into account in determining the hypothetical frequencies for a purely chance relationship. The degrees of freedom for any test of independence are given by the following:

$$d.f. = (A_c - 1)(B_c - 1) \quad [15:6]$$

Number of degrees of freedom for a Test of Independence of cross-tabulated data

where A_c represents the number of classes or categories for one attribute, and B_c the number of classes or categories for the other attribute. For the 2 by 4 cross-tabulation in Table 15:9, *d.f.* equals (1)(3) = 3.

Since there are only 3 *d.f.*'s for these data, it is necessary to calculate the hypothetical frequencies of only three cells (in either row); the others are obtained by subtraction. Thus, f_h for cell *a* is:

$$f_{h_a} = \frac{1685(350)}{2026} = 291.1 \text{ or } (291)$$

$$f_{h_b} = \frac{1685(800)}{2026} = 665.3 \text{ or } (665)$$

$$f_{h_c} = \frac{1685(638)}{2026} = 530.6 \text{ or } (531)$$

By subtraction,

$$f_{h_d} = 1685 - (291 + 665 + 531) = 198$$

$$f_{h_e} = 350 - 291 = 59$$

$$f_{h_f} = 800 - 665 = 135$$

$$f_{h_g} = 638 - 531 = 107$$

$$f_{h_h} = 238 - 198 = 40$$

Table 15:10. Computation of Chi-Square for Test of Independence
(Data from Table 15:19)

Cell	Sample Frequencies f_s	Hypothetical Frequencies f_h	Differences $(f_s - f_h)$	Differences Squared $(f_s - f_h)^2$	Chi-Square Ratio $(f_s - f_h)^2/f_h$
a	230	291	61	3721	12.79
b	660	665	5	25	.04
c	570	531	39	1521	2.86
d	225	198	27	729	3.68
e	120	59	61	3721	63.07
f	140	135	5	25	18.52
g	68	107	39	1521	14.21
h	13	40	27	729	18.23
	$N_s = 2026$	2026			$\chi^2 = 133.40$

From Table 15:2:
When *d.f.* = 3, $P = .001$ for $\chi^2 \geqq 16.27$

The computation of chi-square for these data is shown in Table 15:10, and chi-square is found to be 133.4. For 3 *d.f.*, the probabilities are only 1 in 1000 for chi-squares equal to or greater than 16.27. Hence, we can reject the null hypothesis, i.e., that income status and the respondents' opinions are independent of each other. In other words, there is some correlation between these two variables.

Contingency Coefficient

The Contingency Coefficient, C, was presented in Chapter 4 as a measure for the correlation of polytomous attributes, and its limitations were indicated, especially when the number of cells is small (see Table 4:15). C can be computed by the method used in Table 4:13, or by chi-square. In the latter case, it is equal to

$$C = \sqrt{\chi^2/(N_s + \chi^2)} \qquad [15:7]$$

Contingency Coefficient from chi-square

For the data in Table 15:8, C is equal to .29:

$$C = \sqrt{17.92/(200 + 17.92)} = \sqrt{.0822} = .29$$

and for the data in Table 15:10, to .25:

$$C = \sqrt{133.4/(2026 + 133.4)} = \sqrt{.0624} = .25$$

The latter is approximately the same as the value obtained in Table 4:13.

As was stated in Chapter 4, C was developed by Pearson for categorical distributions on the assumption that each attribute is a continuously distributed variable, that the distributions are similar in form to the normal bell-shaped curve, and that a linear function is adequate or most suitable to describe the correlations. Obviously not all of these assumptions are always satisfied. Four decades ago there was more emphasis than there is today on attempts to summarize correlation in terms of a measure analogous or equivalent to product-moment r. Unfortunately, with the value of C limited by the number of cross-tabulated categories, the measure is somewhat ambiguous in its implications concerning correlation.

A test of the null hypothesis for C, the contingency coefficient, is obviously unnecessary when C is derived from chi-square. As a matter of fact, when the chi-square Test of Significance does not warrant the rejection of the null hypothesis, there is no point in computing C.

Relation Between χ^2 and ϕ

Karl Pearson * pointed out the following fundamental relation between χ^2 and ϕ:

$$\chi^2 = N_s\phi^2 \qquad [15:8]$$

Equivalence of χ^2 and ϕ

or

$$\phi = \sqrt{\chi^2/N_s} \qquad [15:9]$$

Equivalence of ϕ and χ^2

* Karl Pearson, *op. cit.*, p. xxxv.

EXERCISES

1. Describe the kinds of hypotheses for which chi-square tests of significance are relevant.
2. Set up a relevant hypothesis for the sample data in Table 2:4 and test it by chi-square.
3. Set up a relevant hypothesis for the frequencies of males in Table 3:6 and test it by chi-square.
4. Determine by chi-square whether either one of the distributions in Table 6:7 diverges significantly from the normal, bell-shaped distribution.
5. Use chi-square to determine the degree of correlation in terms of the Contingency Coefficient for the following data:*

"If a man was paid $50.00 a week for 48 hours' work in wartime and he is now working only 40 hours a week, should he still be paid $50.00?"

	Socio-Economic Groups			
Respondents' Replies	*D*	*C*	*B*	*A*
Yes	260	450	285	145
No	195	470	420	325

6. Use the conversion formula (15:9) to obtain the phi coefficient for the data in Exercise 5.

* Adapted from H. C. Link, "The Psychological Corporation's Index of Public Opinion," *Journal of Applied Psychology*, 30:1–9, 1946.

CHAPTER 16

The Predictive Meaning of Correlation

The correlation coefficient is an index which summarizes the degree of association or co-variation characteristic of the relationship between two attributes. Several methods for measuring the *degree* of such correlation were presented in Chapters 4, 9, and 10. We saw that many of the correlation problems in psychology and related fields can usually be dealt with by Pearson's product-moment method of linear correlation, or by methods that give an estimate of Pearson's *r*. However, this earlier treatment was presented primarily from the point of view of *descriptive* statistics. Hence we shall now proceed to amplify the implications of correlation for problems in sampling and analytical statistics.

One of the most practical ways of interpreting correlation coefficients is in terms of their predictive meaning.* Although a correlation coefficient is not necessarily obtained for the purpose of predicting values of one variable from given values of the other, the meaning of predictive estimates is basic to an evaluation of the practical usefulness of correlation. Only thus can we evaluate the *practical meaning* of varying degrees of *r*, such as .40, .60, .80, .90, etc.

We saw earlier that the product-moment correlation coefficient is the slope of a straight line that best fits a cross-tabulated set of correlational fre-

Fig. 16:1. Scatter of Correlation Frequencies for $r_{xy} = .00$. Scatter at a Maximum

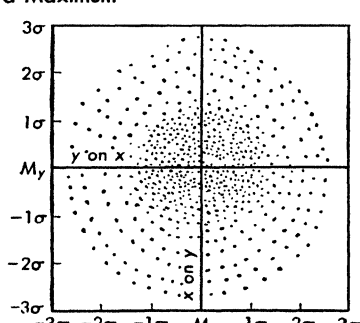
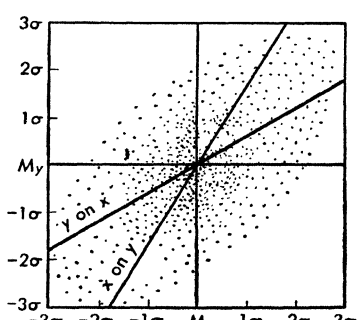


Fig. 16:2. Scatter of Correlation Frequencies for $r_{xy} = .60$



* Cf. J. G. Peatman, "On the Predictive Meaning of Correlation," *Journal of General Psychology*, 22:17–23, 1940.

quencies whose respective measures have been equalized in terms of z scores. We saw, further, that all bi-variates have two regression lines, the regression of $\bar{y}$ on x and the regression of $\bar{x}$ on y, and that the value of product-moment r is the geometric mean of the regression coefficients for each best-fitting straight line function. From the point of view of sampling, the *scatter* of the correlational frequencies about the regression lines gives a basis for determining the *accuracy* with which values of one variable can be predicted from those of another. For example, as the correlation coefficient approaches zero, the scatter of the correlational frequencies approaches a maximum, as indicated in Fig. 16:1. On the other hand, as the correlation coefficient approaches unity (either 1.0 or −1.0), the scatter of the correlational frequencies about the regression line approaches a minimum. In the case of perfect correlation, there is no scatter whatsoever. (See Figs. 16:2, 16:3, and 16:4.) The scatter of

Fig. 16:3. Scatter for $r_{xy} = .60$ as Measured per Class Interval in Terms of σ_{est_y} $(\sqrt{1 - r^2_{xy}})$, and with Equal Variability for Each Class Interval Assumed

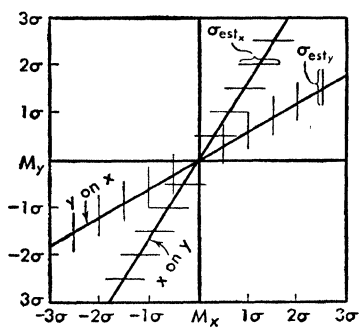
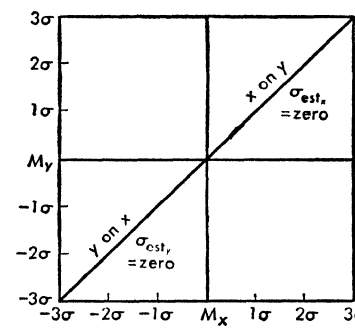


Fig. 16:4. Perfect Positive Correlation, $r_{xy} = 1.0$. No Scatter: σ_{est} = zero



normal correlation surfaces forms an ellipse which becomes narrower and approaches a straight line as the correlation increases, and which widens and approaches a circle as the correlation decreases toward zero.

Two problems arise in interpreting the predictive meaning of a correlation coefficient. The first involves the mathematical procedure for making a prediction. The second concerns the accuracy or efficiency of the prediction made. The making of a prediction is based mathematically on the regression equation of the line that best fits the co-variation of the data. The accuracy or efficiency of the prediction is determined in terms of the *standard error of estimate*. This is based on the standard deviation of the scatter of correlational frequencies about the best-fitting straight line, and is an estimate of the standard deviation of the sampling distribution of predicted values. For a correlation between variables x and y, there are two standard errors of estimate: one for the regression of $\bar{y}$ on x, and the other for the regression of $\bar{x}$ on y.

A. MAKING THE PREDICTION

The prediction of values of one variable from given values of the other is based on the correlation between the sample results. When the correlation is linear, as it is assumed to be for product-moment correlation, predictions can be made by means either of regression (straight-line) equations or of a straight line fitted to the graphic distribution of variations of x with respect to y (or y with respect to x). We saw in Chapter 9, Section B, that the regression equations in deviation score form are as follows:

[16 : 1]

($\bar{y}$ on x): $\bar{y} = r\frac{\sigma_y}{\sigma_x}x$, or $\bar{y} = rb_{yx}x$

Regression of $\bar{y}$ on x in product-moment correlation

[16 : 2]

($\bar{x}$ on y): $\bar{x} = r\frac{\sigma_x}{\sigma_y}y$, or $\bar{x} = rb_{xy}y$

Regression of $\bar{x}$ on y in product-moment correlation

in which the regression coefficients b_{xy} and b_{yx} are respectively

$$b_{xy} = r\frac{\sigma_x}{\sigma_y} \quad \text{and} \quad b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

In practice, however, it is usually more convenient to estimate one variable from given values of the other in terms of original scores rather than in terms of x and y. The regression equations in original score form are as follows:

($\overline{Y}$ on X): $\overline{Y} = r\frac{\sigma_y}{\sigma_x}X - r\frac{\sigma_y}{\sigma_x}M_x + M_y$ [16 : 3]

Regression of $\overline{Y}$ on X

($\overline{X}$ on Y): $\overline{X} = r\frac{\sigma_x}{\sigma_y}Y - r\frac{\sigma_x}{\sigma_y}M_y + M_x$ [16 : 4]

Regression of $\overline{X}$ on Y

The use of these two equations in making predictive estimates will be illustrated with the following data: The correlation between intelligence test scores (x variable) and grades (y variable) was found to be .50 for a college sample of 200 subjects. The means and standard deviations of each variable were approximately:

Intelligence Test Scores (x)	Grades (y)
Mean = 80.0	Mean = 75.0%
σ_x = 15.0	σ_y = 8.0%

Assuming that the relationship is *linear* and that these results are derived from a random sample of a normally distributed universe, we can estimate the *average* intelligence test score for a given grade score or, conversely, the *average* grade score for a given test score. The latter is generally more often required because scholastic achievement (as measured by grades) is usually considered to be a function of intelligence. The interrelation of causal factors, however, is complex.

The regression equation for $\overline{Y}$ on X gives the following result for these data:

$$\begin{aligned}\overline{Y} &= .50\frac{8.0}{15.0}X - .50\frac{8.0}{15.0}80.0 + 75.0 \\ &= .267\,X - 21.360 + 75.0 \\ &= .267\,X + 53.640\end{aligned}$$

This equation is the basis for estimating $\overline{Y}$ (grade scores) from given values of $\overline{X}$ (intelligence test scores) when .50 is the correlation for the sample data. In order to check the summary of the equation, it is well to predict the average intelligence test score for a grade score equal to the mean value of 75.0% (or a grade of C). The average value of X will be equal to the mean intelligence test score, because the two regression lines of a product-moment correlation matrix intersect at coordinates projected from the arithmetical means of the respective distributions. Although an actual prediction from the mean of 80.0 is thus unnecessary, the computation is relevant, since it checks the values obtained in the equation. Thus:

$$\overline{Y} = .267(80.0) + 53.640 = 21.360 + 53.640 = 75.0\%$$

The average grade score of students with intelligence test scores of 95.0, one standard deviation above the mean,

$$M_x + 1\sigma_x = 80.0 + 15.0 = 95.0$$

is predicted as follows:

$$\overline{Y} = .267(95.0) + 53.64 = 25.365 + 53.640 = 79.0\%$$

This value indicates that students with intelligence test scores of 95 will, *on the average*, have a grade score of 79%, or C+.

The average grade score for students with intelligence test scores of 110.0, two standard deviations above the mean, will be equal to the following:

$$\overline{Y} = .267(110.0) + 53.640 = 29.370 + 53.640 = 83.0\%$$

Students with intelligence test scores of 110 will have an average grade score of approximately 83%, or slightly less than B. Students with intelligence test scores of 50.0, two standard deviations *below* the mean, will have, on the average, the following grade scores:

$$\overline{Y} = .267(50.0) + 53.64 = 13.350 + 53.640 = 67.0\%$$

Such students will, on the average, be expected to have grade scores of 67.0%, or slightly better than D.

Such predictions are the best estimates possible with the sample data available. Each prediction, it should be emphasized, is an *average estimate.* Only if the correlation between two variables is perfect will all the values of Y be the same for a given value of X. As the correlation approaches zero,

the average estimate for Y is less and less accurate, because the scatter about the average value increases. The greater the scatter, the less reliable the prediction; and, conversely, the less the scatter, the more accurate or reliable the prediction. When the correlation is zero, the scatter for estimates of Y from any value of X is at a maximum and is equal to the range of deviation of the y variable itself. Before describing a method for computing the *error of estimation* for a given value of X when the correlation is not zero, we shall illustrate graphically the predictive estimates which have been made.

Predictions on a Correlation Matrix

Two types of graphs can be used to show the predictive relationship between two variables. In one type, the geometric field of coordinate axes is used, as in Fig. 16:5. The x and y variables are scaled on the abscissa and ordinate, respectively, in such a way that the scale distances are equalized in terms of standard deviation units. In other words, the scales are the same as those used in Fig. 9:10 for a z score cross-tabulation chart. The scales for both z scores and original scores are given for each variable.

Fig. 16:5. Predicted Values of y from Given Values of x, When $r_{xy} = .50$

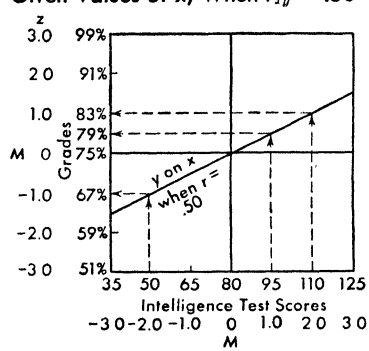
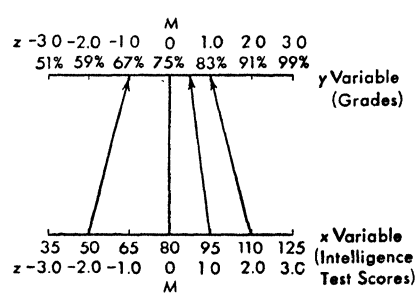


Fig. 16:6. Predicted Values of y from Given Values of x, When $r_{xy} = .50$



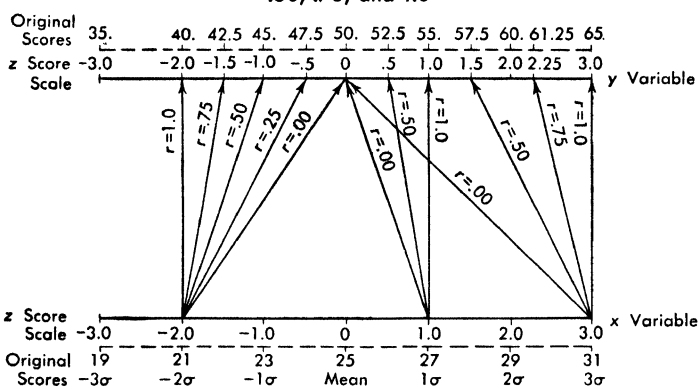
As already indicated, the predicted estimate of Y from the mean value of x is the mean of y, because the regression line intersects the coordinate axes at the means. Also, a grade score one-half a standard deviation above the mean grade score is the value predicted from an intelligence test score one standard deviation above the mean intelligence test score. The predicted value with the regression equation is 79.0%, and this value is also obtained by the graphic method in Fig. 16:5. The other predictions computed above are also shown.

In the second type of graph, two parallel horizontal scales, one for each variable, are used, as shown in Fig. 16:6. Instead of the two scales being laid off as the two axes of a geometric field, they are laid off parallel to each other. The deviation distances on each scale are drawn in z score units that are

equal to each other, with the mean of the y variable directly opposite the mean of the x variable. A point one standard deviation above the mean of the y variable has the same relative position on the scale as a point one standard deviation above the mean of the x variable, etc. The predictions computed above are shown in this figure.

When r is .50, the z score value of one variable predicted from the z score value of the other will be exactly half as great (except at the mean). This follows from the fact that in z score form the regression equation of $\bar{z}_y$ on z_x is $\bar{z}_y = .50z_x$. When r is .50, the value of y predicted from any given value of x is exactly half the latter's standard deviation distance from the mean of y.

Fig. 16:7. Values of y Predicted from Several Values of x When r_{xy} = .00, .25, .50, .75, and 1.0



The prediction of measures of one variable from another variable that is correlated with it is further illustrated in Fig. 16:7. Again, the relationship between two variables is shown by parallel scales drawn so that the corresponding z score values are directly opposite each other. The predictive estimates made from z_x scores of -2.0, $+1.0$, and $+3.0$ are drawn for different degrees of correlation, namely, r = .00, .25, .50, .75, and 1.0. It is apparent that when r is zero, regardless of the z score value from which a prediction is made, the average estimated value on the y variable is the mean of y (a z_y score of zero). On the other hand, when the correlation is positive and perfect (r = 1.0), the predicted z_y values are identical with the z_x values from which the predictions are made. As the correlation decreases from 1.0 to .00, however, the predicted values of z_y "regress" toward the mean value of y, i.e., $z_y = 0$.

In negative correlation, prediction is logically similar to that in positive correlation except that the positive deviations of one variable tend to be associated with the negative deviations of the other. In other words, when there is a perfect negative correlation, z_x values of 3.0 are associated with

z_y values of -3.0, z_x values of -2.0 are associated with z_y values of $+2.0$, etc. As negative correlations approach zero, the regression is similar to that in positive correlation, since the predicted values for one variable progressively approach the mean value of that variable.

B. THE ACCURACY OR EFFICIENCY OF PREDICTIONS

Only a few of the predictive implications of product-moment correlation were illustrated in the preceding section. These are the initial aspects of the total situation. The research worker must also be aware of the *accuracy* or *efficiency* of predictions for varying degrees of correlation. Predictions can always be made from either of two correlated variables, regardless of the degree of correlation between them. As we have seen, when the correlation coefficient itself is zero, a predictive estimate can be made even though it is no more informative than a guess. When r is zero, all predictive estimates for one variable will be the mean of that variable, regardless of the values of the other variable. However, implications of the accuracy or efficiency of predictive estimates vary tremendously for different degrees of correlation, the efficiency being zero when r is zero.

The Standard Error of Estimate

We have already pointed out that the *scatter* of correlational frequencies about either regression line furnishes a graphic picture of the efficiency with which a predictive estimate can be made. When the correlation is zero, the scatter is at a maximum. In the case of bi-variates which are distributed normally, the correlational frequencies are concentrated near the center of the bi-variate distribution and the scatter is distributed circularly from that center. As the degree of correlation increases, negatively or positively, the scatter decreases gradually and forms an elliptic pattern about the regression line.

What is next needed is an algebraic method for expressing the degree of scatter characteristic of different product-moment correlation coefficients. The scatter is measured in terms of the standard deviation, and is the standard error of estimate already referred to.* This estimate of scatter about the best-fitting line is readily obtained from the following formulas:

$$\sigma_{est_y} = \sigma_y\sqrt{1 - r_{xy}^2} \qquad [16:5]$$

Standard error of estimate of $\bar{y}$ on x

$$\sigma_{est_x} = \sigma_x\sqrt{1 - r_{xy}^2} \qquad [16:6]$$

Standard error of estimate of $\bar{x}$ on y

where σ_y and σ_x are the standard deviations of the distributions of the variables correlated, and r_{xy} is the correlation between them.

* The P.E. of estimate is sometimes used; it is equal to .6745 σ_{est}.

The measure of scatter is thus a function of the degree of co-variability between two variables. The expression under the radical is the basic measure of the degree of scatter between two variables and serves to reduce σ_x or σ_y accordingly. T. L. Kelley * called $\sqrt{1 - r_{xy}^2}$ the *coefficient of alienation* and symbolized it by k. This coefficient gives the ratio of the variability of the measures in any class interval of x or y to the variability of y or x as a whole.† Thus, the estimated scatter of any class interval for the y variable is:

$$k = \frac{\sigma_y\sqrt{1 - r_{xy}^2}}{\sigma_y} = \sqrt{1 - r_{xy}^2} \qquad [16:7]$$

Coefficient of alienation, k

An inspection of this formula shows that when r is equal to 1, k is equal to zero, and that consequently the values of the standard error of estimate in Formulas 16:5 and 16:6 also are zero. This is the situation in perfect correlation, since there is no scatter about the regression line. On the other hand, when r is zero, k is equal to 1.0, and the value of the standard error of estimate is identical with the measure of variability of the distribution. In other words, when r is zero, the error of estimate is at a maximum and is equal to the standard deviation of the variable itself.

Between r values of 1.0 or -1.0 and zero, the scatter of correlational frequencies about the regression line decreases gradually as the correlation coefficient increases from zero. The way in which the degree of scatter decreases must be clearly understood because it is the basis for interpreting the predictive *efficiency* or *accuracy* of correlation coefficients. The relationship between the estimate of error, k, and varying degrees of correlation is given in Table 16:1.

As already indicated, if r is zero, the error of estimate is at the maximum, and is equal to the standard deviation of the variable for which predictions are being made. Thus, when r is zero, any predictions of $\bar{y}$ from x or of $\bar{x}$ from y are no better than a guess. Regardless of the value of x, the predicted value of y will be the mean of the y distribution, and the scatter of the y variable above and below the regression line (whose slope is zero) will be equal to the scatter of the y variable as a whole. In other words, when r equals zero, σ_{est_y} is equal to σ_y. If r is .10, we see from Table 16:1 that k is .995. Thus, an increase from zero to .10 in the degree of correlation decreases the error of estimate by only .005 of its maximum value when r is zero. Hence a correlation coefficient of .10 is likewise little better than a guess. When r is .30, the predictive efficiency is only 5% better than a guess because k is .95. When r is .50, k is .866, and the predictive efficiency is about 13%. Table 16:1

* T. L. Kelley, "Principles Underlying the Classification of Men," *Journal of Applied Psychology*, 3:50–67, 1919.

† The interpretation of the standard error of estimate is based on the assumption of a *normal* correlation surface, i.e., that each variable is normally distributed, and of *homoscedasticity*, i.e., that the scatter or variability of all arrays (or class intervals) of a variable is the same.

Table 16:1 Values of k, the Coefficient of Alienation ($\sqrt{1 - r^2}$) for Values of r from Zero to Plus or Minus 1.00 *

r	k	r	k	r	k	r	k
.00	1.000	.25	.968	.50	.866	.75	.661
.01	.999+	.26	.966	.51	.860	.76	.650
.02	.999+	.27	.963	.52	.854	.77	.638
.03	.999+	.28	.960	.53	.848	.78	.626
.04	.999	.29	.957	.54	.842	.79	.613
.05	.999	.30	.954	.55	.835	.80	.600
.06	.998	.31	.951	.56	.828	.81	.586
.07	.998	.32	.947	.57	.822	.82	.572
.08	.997	.33	.944	.58	.815	.83	.558
.09	.996	.34	.940	.59	.807	.84	.543
.10	.995	.35	.937	.60	.800	.85	.527
.11	.994	.36	.933	.61	.792	.86	.510
.12	.993	.37	.929	.62	.785	.87	.493
.13	.992	.38	.925	.63	.777	.88	.475
.14	.990	.39	.921	.64	.768	.89	.456
.15	.989	.40	.917	.65	.760	.90	.436
.16	.987	.41	.912	.66	.751	.91	.415
.17	.985	.42	.908	.67	.742	.92	.392
.18	.984	.43	.903	.68	.733	.93	.368
.19	.982	.44	.898	.69	.724	.94	.341
.20	.980	.45	.893	.70	.714	.95	.312
.21	.978	.46	.888	.71	.704	.96	.280
.22	.976	.47	.883	.72	.694	.97	.243
.23	.973	.48	.877	.73	.683	.98	.199
.24	.971	.49	.872	.74	.673	.99	.141

shows that a correlation coefficient must be .87 for its predictive efficiency to be 50% better than a guess. Even when r is .99, its predictive efficiency is still far from perfect, being 86% better than a guess. The accuracy of prediction thus increases very gradually as correlation coefficients diverge from zero, and then much more rapidly as they approach 1.00 or −1.00.

The accuracy of predictive estimates based upon product-moment correlation can be concretely illustrated by means of the grade and intelligence test score data used for the predictive estimates in Figs. 16:5 and 16:6. The correlation coefficient for these data was .50, and the standard deviation of the grade scores was 8%. The standard error of the grade scores predicted from the test scores is thus:

$$\sigma_{est_y} = \sigma_y\sqrt{1 - r_{xy}^2} = 8.0\sqrt{1 - (.50)^2} = 8.0(.866) = 6.93 \text{ (or 7\%)}$$

Since the standard deviation of the y variable as a whole is 8% and the variation of any predicted value of a grade score is approximately 7%, the prediction when r is .50 is somewhat better than a guess. The proportionate

* From Table V, Appendix B, pp. 516–517.

increase in efficiency of prediction, as against a sheer guess, is thus the difference between an error of 8% (when r equals zero) and an error of about 7% (when r equals .50). The increase in efficiency of prediction is therefore $(8\% - 7\%)/8\% = 12.5\%$.

The Interpretation of the Error of Estimate

The estimate of Y was found to be an average grade of 79% when predicted from an intelligence test score of 95.0. The standard error of this estimate is 7%. Let us now see how this error of estimate is interpreted in relation to these bi-variates. On the assumption that the bi-variates are normally distributed and that the variability of the grade scores in each array is approximately equal, we can infer that in the long run approximately 68 out of 100 grade score predictions from a given intelligence test score will vary within the limits of 79% ± 7%, or between 72% and 86%. This is the range of plus and minus *one* standard error, which in the normal probability distribution includes approximately .68 of the area.

In other words, on the basis of these sample results, students with intelligence test scores of 95.0 will, on the average, have grade scores of 79%. In the long run, approximately two-thirds may be expected to have grades between 72% (C⁻) and 86% (B). These are the limits for a sampling distribution of predicted scores $\pm 1.0\sigma_{est}$ from the grade score of 79%, predicted from an intelligence test score of 95.0. If we take into account the range of a normally distributed sampling distribution—two standard deviation units above and below the predicted grade score value—approximately 95 in 100 such predictions should in the long run lie within the limits of 79% ± 2(7%), i.e., from 65% to 93%, or, in terms of letter grades, from D to A. Finally, since a range of ±2.5 or 3.0 standard deviation units includes nearly 100% of the area of the normal probability distribution, practically all grade score values predicted from an intelligence test score of 95.0 will lie within the range of 79% ± 2.5(7%), or 79% ± 3(7%). By the 2.5σ criterion these limits are 61.5% and 96.5%; and by the 3σ criterion they are 58% and 100% (100% being the upper limit of actual possibility).

These estimated ranges of possible error (sampling variations) in prediction are similar to the confidence limits already described in Chapter 13 for Tests of Significance. In reality, what we have been doing is to set up Tests of Significance for a continuum of hypothetical parameter estimates, the results of which determine the range of acceptable or *likely* hypotheses. At the same time, the limits for *unlikely* hypotheses are also established. Thus, by the T ratio criterion of 2.5, we can reject with confidence the hypothesis that students with intelligence test scores of 95.0 will have grade scores less than 61.5% (D⁻) or greater than 96.5% (A), these being the limits of the predicted grade score value ±2.5 times the standard error of the estimate.

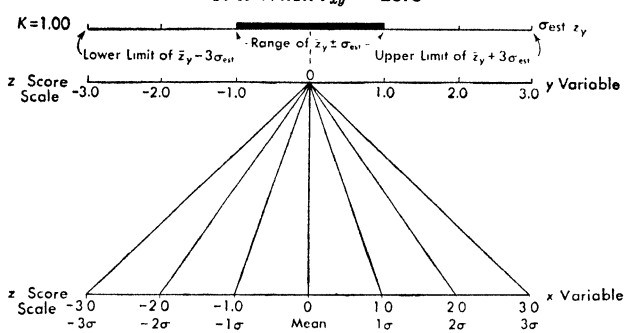
Thus, in predicting values of one variable from a given value of another, we

first estimate the range of most likely hypotheses in the light of the sample result, and we then indicate the limits for unlikely values. We saw in Chapter 12 that values beyond a range of ± 2.5 or ± 3.0 standard error units are *unlikely* in random sampling, and that values within the range of ± 2.0 error units are *likely*. These criteria of 2.0 and 2.5 or 3.0 standard error units determine the range of questionable or doubtful hypotheses; that is, they can be neither accepted nor rejected with confidence. We have seen that the limits of likely hypotheses are determined by the variability of the sample result when the correlation is zero. On the other hand, when the correlation is .87, the limits for *likely* hypotheses are but half as great as they are when r is zero. This is what is meant by the statement that the efficiency of predicted estimates is 50% better than a guess when r is .87.

Graphic Representation of the Accuracy of Predictive Estimates

In Fig. 16:8 the graphic technique employed in Figs. 16:6 and 16:7 has been used further to illustrate the error characteristic of predictive estimates when r is zero. All the predicted values for y are at the mean, regardless of the x

Fig. 16:8. Error of Estimate for y Predicted from Various Values of x When r_{xy} = zero



score from which they are predicted. Furthermore, the range of error is assumed to be distributed normally and over limits similar to those of the y variable itself. In other words, when r is zero we cannot with confidence reject the hypothesis that the limits of the value of one variable which is predicted from a given value of the other will be any smaller than the limits of the entire distribution. Although, on the average and in the long run, we should expect that the most likely value of predicted scores would equal the mean of the distribution, there is less probability of this value coinciding with the mean than lying somewhere else in the distribution. The mean is only the modal point. This situation is analogous to the sampling distribution obtained from tossing 20 coins, for which a result other than 10 heads is more likely to occur than exactly 10 heads. In any event, when r is zero, any prediction

is a guess. Under the circumstances (i.e., a normal sampling distribution) the best guess is the mean of the sample distribution.

Fig. 16:9 shows the range of variability or sampling error to be expected in predictions from the mean of x, for varying degrees of correlation. This

Fig. 16:9. Error of Estimate for Values of y Predicted from the Mean Value of x When r_{xy} = .00, .25, .50, .75, .90, and .99

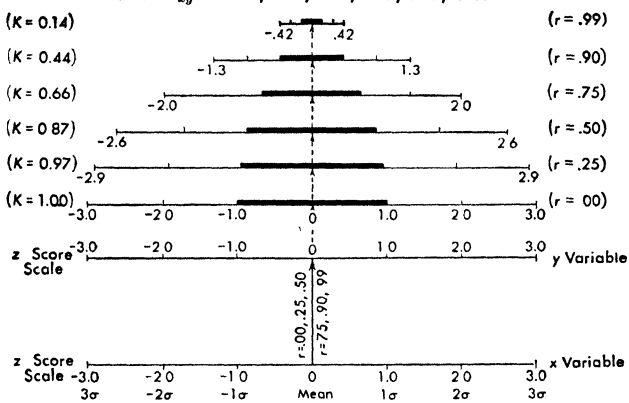
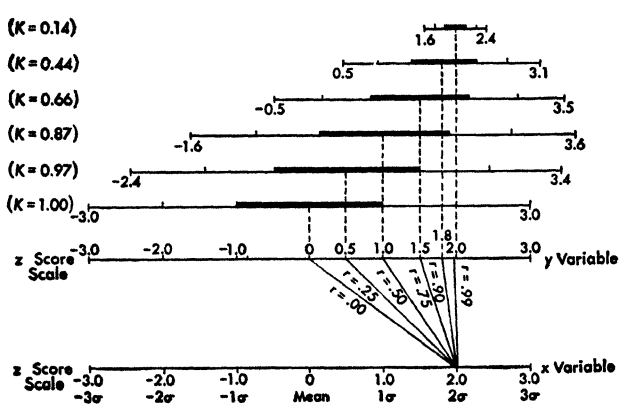


figure further illustrates the fact that the range of the error for these estimates decreases gradually as the correlation coefficient increases from zero. Since all the predicted values of y are made from the mean of the x variable, the $\bar{y}$ estimates are equal to the mean of the y variable.

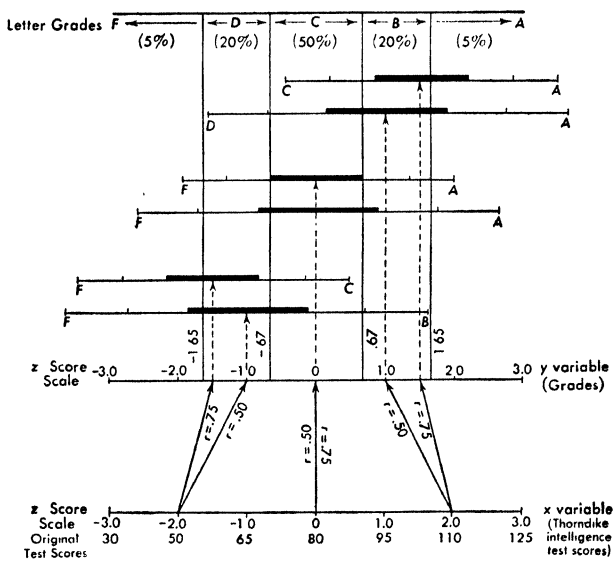
Fig. 16:10 illustrates the effect of correlation on the predicted estimates of

Fig. 16:10. Error of Estimate for Values of y Predicted from a Value of x Equal to a z score of 2.0, When r_{xy} = .00, .25, .50, .75, .90, and .99



$\bar{y}$ from x at a point two standard deviations above the mean of x, and the range of error characteristic of such estimates. Six different degrees of correlation are used, ranging from zero to a correlation of .99. The divergence of the *average* estimate from the mean of the y distribution as r increases is shown in relation to the range of expected sampling error. Again it is apparent that the accuracy of prediction increases gradually as the correlation increases from zero. The sampling error has a sizable range, even with a correlation of .75. Thus the range of *likely* hypotheses for the values of y estimated from z_x equal to 2.0 is fairly great, the limits of the estimate $\pm 2.0\sigma_{est}$ being equal to z_y scores of 0.18 and 2.8. When the correlation is .99, the limits for *unlikely* hypotheses (by the T criterion of 3.0) are z_y scores of 1.6 and 2.4. In other words, y cannot be estimated from x without error even for such a high correlation.

Fig. 16:11. Error of Estimate for Values of y (Grades) Predicted from Thorndike Intelligence Test Scores of 50 and 110, When $r_{xy} = .50$ and .75



The above predictive implications of correlation are brought together in Fig. 16:11 for the intelligence test scores and grades used earlier in this chapter. The grade scores are the y variable; they are scaled at the top of the figure in five broad classes of letter ratings from F to A, with the average expectancy of frequencies for each class given in percentages. The equivalent z score limits of each class are given on the z_y score scale in the middle of the figure. Thus, a grade of C includes a range on the y variable whose limits in terms of z scores are taken as -0.67 and $+0.67$. The *universe* is assumed to be dis-

tributed normally. In such a distribution z score limits of -0.67 and $+0.67$ mark off the range of the middle 50% of the frequencies. The range of B grades is from z score values of 0.67 to 1.65 and is taken so as to include 20% of the frequencies. The range of D grades includes the same proportion of frequencies as the B grade range, the z score limits being -0.67 and -1.65. The remaining 10% of the frequencies are evenly divided between the A and F grades. The A ratings lie beyond the z score limit of 1.65 and the F ratings lie below the z score limit of -1.65.*

The predictive meaning of correlation is illustrated for coefficients of .50 and .75. Estimated grades are predicted from intelligence test scores of 50 ($z_x = -2.0$), 80 (z_x = zero), and 110 ($z_x = 2.0$). The coefficient of .50 is the sample result and is typical of the correlations between intelligence test scores and college grades reported in the literature. The correlation of .75 is higher than any empirical value which has come to the writer's attention.

Fig. 16:11 illustrates both the predicted letter grades and their relation to their respective ranges of variation or error. Thus, the predicted estimate for students with an intelligence test score of 110 is a letter grade within the B range, regardless of whether the correlation is .50 or .75. However, if the correlation is only .50, the lower limit of *unlikely* results is a D. In other words, we can reject as unlikely the hypothesis that individuals with intelligence test scores of 110 will have letter grades lower than D. The upper limit of *unlikely* results is the maximum rating, namely, an A. But the upper limit of *likely* hypotheses also lies within this same grade interval. Consequently, the upper limit really imposes no restriction on what is likely or unlikely for people with such an intelligence test score.

If the correlation were .75 instead of .50, then, according to the figure, the lower limit for *unlikely* hypotheses would lie within the C interval, and consequently it would be most unlikely that students with intelligence test scores of 110 would have grade scores of less than C. In view of the sample results, i.e., r equals .50, the most plausible hypothesis is that most students with intelligence test scores of 110 will have scholastic averages ranging from C to A.

When scores are predicted from values below the mean of x, it is the upper range for unlikely hypotheses, rather than the lower, which is of major concern. Thus, as indicated in Fig. 16:11, grade scores predicted from an intelligence test score of 50 lie in the D interval, and the upper limit for unlikely hypotheses is a grade of B when r equals .50, and a grade of C when r equals .75. In other words, it is unlikely that students with intelligence test scores of 50 will have grade score averages above B when the correlation between the two is .50, and it is likewise not likely that their grade scores will be above C if the correlation is .75.

Finally, when grade scores are estimated from near the mean intelligence

* Such a division of letter grades for distributions assumed to be normal is often used in educational circles, but is not thereby to be condoned. There are no immutable laws operating in the scholastic natures of *students* which dictate that 5% are to fail, etc.

test score, both the upper and the lower limits for unlikely hypotheses are relevant in appraising the accuracy of the estimates. According to Fig. 16:11, for a correlation of either .50 or .75, these limits are in the F and A grade ranges for a T criterion of either 2.5 or 3.0. If a range of ± 2.0 standard errors is taken as the range of likely hypotheses, then when r equals .50 the grade scores of students with intelligence test scores of 80, the mean value, will most likely range from D to B.

The Index of Predictive Efficiency (E)

Since k, the coefficient of alienation, measures the proportionate reduction in error or scatter for estimates of a variable predicted from given values of a second variable, the efficiency of prediction can be measured, or indexed. The index commonly used for this purpose, and symbolized as E by Clark Hull,* expresses as a percentage the proportionate reduction in the error of estimate

Table 16:2. The Index of Predictive Efficiency, E, for Values of r from Zero to Plus or Minus 1.00

$$E = 100\%\left(1 - \sqrt{1 - r_{xy}^2}\right)$$

r	E	r	E	r	E	r	E
.00	0.0%	.25	3.2%	.50	13.4%	.75	33.9%
.01	0.1⁻	.26	3.4	.51	14.0	.76	35.0
.02	0.1⁻	.27	3.7	.52	14.6	.77	36.2
.03	0.1⁻	.28	4.0	.53	15.2	.78	37.4
.04	0.1	.29	4.3	.54	15.8	.79	38.7
.05	0.1	.30	4.6	.55	16.5	.80	40.0
.06	0.2	.31	4.9	.56	17.2	.81	41.4
.07	0.2	.32	5.3	.57	17.8	.82	42.8
.08	0.3	.33	5.6	.58	18.5	.83	44.2
.09	0.4	.34	6.0	.59	19.3	.84	45.7
.10	0.5	.35	6.3	.60	20.0	.85	47.3
.11	0.6	.36	6.7	.61	20.8	.86	49.0
.12	0.7	.37	7.1	.62	21.5	.866	50.0
.13	0.8	.38	7.5	.63	22.3	.87	50.7
.14	1.0	.39	7.9	.64	23.2	.88	52.5
.15	1.1	.40	8.3	.65	24.0	.89	54.4
.16	1.3	.41	8.8	.66	24.9	.90	56.4
.17	1.5	.42	9.2	.67	25.8	.91	58.5
.18	1.6	.43	9.7	.68	26.7	.92	60.8
.19	1.8	.44	10.2	.69	27.6	.93	63.2
.20	2.0	.45	10.7	.70	28.6	.94	65.9
.21	2.2	.46	11.2	.71	29.6	.95	68.8
.22	2.4	.47	11.7	.72	30.6	.96	72.0
.23	2.7	.48	12.3	.73	31.7	.97	75.7
.24	2.9	.49	12.8	.74	32.7	.98	80.1
						.99	85.9
						1.00	100.0

* C. Hull, *Aptitude Testing*, World Book Co., Yonkers, 1928.

from the maximum error characteristic of zero correlation. It is given by the following:

$$E = 100\%(1 - \sqrt{1 - r_{xy}^2}) = 100\%(1 - k) \qquad [16:8]$$

Index of predictive efficiency, E

The E values for varying degrees of correlation from zero to plus or minus 1.00 are given in Table 16:2. Thus, a coefficient of .60 has a predictive efficiency 20% better than a sheer guess. That is, when r equals .60, the range of likely hypotheses for any predictive estimate is 20% less than the range of such hypotheses when r equals zero.

When r is .866, the predictive efficiency is 50%. In other words, a correlation must be .866 in order for the range of error to be only 50% as great as in the case of a guess, i.e., zero correlation. If r equals .95, the predictive efficiency is 69%. If r is 1.00, its efficiency of prediction is 100%, since there is no error or scatter for a perfect correlation.

Standard Error of Estimate for the Mean (σ_{est_M})

A group result in terms of its mean can be predicted more accurately than can a particular score because the standard deviation of a sampling distribution of means is considerably less than it is for a sampling distribution of particular scores. The latter measure is given by the standard error of estimate, $\sigma_{est_{\bar{Y}}}$ and, as we have seen, is a function of the scatter in the bi-variate distribution. The standard error of estimate of the mean of y from the mean of x, on the other hand, is given by the following:

$$\sigma_{est_{M_y}} = \frac{\sigma_y}{\sqrt{N_s - 1}} \sqrt{1 - r_{xy}^2} \qquad [16:9]$$

Standard error of estimate for the mean of one variable predicted from the mean of a correlated variable

where $\sigma_y/\sqrt{N_s - 1}$ is the standard error of a mean (Formula 13:5a) and $\sqrt{1 - r_{xy}^2}$ is k.

When predicted from a correlated variable, the standard error of a mean is thus reduced by the value of k, or, from the point of view of the standard error of estimate of $\bar{y}$ on x, the error in predicting a mean score is reduced by $\sqrt{N_s - 1}$. (Note that when N_s is large, this value can be taken simply as $\sqrt{N_s}$.)

A correlation coefficient of .50 is considerably more efficient for predicting a mean score than a correlation coefficient of .90 for predicting particular scores. When r equals .90, k is .44 (Table 16:1), and E is 56% (Table 16:2). In predicting a mean score, even if r is only .50 for a sample of 100 cases, the standard error of the mean estimate is $k/\sqrt{100}$, which is one-tenth of .87, or .087. The index of predictive efficiency, E, is therefore 100%(1.00 − .087) = 91%.

Thus in the correlation between students' intelligence test scores and their college grades, the standard error of estimate of a predicted mean grade is as follows:

$$\sigma_{est_{M_y}} = \frac{8.0}{\sqrt{100}}\sqrt{1 - (.50)^2} = \frac{6.93}{10} = 0.7\%$$

We can therefore be confident that the mean of the universe from which the sample results were derived will lie within a range of 2.5(0.7%) = 1.75% grade points. This estimate is fairly precise, E being 91% as indicated above.

Tests of Significance for Predictive Estimates

The logic of Tests of Significance for predictive estimates is the same as for the statistics already considered:

$$T = \frac{s - h}{\sigma_s}$$

where s is the sample value, h is the hypothetical parameter value, and σ_s is the standard error of the measure or statistic under consideration. In a Test of Significance for a predictive estimate, s is the predicted value based on the *sample bi-variate distribution*, h is the parameter value of a relevant hypothesis, and σ_s is the standard error of estimate ($\sigma_y\sqrt{1 - r_{xy}^2}$).

We shall present a Test of Significance for a predictive estimate with the data from Fig. 16:11. Is it likely, when $r_{xy} = .50$, that students with intelligence test scores of 50 will have grade scores equal to or greater than A$^-$? The hypothesis to be tested is that the parameter value of the predictive estimate is equal to at least a grade score of A$^-$, which we shall consider as a percentage grade of 90. We have already seen, in Fig. 16:11, that the grade score predicted for students with intelligence test scores of 50 is -1.0 standard deviations below the mean grade. Since the mean grade is 75% and the standard deviation of the grade distribution is 8%, the predicted score in terms of percentages is 67%. The standard error of estimate is $8\%\sqrt{1 - (.50)^2} = 6.9\%$. The Test of Significance for the hypothesis of a grade of A$^-$ or better is therefore as follows:

$$T = \frac{67\% - 90\%}{6.9\%} = \frac{33\%}{6.9\%} = 4.8$$

Since the T ratio is 4.8, we can reject the hypothesis with confidence. In other words, it is most unlikely that students with intelligence test scores of no more than 50 will have grade scores as high as 90%. It should be noted that this conclusion does not, and logically cannot, rule out the possibility of an individual exception. As we have repeatedly emphasized, the application of the theory of probabilities to empirical data is based on what under certain circumstances can be expected to occur *in the long run*, and hence the improbable case is not excluded.

We shall next test the hypothesis that students with intelligence test scores of no more than 50 will have grade scores of C or better. With 75% representing the parameter value of this hypothesis, and with the data from the preceding example, the Test of Significance is as follows:

$$T = \frac{67\% - 75\%}{6.9\%} = \frac{8.0}{6.9} = 1.2$$

This time the test ratio is 1.2, and hence the hypothesis cannot be rejected. Furthermore, since a T ratio of 2.0 or less can be taken as a criterion for *likely* hypotheses, this result is likely. In other words, at least some students with intelligence test scores of 50 will in all likelihood have grade scores of C or better. However, many other hypotheses are also likely. Such hypotheses are denoted by confidence limits set up in terms of criteria equal to ± 2.0 standard error units from the sample value. Since the standard error of estimate is 6.9%, the confidence limits for likely hypotheses will be $67\% \pm 2(6.9\%)$, or approximately 53% and 81% (letter grades of F and B$^-$). Since a T ratio of 2.5 has been used as the criterion for unlikely hypotheses, such hypotheses would lie in the range of possible grade score values below $67\% - 2.5(6.9\%) = 50\%$, and above $67\% + 2.5(6.9\%) = 84\%$. In other words, grade values of below 50% (F) or above 84% (B) are unlikely for students with intelligence test scores of 50.

Summary

From the preceding sections, it should be evident that correlation coefficients of less than .30 have little value for predictive purposes. Even a coefficient of .50 or .60 does not yield a very accurate estimate of y from x. Correlation coefficients in the .80's or .90's are *high* from the point of view of their predictive efficiency.

However, coefficients of .40 or .50 may be useful in predicting upper score limits for variable y from low scores of x, or lower score limits for y from high scores of x. When considered in relation to other coefficients by the multiple correlation method (cf. Chapter 17, Section E), correlations as low as .20 or .30 may even be of value in increasing the predictive efficiency of a battery of tests.

Whether or not a correlation of .50, for example, is low, fair, or high thus cannot be answered categorically; it depends upon the nature of the situation. In the following chapter we shall consider the use of the technique of correlation in evaluating psychological tests, a field in which an understanding of the predictive implications of coefficients of correlation is particularly essential for an adequate interpretation of results.

EXERCISES

1. Distinguish making a prediction and evaluating its accuracy.
2. Describe how the accuracy or efficiency with which predictions can be made is dependent upon the extent of scatter in a bi-variate distribution.

Use the results obtained for Exercises 10 and 13 of Chapter 9 for the following two problems:

3. Determine the efficiency with which the intelligence test scores of the freshmen's best friends can be predicted from the freshmen's intelligence test scores of 90. Draw a graph to illustrate.
4. Determine the efficiency with which the average grade index of the freshmen's best friends can be predicted from the freshmen's average grade of 65. Illustrate with a graph.

CHAPTER 17

Correlation Methods for the Evaluation of Psychological Tests

The technique of correlation has come to be recognized as an indispensable statistical tool for the appraisal and evaluation of psychological test procedures.* The developments in this field are well illustrated by the contrast in the psychological test procedures used in World War I and in World War II. Psychological tests for measuring "intelligence" were used for the first time on a large scale during the First World War, when a group of outstanding psychologists developed the Army Alpha and Beta tests for the purpose of differentiating aptitude for army work. By and large, these tests were based upon considerations more rational than empirical. However, this implies no criticism of the work of these psychologists, for they proceeded in accordance with the best standards, information, etc., available at the time. World War II saw the use of many different kinds of psychological tests and classification procedures by all branches of the Armed Forces, the result of the wealth of empirical knowledge accumulated during the intervening twenty-five years, as well as of the continuing process of validation during the war itself. In all such work in test development the technique of correlation is indispensable to any adequate evaluation of results.

The evaluation of psychological tests is usually approached with two basic considerations in mind, namely, *test reliability* and *test validity;* they will be considered in this chapter. A third aspect of the problem has received much attention since 1930: the organization of abilities or aptitudes, i.e., how they are manifest and interrelated. The treatment of this aspect has been essentially statistical, and the techniques involved are usually known as *factor* or *cluster analysis;* they are considered in Chapter 18.

Reliability and Validity a Question of Degree

A test is said to be reliable if it is accurate or consistent, and to be valid if it measures what it is supposed to measure. For practical purposes, however, these definitions need to be clarified and restated in terms susceptible to empirical analysis. Thus, a test is reliable and valid to the extent that its

* Cf. J. G. Peatman, "On the Meaning of a Test Score," *American Journal of Orthopsychiatry*, **9**:23–47, 1939. For a review of the recent literature, see H. S. Conrad, "Statistical Methods Related to Test Construction and Evaluation," *Review of Educational Research*, **14**:110–126, 1944.

results enable the prediction of various kinds of behavior in educational, occupational, or social situations. This prediction is a matter of degree, and hence the consistency with which a test differentiates performance is likewise a matter of degree. The empirical problem, therefore, involves determining the degree or extent to which a test is reliable and valid, rather than attempting to answer such poorly framed questions as "Is it reliable?" or "Is it valid?"

A. THE RELIABILITY AND VALIDITY OF A BAROMETER AND OF A PSYCHOLOGICAL TEST

Inasmuch as analogies are oftentimes helpful, we shall illustrate the parallel implications of reliability and validity in terms of physical measurement. That a Torricelli barometer provides a relevant analogy is evidenced by the frequent allusion in the social sciences to the "barometric" character of this or that measure or index.

The Barometer

What does a barometer measure? It has been experimentally established that it measures atmospheric pressure. The measurement itself is obtained in terms of the height—in inches, millimeters, or bars—of a mercury column in a tube. The behavior of the mercury column is closely correlated with changes in atmospheric pressure. The correlation is perfect except for errors of observation and errors implicit in the instrument. If the barometer is well designed, the error averages only .005 of an inch when the tube is one-quarter of an inch in diameter, and only .002 of an inch if the diameter is .5 inch. This, then, is the index of the reliability of the barometer, from which it follows that the differentiations of atmospheric pressure are highly reliable. The reliability of the instrument is determined by making repeated readings under experimentally varied and highly controlled conditions. When employed under similar circumstances, a barometer is found to yield highly consistent measures. The accuracy of the measurement varies to some extent with differences in the composition of the atmosphere and the size of the tube. But the fact remains that a well-constructed barometer is so reliable an instrument that if its reliability were expressed as a correlation coefficient, r would approach 1.00.

How about its "validity"? (1) We have already indicated that a barometer measures atmospheric pressure. This can be designated as its *operational validity*, i.e., validity in terms of operations which, in the case of this instrument, correlate perfectly with atmospheric pressure. (2) It can also be used to predict a type of physical behavior which is distinct from atmospheric pressure as such, that is, changes in weather. This is one of the most important practical functions of a barometer, and can be designated as one aspect of its *functional validity*. It can also be used to gauge the altitude above or

below sea level. At sea level, the barometric measure is about 30 inches; at 1000 feet below, the measure is about 31 inches; at 1000 feet above, it is about 29 inches; and at 50,000 feet above, it is about $3\frac{1}{2}$ inches. The practical implications of such an instrument in air navigation are obvious. Its functional validity in relation to weather and altitude is measured by the degree of correspondence between the readings and the other physical factors.

The Psychological Test

Now let us examine the analogy between a barometer and a psychological test. For this we shall use the Minnesota Vocational Test for Clerical Workers,* and we shall take up the analogous points in turn: (1) the immediate character of the measures (test scores) obtained; (2) the reliability of the instrument (test); and (3) the validity of the measures (test scores), with respect to both operational validity and functional implications.

Measures Obtained (Test Scores)

The Minnesota Clerical Test consists of two parts, *number-checking* and *name-checking.* Numbers and names are each listed in pairs, the subject's task being to discriminate dissimilar and identical pairs. Two measures are obtained: (1) the total number of correctly discriminated number-pairs, and (2) the total number of correctly discriminated name-pairs. Since the task is fairly simple for most adults, there is a *time limit,* and *speed* thus becomes an integral part of the meaning of the score.

In some psychological tests, the measures may be obtained in inches (as in *steadiness* tests) or seconds or minutes (as in any amount-limit test, reaction-time tasks, etc.). In the Minnesota Clerical Test, however, the measure is a *count*—an enumeration of the total number of tasks correctly performed. This type of measure is characteristic of the kind obtained with many other psychological tests.

The Reliability of the Measures Obtained

The usual index of reliability of a test is a correlation coefficient which measures the *consistency* with which the abilities sampled are *differentiated.* It is only indirectly analogous to the index of reliability of a barometer (which, as we have seen, is taken in terms of the expected error and may average only .005 or .002 of an inch). Furthermore, the reliability coefficient of a test is obtained by one of several methods to be described in Section B, and is in itself a somewhat unsatisfactory measure because the correlation obtained is affected by the range of ability of the sample used in standardizing the test. Thus, an estimate of test reliability based on a sample of college

* D. M. Andrew, D. G. Paterson, and H. P. Longstaff, *Minnesota Vocational Test for Clerical Workers,* Psychological Corporation, New York, 1933.

freshmen will most likely be lower than one based on all 18-year-olds. Fortunately, however, the standard error of a test score, σ_X (cf. Chapter 13) takes into account the variability of a sample group and at the same time yields an index of reliability more immediately meaningful and useful than the usual reliability coefficients given for tests.

The Standard Error of a Test Score (σ_X). A test is used to obtain measures that can be identified with a scale, or series of measures, which will signify for each test score a relatively consistent placement, or position, on the scale. The scale yielded by a test is assumed to represent a continuum that ranges from the *least* to the *greatest* degree of the abilities being measured. (This assumption of continuity may be difficult to justify in the case of personality and interest inventories.) The behavior of a mercury column in a barometer is similarly assumed to be scaled with respect to such a continuum. The higher the reading, the less the atmospheric pressure; the lower the reading, the greater the pressure. The problem of measuring the reliability of an instrument, therefore, depends on determining the accuracy of the location of each measure on the scale: Is its position subject to a large range of error, or to such a small range as to be negligible for all practical purposes? For any psychological test, the standard error of a test score, σ_X, gives an *estimate* of the accuracy of the result, in terms of its location on the scale. Hence, it is the most practical and meaningful index of the reliability of a test.

If a person receives a score of 150 on number-checking in the Minnesota Test for Clerical Workers, is this significantly greater than the median score of 144 obtained from a sample of employed clerical workers? Can a difference of six units on the scale be expected on the basis of chance, or is the score of 150 significantly greater than the median score value of 144? The following Test of Significance answers these questions:

$$T = \frac{X_s - X_h}{\sigma_X}$$

where X_s is the individual score, X_h is the value of the *hypothesis* tested (in this case, the median of 144), and σ_X is an estimate of the standard error of X_h, which we saw earlier is equal to the following:

$$\sigma_X = \sigma_x\sqrt{1 - r_{xx}}$$

where σ_x is the standard deviation of the distribution of measures, and r_{xx} is a measure of the reliability of the test.* Reliability coefficients as high as .91 have been reported for this test; consequently, this value will be used to simplify the example. If the standard deviation of the distribution is taken as 25 units, the standard error of a score will be:

$$\sigma_X = 25\sqrt{1 - .91} = 25(.30) = 7.5$$

* See Table V, Appendix B, for values of $\sqrt{1 - r}$.

This value of 7.5 is analogous in its implications about reliability to the index of reliability of the barometer. When the reliability coefficient of a test is .90, the standard error of an individual score is about one-third the standard deviation of the distribution.

We now have the necessary values for the above Test of Significance:

$$T = \frac{X_s - X_h}{\sigma_X} = \frac{150 - 144}{7.5} = \frac{6.0}{7.5} = 0.8$$

This T ratio, 0.8, is too small to warrant the conclusion that a score of 150 is significantly greater than the median of 144.

What result would be significantly greater than a median performance of 144? If a T ratio of 2.5 is taken as the criterion for a significant difference, then 2.5 times the standard error of a score will give the desired estimate, 2.5(7.5), which equals 18.75, a difference of 19 units in the scale. If the score is 169 rather than 150, we can conclude that it is significantly greater than the median.

A test is thus more reliable, the narrower, so to speak, the range of possible error of a particular score on the scale. This was also the case for the barometer, whose error is no greater than a few thousandths of a unit (inch) on the scale. If the location of a test score on a psychological scale is subject to a wide margin of error, the test will not be very useful.

The Validity of Test Scores

What about the validity of the Minnesota Clerical Test? As in the case of the barometer, we must distinguish between its operational and its functional validity.

Operational Validity. The Minnesota Test measures two closely related types of psychological functions, viz., the speed of *number-checking* and *name-checking.* We can assume that scores on the test are directly correlated with differences in ability to perform these tasks. Although this correlation should theoretically be perfect, it is not, because of errors of response and those arising in administering and scoring the test. Furthermore, the degree to which these abilities are imperfectly measured is indicated by the reliability of the test scores.

It is difficult, if not impossible, satisfactorily to summarize the operational character of some psychological instruments, as for example, a personality or an interest inventory. On the other hand, a test may have useful functional implications in diagnosing or in predicting behavior in life situations, and at the same time have no agreed-upon operational validity. Fortunately, from the point of view of counseling and measurement, this is not a handicap. When a test of "general mental ability" is shown empirically to be valuable

in predicting scholastic aptitude, or aptitude for a particular occupational situation, it makes little difference what it is called, unless its name has been loaded historically with emotion or misleading implications, such as the term "intelligence."

To summarize, simply *naming* a test does not establish its operational validity; rather, what it measures is dependent upon a sound analysis of the tasks and responses directly involved. The question still remains, however, whether it will satisfactorily predict behavior in this or that occupational situation. This is the problem of its *functional validity.*

Functional Validity. We saw that a barometer is used in the prediction of changes in the weather as well as in the measurement of distance above or below sea level. The latter involves practically no error because of the extremely high correlation with atmospheric pressure. However, weather predictions cannot be so accurate because factors other than atmospheric pressure alone are involved. The situation in psychological measurement is even more complex because behavior in a given situation is influenced by a multiplicity of factors, many of which are unknown or immeasurable. Thus clerical success depends upon many more factors than *number-checking* or *name-checking.* Consequently, the empirical problem is to determine whether a *critical score* for the test can be established that will *generally* differentiate the members of a given population who are likely to succeed from those who are likely to fail.

Critical Scores

In practice, it is desirable to set two critical scores: one score that will permit the selection of individuals most likely to succeed; and another, lower score that will delimit those most likely to fail. The values between these two scores constitute a range whose implications are doubtful. Such a procedure takes into account the unreliability of the test scores. The authors of the Minnesota Clerical Test make no attempt to determine such scores; instead, they say that "the critical scores for the selection of employees for a given occupation should be determined by the hiring standards for the particular company." This is sound; in fact, it is the only practical procedure. Business and industrial organizations should develop their own critical scores on the basis of practical experience.

Still another critical score that is sometimes useful is an *upper critical* score such that values beyond it are more indicative of failure than of success. Many organizations have found that people who have very high scores on a mental aptitude or clerical test do not do well in routine jobs, not because they lack ability, but because the jobs are not sufficiently interesting to continue to motivate them. *Aptitude* is compounded of both potential *abilities for* and *interest in* the tasks to be performed.

B. THE DETERMINATION OF TEST RELIABILITY

Despite the fact that we might "measure" distance with an elastic yardstick, it is obvious that we could not be assured of the consistency of our measurements. Yardsticks must be constructed in such a way that the measurements obtained with them will be consistent. Similarly, a watch or clock will yield a "measure" of time as long as it is running. However, unless its measurement of seconds, minutes, and hours is such that it keeps correct time, it is unreliable and consequently unsatisfactory, despite the fact that *time* itself is "measured."

The measurement of distance, time, weight, etc., is based upon the centimeter-gram-second (c.g.s.) system. For each of these types of measurement, the fundamental problem is not validity but reliability, because the measurement of distance, time, or weight is inherent in the operation performed. That is, by definition, distance, time, and weight are each measured by a series of appropriate and well-standardized operations. In all such cases, there is no doubt that the measuring instrument is yielding an observation of the required kind. In other words, there is no question about the *operational* validity. The real question in each case is the degree of reliability of each type of measurement.

We have seen that the operational validity of the Minnesota Clerical Test is analogous to that of measures based on the barometer. Another example is the operations comprising the reaction-time experiment, which measures the speed with which an individual can react to a stimulus. The result is given in terms of time, and there is no question but that reaction time is measured. The basic question concerns the reliability of the result.

The operational validity of an intelligence test or of a personality inventory is much more difficult to define. It is not sufficient to say that "intelligence" is that which is measured by an intelligence test. Such a definition is useless unless the operations actually involved in the test are described and understood in detail. Even then, the definition is often unsatisfactory because the avowed purpose of an intelligence test is to differentiate what people can be expected to do in various life situations. The necessary empirical approach to determining the validity of an intelligence test is the *functional* approach of ascertaining the kinds of behavior which are predictable, at least to some degree, from such tests.

The Correlation Index of Test Reliability

Since the product-moment correlation coefficient is an index of the degree of co-variation between bi-variates, the reliability of a test can be expressed in terms of it. The greater the consistent differentiation of the quality or trait measured by the test, the more reliable the test. If, for example, a test is administered and the results obtained are correlated with those obtained from the same people on a readministration of the same test, the technique

of correlation can be applied to ascertain the degree to which the test differentiates the individuals in *relatively* the same way on both administrations. If the correlation is low, the test is no more reliable than an elastic yardstick. If it is high (.90 or more), it should give a fairly satisfactory differentiation.

Unfortunately, there is no one best procedure for obtaining an index of the reliability of a test. The most commonly used methods are:

1. The test-retest, by which a group's performance on a test is correlated with its performance on a readministration of the same test.
2. The correlation of a group's performance on alternate forms of the same test.
3. The split-half technique, or correlation of individuals' scores on odd and even items of a test.

A fourth method, item inter-correlation, is also sometimes employed.

All these techniques attempt to measure the reliability of a test in terms of the consistency with which it differentiates the attribute or trait measured, irrespective of the attribute or trait concerned. When alternate forms of a test are available, the second method is one of the best. The split-half technique is not entirely satisfactory because the index of reliability derived from it is not necessarily meaningful over a period of time; i.e., it measures the consistency of individual differentiation at the time the test is administered. The first and second methods are often superior in this respect. This is an important consideration in tests of ability or aptitude. In attitude questionnaires and personality and interest inventories, however, it may not be so important because the psychological qualities involved may be expected to change more than abilities and aptitudes, at least during the period of growth. At any rate, the split-half technique has come to be used extensively for estimating the reliability of an inventory or questionnaire.

Test Reliability by the Method of Test-Retest (r_{xx})

In estimating the reliability of a psychological test by the test-retest method, the "accuracy" of the differentiation is checked by means of the readministration of the same test to the same group. The people who take the test should be chosen randomly from the universe for which the test is designed. An index of reliability is obtained by correlating the two sets of results. This correlation coefficient provides a measure of the consistency with which the differentiation of the test results is maintained over the time between the two tests. The coefficient itself is symbolized by r_{xx}, where r is the product-moment correlation coefficient, the subscripts x and x standing for the test variable correlated with itself.

The consistency of the differentiation of individual results over a period of time is important *relatively* rather than absolutely. A person might well achieve a higher score on the second administration of a test, but the real

question from the point of view of test reliability is whether his performance remains relatively the same—whether his relative position in the test scale is unchanged. The product-moment correlation coefficient measures the degree of co-variation between two variables irrespective of the absolute size of the scores or measurements, because the scales that measure both variables are made comparable in terms of units of standard deviation. Consequently, product-moment correlation is well adapted to measuring the extent to which the relative position on a scale remains unchanged.

The test-retest method is likely to yield too high rather than too low an estimate of reliability because of the possible correlation of "memory factors" or of "errors." That is, unless the time interval between the tests is sufficiently long, the responses to many items in the test (whether correct or not) may be remembered. This makes for positive correlation. For this reason, it is best to use this method only when there can be a sufficient time interval between the two administrations of the test.

Test-Retest Reliability of a Digit-Span Test

Test-retest reliability will be illustrated by the digit-span test which has been long used in Binet intelligence testing.* A coefficient of .86 was obtained for two administrations, at an interval of about two months, of the same digit-span test to 142 college students. None of the subjects knew in advance that the same material would be used for the second test. The reliability was satisfactory, despite the fact that the subjects consisted of college students, and hence represented a rather restricted range of general ability. With samples less restricted in this respect, the coefficient should be considerably higher—well in the .90's. Nevertheless, the test-retest method was appropriate for determining the reliability of this type of test material because it was unlikely that the subjects would remember any particular item from one test to the other. However, the result may have been influenced by some correlation between such factors as mnemonic techniques developed during the first testing and carried over into the second. But the correlation would have been lowered rather than increased, if other individuals employed such techniques during the second test but not during the first.

This latter point suggests one of the shortcomings of the methods used to estimate the reliability of any test. Actually, of course, a test has no meaning as a measuring instrument unless it is considered in relation to a person. Consequently, its reliability cannot be appraised or evaluated independently of people's responses to it. A test might at a particular time yield a highly consistent differentiation of the abilities or psychological factors tested, whereas, when given again to the same subjects, the reliability coefficient might be lowered as the result of factors which are desirable and in them-

* J. G. Peatman and N. M. Locke, "Studies in the Methodology of the Digit-Span Test," *Archives of Psychology*, No. 167, 1934.

selves of psychological significance. This is to some extent true of any test of ability or capacity, but it is particularly true of the various personality and interest inventories which have been developed for psychological diagnosis, because the level of ability or capacity generally varies less than personality, interest, attitude, etc., at least during the period of growth. A high test-retest coefficient for a personality inventory such as the Bell * or an interest inventory like the Kuder † would indicate relatively no change in the attributes inventoried. If the time interval between the two administrations of the same inventory is long, a high coefficient may be significant independently of the reliability of the instrument itself. On the other hand, a low coefficient does not in itself necessarily reflect on the reliability of the instrument. Changes in the personalities, interests and attitudes of growing boys and girls are to be expected. The test-retest method is consequently the least satisfactory technique for estimating the reliability of personality and interest inventories.

Test Reliability by the Method of Alternate Forms ($r_{xx'}$) ‡

A second commonly used method for estimating the reliability of a test consists in (1) administering one form of the test to a sample, (2) administering an alternate but equivalent form to the same sample, and (3) correlating the results on the two forms to measure the reliability. The reliability coefficient obtained by this method is symbolized by $r_{xx'}$, the subscript x representing the variable measured by the first test, and the subscript x' representing the alternative form of the test. The latter subscript differentiates this coefficient from the test-retest coefficient (r_{xx}).

Reliability coefficients obtained from alternate forms of a test are likely to be too low rather than too high because of the impossibility of devising two forms of the same test that are really equivalent. In fact, *equivalence* is itself a complicated concept, so far as psychological tests are concerned. Two tests are equivalent if they differentiate a population in exactly the same way. However, if this is taken to mean that the individual differentiation must be consistent, the reasoning is circular, because it implies that the correlation between the two forms must be perfect, or at least very high. But this correlation is what is being sought as a measure of the reliability of the test. In practice, therefore, two forms of a test are considered fairly equivalent if they yield similar means, variations, and distributions for appropriate samples of the universe for which the test is designed. Two items of a test are usually considered equivalent if, for a given sample, the task involved is

* H. M. Bell, *The Adjustment Inventory: Adult Form*, Stanford University Press, Stanford University, 1938.

† G. F. Kuder, *Preference Record*, Science Research Associates, Chicago, 1942.

‡ Such intelligence tests as the Army Alpha and the revised Stanford-Binet are available with alternate forms. Often, however, alternate forms of a test are not available.

similar, and a similar test result in terms of errors and correct performances is obtained for each.

Test Reliability by the Split-Half Method ($r_{\frac{x}{2}\frac{x'}{2}}$)

Perhaps the most widely used method for estimating the reliability of tests is the split-half technique. It has the advantage over the other two in that an estimate can be obtained from only one administration of a test. This obviates the need for an alternate form, and also the possibility of the result being distorted by memory factors, as in the test-retest technique. Unfortunately, however, the split-half method does not always yield the result actually sought. That is, an investigator who wishes to measure the consistency with which a test differentiates the relative abilities or capacities of a universe over a period of time may find this method inadequate because it yields a coefficient of reliability for the test only at a *particular time*. Furthermore, the method is subject to manipulation in that the longer the test or the more parts it has, the higher the coefficient obtained by this method. That this is the case will be clearer after the technique has been described.

The basic principle underlying the split-half method is the division of each person's results into halves, and the correlation of the group's results for each half of the test. The basis on which the division is made is, of course, important. The division is usually made by obtaining each individual's results on the odd and on the even items of the test; hence the name, the *odds-even method* of reliability sometimes given this technique. The reliability coefficient obtained by this method is symbolized by $r_{\frac{x}{2}\frac{x'}{2}}$, the subscript $\frac{x}{2}$ representing one-half of the test results, and the subscript $\frac{x'}{2}$ the other half.

This method is similar to the alternate test method, in that each half is analogous to an alternate form of a test. However, in the split-half method, the two halves are administered not consecutively but simultaneously; i.e., the subject does an odd-numbered item, then an even-numbered, etc.

This method may be exemplified by a vocabulary test administered to a group of 181 college students. The test consisted of a list of 80 words to be defined in terms of multiple-choice alternatives. Two scores were obtained for each subject: the total number of correct responses (1) for the 40 odd-numbered items and (2) for the 40 even-numbered items. These results were then correlated for the group, a product-moment r of .77 being obtained.

Spearman-Brown Prophecy Formula *

The correlation coefficient obtained by the split-half technique provides an index of reliability which is too low for a test as a whole, since the coefficient

* C. Spearman, "Correlation Calculated from Faulty Data," *British Journal of Psychology*, 3:281, 1910; W. Brown, "Some Experimental Results in the Correlation of Mental Abilities," *ibid.*, 3:299, 1910.

is based upon a bi-variate composed of the two halves of the test rather than on two whole tests, as in the test-retest method. The reliability of the test *as a whole* can be estimated by means of the Spearman-Brown prophecy formula, which is based on the assumption that increasing the length of a test is theoretically possible without changing the difficulty, character, or any other relevant conditions attendant upon the administration of the test. The generalized formula is as follows:

$$r_L = \frac{Lr_{xx'}}{1 + (L - 1)r_{xx'}} \qquad [17:1]$$

Spearman-Brown prophecy formula

where L symbolizes the ratio between the desired length and the actual length of the test employed, and $r_{xx'}$ symbolizes the reliability coefficient derived from the administration of alternate forms of the test.

This formula can be used to estimate the reliability of a test whose length is increased as many times as is desired. Thus, with the split-half correlation coefficient, L is equal to 2, because the whole test is, of course, twice as long as either of its halves. Hence, for estimating split-half reliability, the generalized formula simplifies to the following:

$$r_{xx'(2L)} = \frac{2r_{\frac{x}{2}\frac{x'}{2}}}{1 + r_{\frac{x}{2}\frac{x'}{2}}} \qquad [17:2]$$

Spearman-Brown prophecy formula for estimating the reliability of a test as a whole

Applying this formula to the split-half coefficient of .77 obtained for the vocabulary test gives the following *estimate* of the reliability of the vocabulary test as a whole:

$$r_{xx'(2L)} = \frac{2(.77)}{1 + .77} = .87$$

It is this estimated reliability for the test as a whole that is usually reported in the literature as the split-half reliability coefficient.

Split-Half Reliability by Method of Differences for $r_{xx'}$

Earlier (pages 248–249) we presented another method of correlation which is often convenient to use to obtain a split-half reliability coefficient. The method of differences for r has an advantage over the above procedure in that the split-half coefficient can be computed directly from the differences between the paired, original odd and even scores. The r coefficient which is obtained is for the two halves of the test, and r for the test as a whole can be estimated by Formula 17:2. It will be recalled that the method of differences consists in obtaining the sum of the squared differences between each individual's original scores on each half of the test. The greater these differences, the less reliable the test. The ratio of the squared differences to the variance of the test as a whole is subtracted from 1.0:

[17 : 3]

$$r_{\frac{x}{2}\frac{x}{2}} = 1 - \frac{\Sigma(D^2)}{2N_s\sigma_x^2}$$

Split-half reliability by the method of differences

where D is the difference between each person's scores on the two halves of the test, squared and summed for the entire group; N_s is the size of the sample; and σ_x^2 is the variance of the distribution of scores for the test as a whole.

The method of sums for r is also convenient to use to obtain r between alternate forms of a test, as was done in Table 9:7, page 249.

Test Reliability by the Method of Item-Intercorrelation

The reliability of a test can also be estimated by the method of item-correlation. This technique is more analogous to the split-half procedure than to either of the other two because it measures the reliability of a test at the time of its administration. However, this method is cumbersome and is usually not worth the excessive amount of statistical computations required unless the intercorrelations between all the items of a test are needed for some other purpose, as in item analysis.

The intercorrelation of responses to items can be facilitated by means of Thurstone's Diagrams for Tetrachoric Coefficients.* The correct and incorrect responses for the items taken two at a time must be cross-tabulated. Once all the intercorrelations between items are obtained, the next step is to compute the average intercorrelation. If all the coefficients are of about the same order, they can be averaged directly with little error; but if they vary considerably (say from .10 to .90), they should be converted to Fisher's z function before they are averaged (see Table VI, Appendix B) or the median intercorrelation coefficient can be employed instead of the mean of the coefficients.

Once the average intercorrelation coefficient is obtained, the reliability of the test as a whole can be estimated by the Spearman-Brown prophecy formula. If a test has 100 items and the average of the intercorrelations between items is .30, the reliability coefficient of the test as a whole is determined by estimating r for 100 items—in other words, for a test 100 times as long as a single item. The Spearman-Brown formula gives the following reliability coefficient:

$$r_{xx'(100L)} = \frac{100(.30)}{1 + 99(.30)} = .98$$

Effect of Range of Ability on Test Reliability

The correlation coefficient for test reliability is definitely affected by the range of ability among the individuals tested. The correlation coefficient for two administrations of the same test obtained from subjects fairly homo-

* L. Chesire, M. Saffir, and L. L. Thurstone, *Computing Diagrams for the Tetrachoric Correlation Coefficient*, Univ. of Chicago Bookstore, Chicago, 1933.

geneous in their abilities will not be as high as it would if the subjects were more heterogeneous in this respect. In other words, a psychological test that differentiates broad ranges of ability fairly well may have little or no value in differentiating narrow ranges. The finer the differentiating power, the more reliable the test. But a low reliability coefficient for a result obtained from a rather homogeneous sample does not mean that the test is entirely useless. The basic question is whether it differentiates abilities sufficiently consistently for the situations in which it is to be used.

The following formula makes it possible to estimate the reliability of a test if the variability of the sample were increased; only the reliability coefficient for the restricted variable and its standard deviation are required:

[17 : 4]

$$r_{xx_L} = \frac{\sigma_{x_L}^2 - \sigma_{x_g}^2(1 - r_{xx})}{\sigma_{x_L}^2}$$

The effect, on the reliability coefficient, of increasing the variability of the universe

in which r_{xx_L} is the reliability coefficient of the variable x with its variability increased; $\sigma_{x_L}^2$ is the variance of the increased, hypothetical variable; $\sigma_{x_g}^2$ is the variance obtained from the sample with the restricted range of ability; and r_{xx} is the reliability coefficient obtained for the sample result.

The use of this formula will be illustrated by the digit-span test data of Table 9:2. The standard deviation of the test was 1.3 and the test-retest correlation coefficient was .84. What would be the coefficient of reliability for this test if the variability were twice as great? The estimated coefficient is determined as follows:

$$r_{xx'(2L)} = \frac{(2.6)^2 - (1.3)^2(1 - .84)}{(2.6)^2} = .96$$

Thus, if the test results had been derived from a random sample of a broader population rather than from college students with their restricted range of ability, the reliability of the test would have been well in the .90's.

All the above considerations are important in appraising the reliability of any test. If the reliability coefficient for a test has been determined from a sample with a rather restricted range of ability, and if the coefficient is relatively low, it does not necessarily follow that the test is worthless as a means of differentiating consistently the functions or qualities being measured. Thus college students, with their restricted ranges of ability, have often been used in the evaluation of tests, and consequently, the general value of some tests has not been clearly recognized.

On the other hand, a test which has an estimated reliability coefficient of .96 computed on the basis of a sample whose variability is theoretically increased is not thereby better than it was originally, so far as its use with the original restricted sample is concerned. In other words, the estimate of increased reliability provides a basis for judging whether the test will be

satisfactory for differentiating the abilities of a broader range of talent in the population.

C. THE DETERMINATION OF TEST VALIDITY

The problem of validity has two aspects: operational validity and functional validity, mentioned earlier. Both of them have been used in psychological measurement, but only in recent years has functional validity received the attention it deserves.

Operational Validity

Whether a test developed to measure clerical ability, for example, is sufficiently valid depends to a great extent upon the way in which the question concerning its validity is framed. If the test includes a series of tasks, such as number-checking and name-checking, and is satisfactorily reliable (the reliability coefficient being .90 or more), a rational appraisal of the operations involved in the test itself might lead to the conclusion that the test is valid (operationally) for clerical ability. Whether it will differentiate clerical ability in actual working situations is, however, another question.

In the operational approach to validity, the specific nature of the tasks or functions comprising the test is described, and the test is defined in these terms. Unfortunately, however, the logic of this approach is not always followed. For example, on the basis of the logic of the operations involved, the Minnesota Clerical Test should be called a *name- and number-checking test*, rather than a test of clerical ability. If such a test proves to be a reliable yardstick for differentiating ability to perform these kinds of tasks, then, *by definition*, it is operationally valid. But its validity is established only for these two operations, and this does not necessarily imply validity for differentiating clerical ability as a whole.

Functional Validity

An empirical appraisal of the functional validity of a test consists in determining whether, in fact, it does differentiate a given ability in actual working situations and, if so, the degree to which it does. This appraisal requires an adequate sample of subjects, and in the case of clerical ability, a measure of each subject's clerical proficiency in the actual working situation, so that these criteria can be readily correlated with his results on the clerical ability test. A high correlation means that individuals who manifest a great deal of clerical ability in their work do well on the test, and that those who manifest less clerical ability in their work do not do as well on the test. The usefulness of a test whose validity coefficient is .80 or .90 should be apparent. Unfortunately, however, no single test of clerical ability has been devised which yields a functional validity coefficient as high as this, because success in clerical occupations depends upon much more than the ability to perform

two or three relatively simple tasks; it depends upon many kinds of abilities, as well as on the individual's personality make-up.

Test Validities

An important implication of the functional aspect of validity is the fact that a test may be valid for more than one type of situation. That a test may have different validities for different situations rather than simply "a validity" is well illustrated by the so-called general intelligence test. This type of test was originally constructed to differentiate the "intelligence" of individuals. From the operational point of view it was assumed to do this because it included a variety of functions or tasks which require "intelligence" for their successful performance. One leading psychologist seriously proposed, in the early 20's, that intelligence be defined as that which intelligence tests measure. From a practical point of view, this reasoning is circular, and hardly resolves the problem. As a result of greater emphasis on the problem of functional validity since that time, it is recognized today that an intelligence test may be more valid for some purposes than for others, and that in any event its validity has a pluralistic rather than only a single aspect. An intelligence test like the Wechsler-Bellevue has been found to be reliable and useful in predicting aptitude for various types of work as well as scholastic aptitude. The validity of this test lies not in its definition as an intelligence test, but rather in the fact that aptitude for many different types of work or activity can to some extent be predicted with it.

Validity Criteria—Abilities vs. Aptitudes

The chief problem in appraising the functional validity of a test is to obtain satisfactory criteria for checking or measuring its validity for a particular situation. This problem is much more complex for the measurement of aptitudes than of abilities because aptitudes, by definition, are abilities not yet fully developed. An aptitude is potential ability and interest, rather than proficiency after training and experience. If, for example, a test is to be designed which will satisfactorily differentiate the mechanical aptitude of high-school students, this requires a test which will measure *capacities for the development* of mechanical abilities.

Aptitude measurement thus involves the determination of whether any abilities are present which, when measured, will serve as a basis for the prediction of later achievements. In a situation of this kind, there are obviously no observations or measurements of abilities in the actual working situation which can be used as criteria for validating the test. Hence, an appropriate sample of subjects is set up and tested, and an index of their particular ability is obtained later, after they have had an opportunity to develop it. With such measures, their actual achievement can be compared with their earlier performance on the aptitude test. If the correlation is high enough

for predictive purposes, then and to that degree only is the test valid for differentiating the particular aptitude.

Effect of Range of Ability on Test Validity

We pointed out earlier that a correlation coefficient is definitely affected by the variability or range of ability characteristic of the sample, and we presented a formula that permits an estimate of the reliability of a test if used with a broader range of ability. Test validity can be estimated similarly. Such an estimate is important whenever there is evidence that a test has been validated with a sample whose range of ability is more restricted than would be characteristic of the general use of the test. For example, many tests have been developed and appraised with samples of college students whose range of ability is necessarily restricted, at least for some kinds of abilities.

If a test is to be used with people whose range of ability is twice as great as that of the group on which the test was standardized, the validity of the test should be considerably greater for individuals with the wider range. This can be estimated by the following formula:

[17 : 5]

$$r_{cx_L} = \sqrt{\frac{\sigma_{x_L}^2 - \sigma_{x_s}^2(1 - r_{cx_s}^2)}{\sigma_{x_L}^2}}$$

The effect, on a validity coefficient, of increasing the variability of the universe

in which r_{cx_L} is the coefficient of validity, i.e., the correlation between the criterion c and the variable of the test x when the variability of the sample is increased; $\sigma_{x_L}^2$ is the variance for the universe of increased variability; $\sigma_{x_s}^2$ is the variance of the test for the universe of the restricted sample; and r_{cx_s} is the validity coefficient obtained from the sample. This formula is based upon the assumption that the standard errors of estimate of both universes are equal:

$$\sigma_{x_L}\sqrt{1 - r_{cx_L}^2} = \sigma_{x_s}\sqrt{1 - r_{cx_s}^2}$$

If the validity coefficient for the sample is .40, the variance of the rest results for the restricted sample is 5.0, and the variance of the universe is twice as great, i.e., 10.0, the validity coefficient will be:

$$r_{cx(2L)} = \sqrt{\frac{10.0 - 5.0(1 - .40^2)}{10.0}} = \sqrt{.58} = .76$$

Test Battery Validity

In many of the most useful procedures for psychological measurement a battery of tests, rather than a single test, is used. In such cases the validity of several tests must be appraised by considering the effectiveness of the battery of tests *as a whole* rather than singly. The correlation technique is used for this, but in appraising the validity of the tests as a whole, simple

correlation is supplemented by *multiple correlation.* This latter, as well as *partial correlation,* will be considered briefly in Sections E and F respectively.

D. TEST ITEM ANALYSIS

Basic research problems in psychological measurement include the evaluation not only of tests but also of the items or tasks of a test. The methods to evaluate test items are basically very similar to those described for appraising tests.

Item Reliability and Validity

The logic underlying the evaluation of test items is straightforward, even though in practice it is often overlooked. Basically, a test item is reliable if its results correlate highly with the total score of the entire test, provided of course the test itself is highly reliable. Furthermore, a test item is operationally valid if it correlates well with the total test which has itself been demonstrated to have satisfactory operational validity. Finally, a test item is functionally valid if it correlates highly with an independent criterion. Even though the test as a whole may not be satisfactory as far as its functional validity is concerned, each item in it can be correlated with an independent functional criterion of validity, and the poor items thus be eliminated.

The validation of psychological test items is too often approached from the point of view of only an internal analysis of operational validity; that is, the total test score is too frequently taken as the only criterion of validity. The particular items that correlate well with the total score are considered relatively valid, and those that correlate low or negatively are considered unsatisfactory or invalid. The soundness of this procedure for the practical problems of measurement obviously depends upon the functional validity of the test as a whole.

Item analysis offers a means of developing a standardized test, for it enables the selection of items of the proper levels of difficulty, and those which yield the best predictions of the criterion.

Biserial and Fourfold Correlation Techniques

Although the statistical techniques employed for item validations are varied, they are generally based on biserial correlation or on correlation techniques for fourfold tables,* because the *answers* to test items are usually dichotomized as "True" or "False," "Correct" or "Incorrect," "Yes" or "No." When operational (internal) validity is to be determined, biserial r is often employed because the criterion is the distribution of scores for the test as a whole.

* Cf. J. C. Flanagan, "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from Data at the Tails of the Distribution," *Journal of Educational Psychology,* 30:674–680, 1939; J. P. Guilford, "The Phi Coefficient and Chi Square as Indices of Item Validity," *Psychometrika,* 6:11–19, 1941, and "A Simple Scoring Weight for Test Items and Its Reliability," *ibid.,* pp. 367–374.

When functional validity is to be determined, tetrachoric r, the ϕ coefficient, or chi-square is usually employed because the independent criterion itself is often dichotomized into "success" or "failure," "satisfactory" or "unsatisfactory," "x present" or "x absent," etc.

These methods of correlation have been described in earlier chapters and consequently will not be described again. Generally, however, test items are valueless unless their correlation with an adequate criterion is significantly greater than zero. An item that every subject passes or fails has no differentiating and, consequently, no test value. An exception to this is the few easy items inserted at the beginning of a test so that the subject will build up self-confidence.

E. MULTIPLE CORRELATION (*R*)

Multiple correlation is a statistical technique that makes it possible to determine the correlation between two or more variables taken together and a single variable—for example, the correlation between several tests and a criterion of proficiency or accomplishment. Multiple correlation not only yields an over-all single coefficient (symbolized by R) but is valuable for determining the effectiveness with which a battery of tests can predict the criterion. The technique is also employed to weight each test in a battery according to its efficiency in this respect.

Although the computations necessary for a multiple correlation problem involving more than three variables are considerable and beyond the scope of this book,* the essence of the technique can be illustrated by a three-variable problem which requires only relatively simple statistical procedures.

Predicting Academic Success from Two Variables

On entering college, students were given a scholastic aptitude test (variable x) and a test of "social intelligence" (variable y). At the end of their sophomore year, criteria of academic success were obtained in terms of each student's average grade for the first two years (variable c, the criterion). The correlation, r_{cx}, between scholastic aptitude test results and academic success was .60. The correlation, r_{cy}, between scores on the social intelligence test and academic success was .40. The efficiency of prediction for a correlation of .60 is 20% (see Table 16:2), but only 8% for a correlation of .40.

The *combined* predictive efficiency of the two tests cannot be obtained simply by summing the correlation coefficients, averaging them, and using this average for an efficiency index. Whether or not the efficiency of prediction is greater when the two test variables x and y are taken together than when each is considered separately depends upon their relationships with the criterion as well as between themselves. That this is the case will be readily seen

* Cf. C. C. Peters and W. R. Van Voorhis, *Statistical Procedures and Their Mathematical Bases*, McGraw-Hill. New York, 1940, chap. 8.

if we cite an extreme situation. If the correlation between variable x and the criterion were 1.00, and that between variable y and the criterion were also 1.00, the second test obviously would not increase the predictive efficiency of the first. Conversely, if each variable correlated zero with a criterion, the predictive efficiency of the two variables taken together would remain zero. The predictive possibilities between these extremes of perfect and zero correlation are varied.

The multiple correlation of two variables with a criterion is computed by the following formula:

$$R_{c \cdot xy} = \sqrt{\frac{r_{cx}^2 + r_{cy}^2 - 2r_{cx}r_{cy}r_{xy}}{1 - r_{xy}^2}} \qquad [17:6]$$

Multiple correlation of two variables, x and y, with a criterion, c

where $R_{c \cdot xy}$ symbolizes the multiple correlation of the test variables x and y with c, the criterion variable; r_{cx} is the correlation of x with the criterion; r_{cy} is the correlation of y with the criterion, and r_{xy} is the correlation between the two test variables.*

In computing the multiple correlation for the data cited above, we already have the values of r_{cx} and r_{cy}. The only additional information needed is the correlation between the two test variables themselves, which was .50. To summarize:

r_{cx} = .60 (correlation of scholastic aptitude test scores with the grade criteria of academic success)

r_{cy} = .40 (correlation of the social intelligence test scores with the criterion)

r_{xy} = .50 (correlation of scholastic aptitude test scores with the social intelligence test scores)

Substituting these three values in the above formula for R gives the following multiple correlation coefficient:

$$R_{c \cdot xy} = \sqrt{\frac{(.60)^2 + (.40)^2 - 2(.60)(.40)(.50)}{1 - (.50)^2}} = \sqrt{.3733} = .61$$

This multiple correlation coefficient is not significantly different from .60, the correlation between the scholastic aptitude test and the criterion. The efficiency of prediction of the multiple R is 21%, as compared with 20% for r_{cx}. That the addition of the y variable, the "social intelligence" test scores, makes no appreciable difference in the predictive efficiency of the battery as a whole may be somewhat surprising.

Ideally, an effective battery of tests for predicting a criterion such as academic or vocational success would be composed of several tests, each of which correlates fairly high with the criterion but hardly at all with the others. The greater the intercorrelation of the variables in a battery of tests, the smaller the increase in the predictive efficiency of the battery as a whole.

* See Table V, Appendix II, for values of $1 - r^2$.

Predicting Clerical Efficiency from Two Variables

We shall now present a three-variable multiple correlation problem in which the predictive efficiency of two test variables combined is appreciably greater than when either is taken alone. The criterion variable was a measure of clerical proficiency based upon ratings over a period of several months. The test variables were results on a number-name-checking clerical test (x), and a vocabulary-arithmetic test of the omnibus type (y), administered as part of a battery for predicting clerical aptitude. The following correlations were obtained for these three variables:

$$r_{cx} = .40 \text{ (predictive efficiency} = 8\%)$$
$$r_{cy} = .50 \text{ (predictive efficiency} = 13\%)$$
$$r_{xy} = .10$$

The multiple R between the two test variables and the criterion is .61, computed as follows:

$$R_{c \cdot xy} = \sqrt{\frac{(.40)^2 + (.50)^2 - 2(.40)(.50)(.10)}{1 - (.10)^2}} = \sqrt{.3737} = .61$$

Whereas the higher of the correlations between the test variables and the criterion is .50(r_{cy}), the multiple R is .61. For r_{cx} the efficiency of prediction is 8%, and for r_{cy} it is 13%, whereas for $R_{c \cdot xy}$ it is 21% (Table 16:2). Thus, the efficiency of prediction of the combined x and y variables is nearly twice as great as it is for variable (y) alone, which correlated .50 with the criterion. The two tests combined are more efficient because the correlation between them was only .10, i.e., little more than zero, and each had a fair correlation with the criterion.

The Multiple Regression Equation and the Standard Error of Estimate for R

A multiple correlation coefficient not only is valuable for determining the efficiency of prediction of a battery of tests, but is also of considerable significance in providing an empirical basis for appropriately weighting each test. In fact, if the battery is to have the predictive efficiency signified by R, each test must be weighted on the basis of the multiple correlation. Thus, if a battery is composed of two tests and Test x has twice the predictive efficiency of Test y, then x should be given twice as much weight as y in the total score for the battery. The weights required can be directly obtained from the regression equation of the multiple correlation coefficient, provided the equation is expressed in z score form. For a three-variable problem, such as that in the preceding examples, the regression equation is as follows:

[17:7]

$$z_c = \frac{r_{cx} - r_{cy}r_{xy}}{1 - r_{xy}^2} z_x + \frac{r_{cy} - r_{cx}r_{xy}}{1 - r_{xy}^2} z_y$$

Multiple regression equation for a three-variable problem

For the preceding example, this is:

$$\begin{aligned} z_c &= \frac{.40 - (.50)(.10)}{1 - (.10)^2} z_x + \frac{.50 - (.40)(.10)}{1 - (.10)^2} z_y \\ &= .35z_x + .46z_y \end{aligned}$$

The regression coefficient for the x variable is .35, as compared with .46 for the y variable. Therefore, to obtain the most efficient single total score for a battery consisting of these two tests, scores on Test x should be given about three-fourths as much weight as scores on Test y, since .35/.46 = 3/4.

The error of prediction, however, must also be considered. The standard error of estimate for R is the same as the standard error of estimate for the correlation between two variables:

[17 : 8]

$$\sigma_{R_{c \cdot 1,2,3,4 \ldots n}} = \sigma_c \sqrt{1 - R_{c \cdot 1,2,3,4 \ldots n}^2}$$

Standard error of estimate of a multiple correlation coefficient

where $R_{c \cdot 1,2,3,4 \ldots n}$ is the multiple correlation between any number of variables and the criterion, c. When R is .61, as in the preceding examples, the standard error of estimate is 79% as large as the standard deviation of the distribution of criterion scores, σ_c. Hence the efficiency of prediction is 21%. (See Table 16:2.)

F. PARTIAL CORRELATION

Partial correlation is a statistical technique which, in appropriate circumstances, can yield information otherwise obtainable only by experimental methods. The technique is based upon the assumption that the effect of a variable on a bi-variate relationship can be held constant. This assumption implies that the variable represents a relatively unitary function or trait. Because such an assumption holds only rarely for psychological test variables, the use of the technique is limited. But there is one variable, *age*, which has been assumed to be legitimately subject to the technique. Psychologically, the age of individuals provides an index, but a very rough one of maturity.

Partial Correlation with Scholastic Aptitude Held Constant

To illustrate the technique of partial correlation for a three-variable problem in which one variable is to be held constant, we shall use the multiple R of the scholastic aptitude scores (x) and "social intelligence" test scores (y) with the grade criteria of academic success (c). These three variables were correlated as follows:

$$r_{cx} = .60 \qquad r_{xy} = .50$$

$$r_{cy} = .40 \qquad R_{c \cdot xy} = .61$$

We saw above that the addition of the y variable did not appreciably increase the predictive efficiency. We can reverse the procedure and, by partial correlation, estimate the correlation between the "social intelligence" test scores

(y) and academic success, (r_{cy}), with the effect of the scholastic aptitude test results (x) held constant. We can estimate what it would be from the following formula: *

$$r_{cy \cdot x} = \frac{r_{cy} - r_{cx}r_{xy}}{\sqrt{1 - r_{cx}^2}\sqrt{1 - r_{xy}^2}} \qquad [17:9]$$

Partial correlation coefficient for a three-variable problem

The correlation between academic success and social intelligence, with the scholastic aptitude variable held constant, is as follows:

$$r_{cy \cdot x} = \frac{.40 - (.60)(.50)}{\sqrt{1 - (.60)^2}\sqrt{1 - (.50)^2}} = \frac{.100}{.693} = .14$$

The correlation between the social intelligence test results and academic success is reduced from .40 to .14 when the effect of the scholastic aptitude test results is held constant. This was of course expected in view of the multiple correlational analysis already made. The partial correlation coefficient shows that the residue of factors in the "social intelligence" test, beyond those involved in the relationship between the scholastic aptitude test and the criterion, amounts to very little. Social intelligence may have some relation to academic success, but this particular test, as standardized and scored, shows that the relationship is not of much consequence.

Partial Correlation with Age Held Constant

That there is a relationship between test achievement and age is well known. Unfortunately, however, this fact is sometimes neglected in the correlations between psychological variables that are reported. Consider, for example, a correlation of .50 between vocabulary test scores (variable x) and arithmetic test scores (variable y) for a sample of boys and girls ranging in age from 9 to 12 years. At least part of the correlation is doubtless attributable to this heterogeneity in age. That is, because of their greater maturity the older boys and girls should accomplish more than the younger children. These differences in age should therefore increase the correlation.

The basic problem is that of estimating the correlation between vocabulary and arithmetic ability if the sample were all the same age. There are two approaches to this problem. The better procedure consists in using samples that are more homogeneous in age. Thus, four or five samples might be used in order to show the correlation between vocabulary and arithmetic ability for 12-year-olds, for 11-year-olds, for 10-year-olds, etc. When this approach is inconvenient, as it often is, an estimate of the correlation, if all the children had been of the same age, can be obtained by means of partial correlation.

* The terms in the denominator are coefficients of alienation, k. Hence Table V, Appendix B, considerably simplifies the computations.

In other words, an estimate of the correlation between x (vocabulary test) and y (arithmetic test) can be obtained with a (age) held constant. The data are as follows:

r_{xy} = .50 (correlation between vocabulary test scores and arithmetic test scores for the heterogeneous age group)

r_{xa} = .65 (correlation between vocabulary test scores and age) [of subjects]

r_{ya} = .55 (correlation between arithmetic test scores and age)

When age is not held constant, the correlation between x and y is .50. But when it is held constant, the correlation is considerably reduced:

$$r_{xy \cdot a} = \frac{r_{xy} - r_{xa}r_{ya}}{\sqrt{1 - r_{xa}^2}\sqrt{1 - r_{ya}^2}} = \frac{(.50) - (.65)(.55)}{\sqrt{1 - (.65)^2}\sqrt{1 - (.55)^2}} = .22$$

Thus, the correlation between vocabulary and arithmetic ability would more likely have been about .20 instead of .50 if the subjects had been of the same age.

Spurious Correlation

The difference between partial correlation coefficients and original correlation coefficients (.22 and .50, respectively, in the preceding example) has sometimes been held to illustrate the difference between non-spurious and spurious correlation. However, to call the correlation of .50 spurious is misleading. Instead, the effect of age on the result should be ascertained and the partial coefficient interpreted accordingly. After all, age has a very real effect on the correlation between vocabulary and arithmetic ability; there is nothing spurious about it. The real problem involves determining the effect of such a factor, and for this the technique of partial correlation provides a statistical short-cut. The introduction of experimental techniques in the original design of an investigation gives a sounder basis for determining the role of otherwise "hidden" factors that may make the correlation coefficient between two variables higher than it otherwise would be.

EXERCISES

1. What statistical considerations enter into the evaluation of a psychological test?
2. Define test reliability and describe the methods used to determine it, indicating the relative advantages and disadvantages of each.
3. Define test validity, distinguish between the operational and functional validity of a test, and describe methods used to obtain indexes of validity.
4. How does a critical score replace a regression equation in the use of test results for predicting success and failure?
5. Describe how a test may have more than one index of validity.

6. What is the effect of the range of ability of the sample on (a) the reliability coefficient, and (b) an index of validity?
7. What statistical techniques are used in test item analysis? Describe three situations in which different statistical techniques are employed.
8. What does the multiple correlation coefficient measure, and how is multiple correlation utilized in evaluating a battery of tests?
9. What is a partial correlation coefficient? Of what value is the technique of partial correlation in research?

CHAPTER 18

Cluster and Factor Analysis

A. THEORY OF THE ORGANIZATION OF HUMAN TRAITS

Statistical methods of cluster or factor analysis represent a significant development during the past several decades in the appraisal and evaluation of psychometric procedures. We have seen that multiple correlation makes it possible to determine the predictive efficiency of a battery of tests and to weight each test on the basis of its efficiency for predicting a criterion. Factor analysis, on the other hand, gives a basis for insight into the organizational role of the traits, abilities, etc., which enter into performance on a series of tests. This is not to say that multiple correlation is not useful for evaluating a test battery; on the contrary, a most significant step in evaluating any test of ability or aptitude is the determination of its functional validity in the light of an independent empirical criterion. From an empirical point of view, the latter is just as important as a factor analysis of the abilities operating in a battery of tests. Nevertheless, a picture of the way in which abilities are organized provides valuable insights for the construction and administration of tests.

The Coefficient of Determination (r^2)

The principle underlying factor analysis is the association of component factors in two or more correlated variables. For example, when the correlation of two variables is significantly greater than zero, the non-chance factors that account for the correlation are common to both. Thus, the correlation between weight and height is accounted for by factors of organic development *common* to both of these variables. Similarly, the correlation between scholastic aptitude and academic achievement is accounted for by factors of intellectual development which are common to both of these two variables.

Correlations significantly greater than zero can be described as *causal* relationships provided there is a logical basis for the association. Causation itself, however, is complex. For one thing, the determining factors underlying a causal relationship are rarely, if ever, simple. Furthermore, the factors accountable for a causal relation between two variables, x and y, may operate mainly in one direction, or in several directions: (1) y may be a function of x; (2) x may be a function of y; (3) both may be a function, reciprocally, of each other; or (4) both may be a function of a third set of factors. The fourth is illustrated by the correlation between vocabulary and arithmetic ability of children heterogeneous in age, referred to in Chapter 17.

The proportion or percentage of the variance (standard deviation squared) of one variable that is *associated with* the variance of another variable can be estimated from a coefficient that is usually called the *coefficient of determination*, and is the square of the correlation between the two variables, r_{xy}^2.* If the correlation between scholastic aptitude (x) and academic achievement (y) for a given universe is .60, then $(.60)^2$ or .36, is the proportion of the variance of x that is *associated with* the variance of y. If academic achievement could be logically assumed to be a function of scholastic aptitude, and not vice versa, then 36% of the variance characteristic of academic achievement would be determined by factors measured by the scholastic aptitude test. Generally, however, such an assumption of causality from one variable to another is hazardous. Consequently, it is preferable to interpret a coefficient of determination as measuring the proportion or percentage of factors that the two variables have in common. Thus, there is a perfect correlation between the circumferences and diameters of circles; r^2 therefore equals 1.00, and 100% of the variance of either variable is associated with, or a function of, the variance of the other. Both circumference and diameter are properties of "circularity"; neither circumference nor diameter is the causal determinant of the other. The association is invariant, i.e., r is 1.00, but both variables are "determined" by the character of the whole of which they are integral aspects or properties.

The Coefficient of Non-Determination

The proportion or percentage of the variance of one variable not accounted for by a second variable that is correlated with it to some degree is measured by k^2, the coefficient of non-determination. This is the square of k, the coefficient of alienation.†

Spearman's Two-Factor Theory

Interest in how psychological abilities may be organized dates back to Spearman's early work in England.‡ Spearman advanced a theory which came to be known as the two-factor theory, and which held that all human abilities are basically dependent upon (1) a factor of general mental energy and (2) abilities specific to each kind of task situation. The general factor he symbolized as G, and the second, being pluralistic, he symbolized as s, for *specific* factors. He developed the thesis that the extent to which an individual manifests the general factor, G, is a function of his heredity, whereas the specific factors, s_1, s_2, s_3 . . . s_n, represent his specific acquisition of learning and experience.

* See Table V, Appendix B, for r^2 for given values of r.

† See Table V, Appendix B, for $1 - r^2$, or k^2, for given values of r.

‡ C. Spearman, "The Proof and Measurement of Association Between Two Things," *American Journal of Psychology*, 15:72–101, 1904. See also his *The Abilities of Man*, Macmillan, New York, 1927, and *Psychology Through the Ages*, Macmillan, New York, 1938.

Spearman defined intelligence in terms of *G*, and consequently encouraged the development of intelligence tests which would differentiate an individual's abilities with respect to his *G* capacity. He believed that the following three psychological functions were most directly indicative of *G*: (1) introspective capacity; (2) the eduction of relations, and (3) the eduction of correlates. He further contended that the so-called general intelligence tests measure the latter two functions fairly well but do not adequately measure the first. In papers presented over a period of years, Spearman and his students attempted to educe empirical evidence in support of the two-factor theory.

Multiple-Factor Theories

As additional techniques of statistical analysis were developed in the United States by Kelley,* Hotelling,† and Thurstone,‡ Spearman's two-factor theory was found to be increasingly unsatisfactory as an explanation of the organization of human abilities. In fact, Spearman himself, in later works, recognized that a unitary *G* factor would not entirely account for the way in which human abilities manifest themselves. The outcome of his own work, as well as that of American investigators, was the development of a multiple-factor theory which postulated *group factors* in addition to a possible general factor and specific factors. Group factors are psychological functions common to a number of behavior situations but not to all. The general factor, *G* (we prefer the term *common factors*), represents psychological functions common to all situations that demand mental activities. The specific factors, *s*, represent psychological functions peculiar to a particular situation. Whether or not these three types of factors represent innate or acquired abilities is of course beside the point here, for the present discussion is concerned only with the organization and interrelation of psychological functions or "factors."

Sampling Theory and Cluster Analysis

Thomson § and Tryon,‖ following the lead of E. L. Thorndike, have contended that the fundamental functions or factors underlying human behavior are practically infinite in number and relatively independent of each other. The problem in connection with a group of tests is to determine how this myriad of factors is sampled, or drawn upon, in the functioning of behavior. *Cluster analysis*, a statistical technique devised by Tryon, enables

* T. L. Kelley, *Cross Roads in the Mind of Man*, Stanford Univ. Press, Stanford University, 1928.

† H. Hotelling, "Analysis of a Complex of Statistical Variables into Principal Components," *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

‡ L. L. Thurstone, *Vectors of the Mind*, Univ. of Chicago Press, Chicago, 1935; and also, *Primary Mental Abilities*, Psychometrika Monograph, 1938.

§ G. H. Thomson, *The Factorial Analysis of Human Ability*, Houghton Mifflin, Boston, 1939.

‖ R. C. Tryon, *Cluster Analysis. Correlation Profile and Orthometric (Factor) Analysis for the Isolation of Unities in Mind and Personality*, Edwards Bros., Ann Arbor, 1939.

the investigator to determine how psychological functions such as abilities, attitudes, etc., manifest themselves and are interrelated and organized.

Once the facts concerning the organization of psychological functions are ascertained, the theoretical and practical implications for the development and use of psychological tests will be clear. However, not all the facts are known,* and therefore factorial analysis should be viewed as a useful statistical technique for the evaluation and appraisal of test batteries as they are developed.

These various theories of mental organization have practical implications for testing procedures. Thus, if Spearman's two-factor theory were adequate, psychologists should concentrate on developing a battery of tests that would give as reliable and valid a differentiation of *G* as is possible. For example, some investigators have used the I.Q. as if it were the answer.

Thurstone speaks of *primary* mental factors such as memory, word relations, number relations, etc. A well-rounded test battery designed to measure such functions should include at least one reliable test for measuring each one. Furthermore, by the theory of multiple factors, the results of each test in such a battery would not be pooled in order to obtain a composite single score but would be kept separate so as to yield a measure for each function. Nevertheless, there is still the question of the *practical* implications of such a battery for educational and occupational situations. As has been indicated, the practical value must be determined through the evaluation of test results in relation to independent criteria of proficiency or success in the life situation.

B. METHODS OF FACTOR ANALYSIS

The methods of factor analysis which have been used generally in the United States during the past decade are principally those devised by Thurstone † and Hotelling.‡ However, they require complicated mathematical computations and are beyond the scope of this book. Fortunately, *correlation profile analysis*, a relatively simple technique developed by Tryon,§ makes it possible to determine whether the intercorrelations between a group of variables can be explained more satisfactorily in terms of a set of functions common to all of them or in terms of two or more sets of functions (group factors), or whether neither of these explanations is adequate.

Tryon's Method of Correlation Profile Analysis

Tryon's approach to the organization of human abilities, traits, etc., introduces such concepts as "components," "operational unities," and "clusters." He emphasizes that all a research worker can achieve is to "discover general

* Cf. E. E. Cureton, "The Principal Compulsions of Factor-Analysts," *Harvard Educational Review*, 9:287–295, 1939.

† L. L. Thurstone, *op. cit.* ‡ H. Hotelling, *op. cit.* § R. C. Tryon, *op. cit.*

components which act *as if* they were common determiners in different behaviors. Such common determiners are called operational unities. These are defined as those component factors which operate when two or more variables show the same pattern of correlation coefficients with all the other variables of a group. Two variables, *A* and *B*, are said to be wholly or partially determined by an operational unity if both correlate highly with variable *M*, low with *N*, intermediate with *O*, and so on throughout the other variables. In such a case, clearly what is common to *A* is common to *B*, since they are behaving in an identical and unitary fashion. Correlation profile analysis is a simple method for discovering and grouping together variables which have identical patterns or profiles of correlations." *

Cluster Analysis of Body Measurements

The intercorrelation coefficients between the following 12 variables, measures of body dimensions, will be used to illustrate Tryon's correlation profile analysis: †

A. Waist height
B. Hip height
C. Weight
D. Stature
E. Cervical height
F. Tibial height
G. Bitrochanteric diameter
H. Waist girth
I. Hip girth
J. Upper arm girth
K. Posterior arm length
L. Thigh girth

Body measurements for each of these 12 variables were obtained from 32,165 boys aged 4 to 14, and from 31,919 girls of a similar age range; only the data for the boys are presented here. At least part of the correlation between any two of these variables will be due to the sample's heterogeneity in age. Since the correlation of each variable with age was included in the original article, we have been able to hold constant the effect of age by means of partial correlation.

The 66 intercorrelations of these 12 variables are presented in Table 18:1. The intercorrelations above the diagonal are the coefficients presented in the original article, with the effect of age heterogeneity included; those below the diagonal are the partial coefficients with the effect of age held constant.

An examination of this table may not be particularly informative at first. It will be noted, however, that the coefficients below the diagonal are generally lower than those above the diagonal. Furthermore, closer inspection of the latter coefficients reveals that the correlations between measurement and age (the first row) are generally lower than the intercorrelations among the 12 variables themselves, and that all the correlations are *positive* and fairly

* *Ibid.*, p. 2, n.

† The data for this analysis were obtained from a study, "Children's Body Measurements for Sizing Garments and Patterns," made by R. O'Brien and M. A. Girschick, U. S. Department of Agriculture, Miscellaneous Publications No. 36, 1939.

Table 18:1. Intercorrelations of Body Measurements of a Sample of 32,165 Boys, Age 4 to 14

(Original Correlation Coefficients Above the Diagonal; Partial Correlation Coefficients, with the Effect of Age Held Constant, Below the Diagonal)

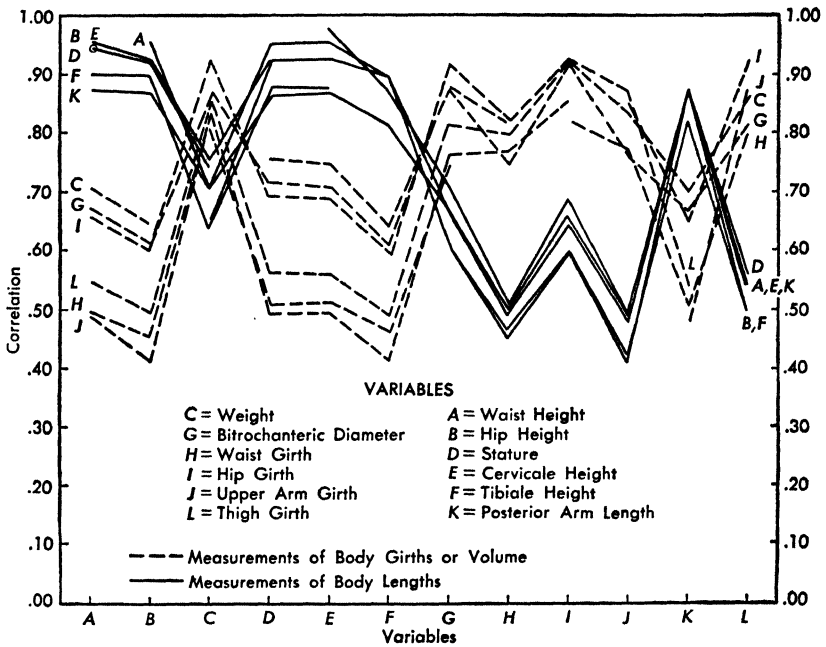
Variables	A	B	C	D	E	F	G	H	I	J	K	L
Age	.899	.900	.822	.897	.898	.879	.826	.690	.807	.688	.887	.736
A. Waist height		.991	.915	.990	.991	.977	.909	.777	.896	.766	.974	.824
B. Hip height	.952		.901	.985	.986	.978	.894	.764	.881	.750	.973	.808
C. Weight	.707	.649		.927	.926	.897	.961	.903	.975	.912	.914	.937
D. Stature	.948	.922	.754		.996	.972	.918	.782	.904	.776	.973	.828
E. Cervical height	.953	.927	.749	.979		.973	.918	.783	.904	.775	.974	.828
F. Tibial height	.895	.899	.640	.872	.876		.890	.767	.877	.750	.960	.806
G. Bitrochanteric diameter	.672	.614	.878	.711	.710	.610		.875	.973	.880	.907	.921
H. Waist girth	.495	.452	.816	.509	.512	.464	.748		.909	.879	.782	.900
I. Hip girth	.660	.603	.928	.690	.688	.596	.919	.824		.922	.893	.964
J. Upper arm girth	.492	.414	.838	.495	.492	.419	.763	.770	.855		.771	.936
K. Posterior arm length	.876	.871	.703	.868	.872	.818	.669	.509	.648	.480		.821
L. Thigh girth	.549	.495	.860	.562	.560	.492	.819	.800	.925	.876	.546	

high. Tryon's correlation profile analysis will give a graphic picture of the interrelationships of these variables. We shall of course use the partial coefficients below the diagonal.

The Correlation Profile

Fig. 18:1 is a correlation profile which brings together the relationships among the 12 body measurements. The graph is constructed by plotting in succession the correlation of each variable with all the other variables. The

Fig. 18:1. Correlation Profiles for Twelve Body Measurement Variables. (Data from Table 18:1)



value of each coefficient is plotted on the ordinate, and the 12 variables, *A* to *L*, are scaled in equally spaced intervals on the abscissa. There are thus 12 line graphs in the figure, one for each variable. Each line graph is continuous, except where the variable is correlated with itself; self-correlation is the reliability coefficient and is not shown.*

* Tryon's method is developed on the assumption that the reliability of the method used for measuring each variable is high, and that at least one of the correlations of each variable with the others is significantly greater than zero. In the intercorrelations in Fig. 18:1, all the coefficients are significantly greater than zero.

Ordinarily, in correlation profile analysis, the profiles of all the variables are not drawn on a single graph; rather, a systematic method for analyzing the table of intercorrelations is used which permits the investigator to plot on separate graphs each group of variables most likely to cluster together.* However, when only 12 variables are used, the correlation profile of each can be drawn on one graph. Fig. 18:1 shows which of the variables, if any, form one or more operational unities.

The implications of the correlation profile in this figure are rather clear. It will be observed that the following six variables (shown by solid lines in the figure) have similar correlation profiles, and hence constitute an operational unity which we shall call Cluster I:

Cluster I

A. Waist height
B. Hip height
D. Stature
E. Cervical height
F. Tibial height
K. Posterior arm length

The trend of the correlations between each of these six variables and all of the variables is similar; that is, all the line graphs for these six rise and fall together.

The remaining six variables also have correlation profiles similar to each other, as shown by the broken lines in the figure; we shall call these six Cluster II.

Cluster II

C. Weight
G. Bitrochanteric diameter
H. Waist girth
I. Hip girth
J. Upper arm girth
L. Thigh girth

The curves of these six variables rise and fall together, even though the actual coefficients are not as similar as were those in Cluster I. However, the six correlation profiles for Cluster II are not only similar but, in contrast to those in Cluster I, have a different pattern and hence provide the basis for a second operational unity.

Examination of the variables in Cluster I reveals that all are measures of *length*, whereas all those in Cluster II are measures of *volume* or *girth*. This empirical result supports the hypothesis that at least two significantly different physical dimensions—length and volume or girth—must be taken into account in measurements of body build.

---

* R. C. Tryon, *op. cit.*, pp. 4–8.

The line graphs in Fig. 18:1 are thus evidence for the existence of two major operational unities among the 12 variables. But other implications are evident from further inspection of these profiles. For example, the correlation profiles of Cluster I are more congruent than are those of Cluster II. This is consistent, of course, with the information on body development and body build revealed by many independent investigations: Growing children, as well as adults, are more variable with respect to measures of volume and girth than to measures of length. That the six volume or girth variables signify greater relative variability among themselves than the six length variables is evidenced by the greater range of their intercorrelations with any single variable.

The correlation profiles also suggest (although they do not necessarily demonstrate) which variable should provide the best single measure of each cluster. In each case, they will be the variables which are at the peak in their respective correlation profiles. The variable which will provide the best measure of the length functions measured by Cluster I as a whole is *A* (waist height), *B* (hip height), *D* (stature), or *E* (cervical height). From only an inspection of the correlation profiles, there is not much basis for selecting one of these four, for any of them apparently represents the cluster as well as the other three. The average of the intercorrelations of each of these four variables with the remaining five in the cluster is about the same, approximately .90. The one best variable could be determined by means of more complicated mathematical methods; * but from a practical point of view, the variable that can be most readily and reliably obtained would be chosen. Since *stature* is the most practical of the four for ordinary measurement, it would be the one to be used.

The choice among the six variables in Cluster II would lie between *C* (weight) and *I* (hip girth). These two have the highest average intercorrelations with the other variables in the cluster, although *L* (thigh girth) also is fairly high, with an average intercorrelation of about .85. A mathematical factor analysis will reveal whether there are any significant differences in the predictive efficiency of one of these variables over the others; but since weight is most readily attainable in ordinary situations, it would be chosen.

The application of correlation profile analysis to the intercorrelations of the 12 variables thus not only reveals two significantly different operational unities or clusters, but also suggests that from a practical point of view stature and weight best represent each operational unity in Clusters I and II respectively.

The intercorrelations have a further implication, viz., that all 12 variables measure certain factors or functions common to all 12. This is evidenced by the fact that all the intercorrelation coefficients are positive and fairly high.

---

* Tryon's orthometric analysis serves this purpose; such an analysis could also be made by other methods of factor analysis.

An inspection of Fig. 18:1 indicates that, on the average, the six variables in Cluster I have correlations of from about .50 to about .70 with the six variables in Cluster II, and that the six variables in Cluster II have average correlations of from about .50 to about .65 with the six variables in Cluster I. In other words, the implication is clear that there is a communality of functions for all 12 variables; this is consistent with the fact that all of them represent measures of body build. Although two important physical dimensions are differentiable, nevertheless organic unity underlies the interrelations of all these variables.

In summary, the correlation profile analysis indicates that the various factors underlying these measurements are interrelated and organized as follows:

Common Factors: Factors or functions *common* to all 12 measurements of body build.
Length Factors: Cluster I—factors or functions common to all the measures of *length*, but not to measures of volume or girth.
Volume Factors: Cluster II—factors or functions common to all measures of *volume* or girth, but not to measures of length.
Specific Factors: Factors or functions common to only a particular variable but not to any other.

With respect to the organization of mental functions rather than body measures, the common factors would be somewhat analogous to Spearman's *G* factor; the length and volume factors would be analogous to the group factors of multiple-factor theory; and the specific factors would be analogous to those referred to by Spearman. However, this analysis is not based on the rigid preconceptions of any theory; rather, it represents an empirical approach to the problem of functional organization.

Cluster Analysis of Psychological Variables

It will be well to illustrate the application of cluster analysis to a series of psychological variables, for the results will not be so unambiguous in their implications as were those obtained with the body measurements discussed above. For this purpose intercorrelations between ten achievement tests administered to 1046 Bucknell College sophomores will be used.

The data presented in Table 18:2 represent the interrelationships of student abilities on a comprehensive survey of achievements in general culture, science, English, and mathematics. The reliability of each test was satisfactory; the lowest was .90 (for the grammar test), the highest .986 (for mathematics). Interpretation of the intercorrelations in this table is somewhat more complicated than was true of the body measurement data in Table 18:1. The achievement test intercorrelations include a few negative coefficients and many correlations not significantly greater than zero. The punctuation test and the grammar test have the highest coefficient, .742.

Table 18:2. The Intercorrelations Between the Scores of 1046 Bucknell College Sophomores on 10 Achievement Tests *

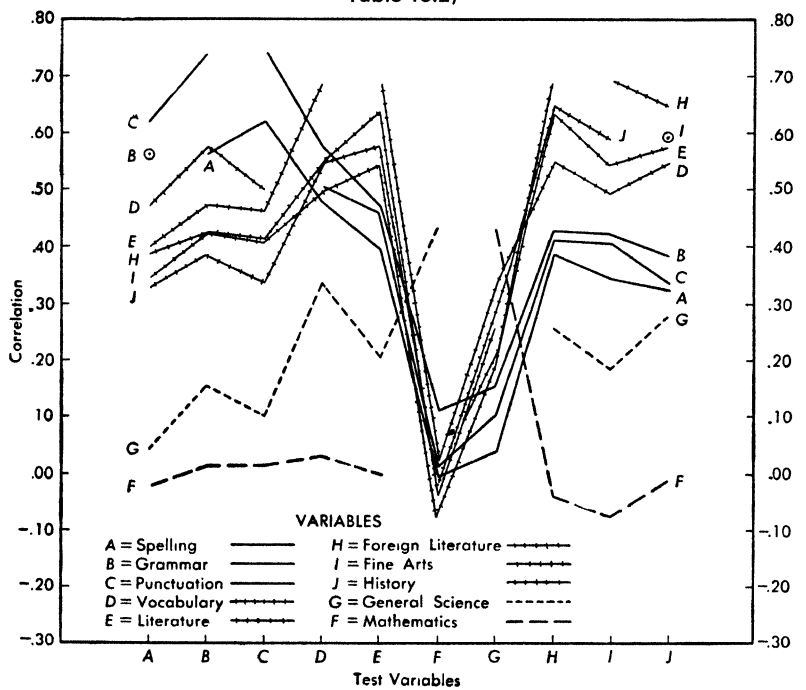
	A	B	C	D	E	F	G	H	I	J
A. Spelling	*.932*	.564	.621	.476	.394	−.022	.044	.389	.344	.328
B. Grammar	.564	*.904*	.742	.577	.472	.013	.158	.429	.426	.383
C. Punctuation	.621	.742	*.907*	.503	.461	.014	.102	.411	.407	.339
D. Vocabulary	.476	.577	.503	*.966*	.688	.030	.334	.548	.494	.546
E. Literature	.394	.472	.461	.688	*.961*	.002	.202	.639	.541	.574
F. Mathematics	−.022	.013	.014	.030	.002	*.986*	.430	−.035	−.075	−.012
G. General science	.044	.159	.102	.334	.202	.430	*.943*	.258	.183	.276
H. Foreign literature	.389	.429	.411	.548	.639	−.035	.258	*.915*	.691	.646
I. Fine arts	.344	.426	.407	.494	.541	−.075	.183	.691	*.920*	.589
J. History	.328	.383	.339	.546	.574	−.012	.276	.646	.589	*.948*

* J. C. Flanagan, *Factor Analysis in the Study of Personality*, Stanford Univ. Press, 1935. Data originally published by C. C. Brigham in *A Study of Error*, College Entrance Examinations Board, New York, 1932.

The reliability coefficient of each variable is given in **italics** on the diagonal, beginning with $A = A$.

The chief implication to be drawn from this table is perhaps that no important set of psychological functions is common to all 10 variables. In other words, the slightly negative and zero correlation coefficients suggest the absence of Spearman's G function or of any other defined communality for the battery of tests as a whole. This table, even more than the preceding table, emphasizes the need for statistical techniques which will enable the investigator to ascertain whether the abilities represented by a group of variables have anything in common—whether they are interrelated and organized.

Fig. 18:2. Correlation Profiles for Ten Achievement Test Variables. (Data from Table 18:2)



Correlation profiles for each of these 10 variables are presented in Fig. 18:2. Although the interpretation of these results is more complicated than was true of those in Fig. 18:1, close scrutiny should yield relevant hypotheses about the psychological functions involved. Thus, several fairly distinct operational unities appear to be characteristic of the 10 variables. The first consists of A (spelling), B (grammar), and C (punctuation). This group can be called Cluster I. A second possible operational unity apparently includes H (foreign literature), I (fine arts), J (history), and possibly D (vocabulary) and E (literature). These will be called Cluster II. There remain Variables F

(mathematics) and *G* (general science). Since there is not much congruence between their profiles, they will be considered independent variables. If the test battery had included several tests of mathematical proficiency, they might have yielded congruent correlation profiles and would therefore have formed a "mathematics cluster." Presumably the same would be true if there had been several tests of science ability.

In summary: the correlation profiles in Fig. 18:2 suggest the following organization of mental functions, so far as they were sampled by the 10 achievement tests:

Practical English Usage Factors: Cluster I—Psychological factors or functions common to *A* (spelling), *B* (grammar), and *C* (punctuation); an operational unity composed of three tests which evidently sample *practical English usage.*

Literature Factors: Cluster II—Psychological factors or functions common to *D* (vocabulary), *E* (literature), *H* (foreign literature), *I* (fine arts), and *J* (history); an operational unity composed of five tests which evidently sample an *understanding and appreciation of literature.*

General Science Factors: Psychological factors or functions common to *G* (general science); a relatively independent test which evidently samples *general science information.*

Mathematics Factors: Psychological factors or functions common to *F* (mathematical ability); a relatively independent test which evidently samples *mathematics ability.*

Specific Factors *S*: Psychological factors or functions common to a particular variable but not to any of the others—*specific factors.* (The correlations are far from perfect.)

As already indicated, these results yield no evidence whatsoever for the existence of Spearman's *G* factor or of any other factors or functions common to the battery as a whole. The correlation profile of the mathematics variable (*F*) is especially important as evidence in support of this point. The mathematics test results had practically a zero correlation with all the others except general science (*G*), and even here the correlation was only .43. On the other hand, the *trend* of the correlation profile of the *G* variable was, in some respects, similar to that of the other variables in Cluster II. Hence the result suggests that ability in general science is composed of abilities in part common to the mathematics variable and in part common to the other variables in Cluster II. Such an inference is not contrary to a common-sense appraisal of the characteristic content of elementary general science courses and the manifold abilities called for.

From this cluster analysis the following practical implication for testing procedures should be clear. It should be more useful to summarize an individual's performance on the battery by differentiating his scores into clusters and independent variables than by pooling them to obtain a composite single score for the battery as a whole. Pooling them would give ambiguous indices of achievement because a single score would fail to reveal his relatively proficient and relatively non-proficient areas. Furthermore, pooling is clearly not

warranted because the correlation profiles in Fig. 18:2 give no evidence of the existence of any important factor or set of factors common to the battery as a whole.

Some General Implications of Factor Analysis

The chief contribution to psychological measurement resulting from factor analyses of psychological variables during the past several decades has been the development of a body of knowledge and theories concerning the organization and interrelations of mental abilities and other attributes. Such theories are no longer based on rational considerations alone; they are fortified by empirical data. Factor analyses of many variables derived from human behavior have provided an empirical foundation for test procedures.

Although some of the earlier theories of mental organization have been demonstrated to be inadequate, no theory that is generally acceptable has yet been established. However, it is well established that many people have unusual capacities and attainments in some respects but not in others. In other words, there are important ability and personality differences within the average person as well as between persons. Recognition of these distinctions is essential to the appraisal of the abilities and aptitudes of individuals as a basis for adequate counseling, guidance, and placement.

EXERCISES

1. Define the coefficient of determination. What does it measure?
2. What is the usefulness of cluster or factor analysis in psychological research?
3. What fundamental principle underlies the usefulness of cluster or factor analysis?
4. Set up a hypothetical battery of ten test variables and describe the kind of result that you would need to obtain from an intercorrelational analysis which would support: (a) Spearman's two-factor theory, (b) the theory of group factors.
5. Set up correlation profiles for the data in Table 18:3, and interpret the results.

Table 18:3. Intercorrelations Between the Scores of 108 Nine-Year-Old Boys on Eleven Psychological Tests *

	A	B	C	D	E	F	G	H	I	J
A. Motor speed		.304	.371	.269	.395	.185	.218	.342	.159	.254
B. Vocabulary	.304		.400	.239	.523	.237	.160	.339	.056	.284
C. Arithmetic	.371	.400		.462	.489	.273	.306	.318	.292	.217
D. Paper form board	.269	.239	.462		.248	.214	.140	.310	.289	.185
E. Logical prose memory	.395	.523	.489	.248		.337	.379	.457	.233	.425
F. Word-word memory	.185	.237	.273	.214	.337		.388	.288	.148	.290
G. Word retention	.218	.160	.306	.140	.379	.388		.307	.203	.364
H. Digit span	.342	.339	.318	.310	.457	.288	.307		.267	.207
I. Geometric form memory	.159	.056	.292	.289	.233	.148	.203	.267		.129
J. Object memory	.254	.284	.217	.185	.425	.290	.364	.207	.129	

*From H. E. Garrett, A. I. Bryan, and R. E. Perl, "The Age Factor in Mental Organization," *Archives of Psychology*, No. 176, 1935, Table 8, p. 19.

APPENDIX A

Bibliography of Statistical Tables and Nomographs, Periodical Literature, and Chief References in Mathematical and Advanced Statistics

STATISTICAL TABLES AND NOMOGRAPHS

Buros, O. K. (ed.), *The Second Yearbook of Research and Statistical Methodology*, The Gryphon Press, Highland Park, New Jersey, 1941.

Chesire, L., Saffir, M., and Thurstone, L. L., *Computing Diagrams for the Tetrachoric Correlation Coefficient*, University of Chicago Bookstore, Chicago, 1933.

Dunlap, J. W., and Kurtz, A. K., *Handbook of Statistical Nomographs, Tables, and Formulas*, World Book Company, Yonkers, 1932.

Fisher, R. A., and Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver & Boyd, London, 1938.

Kendall, M. G., and Smith, B. B., "Randomness and Random Sampling Numbers," *Journal of the Royal Statistical Society*, **101**:147–166, 1938.

Kurtz, A. K., and Edgerton, H. A., *Statistical Dictionary of Terms and Symbols*, Wiley, New York, 1939.

Pearson, Karl, *Tables for Statisticians and Biometricians*, Cambridge University Press, Cambridge, 1914.

PERIODICALS AND GOVERNMENT PUBLICATIONS

Biometrika: A Journal for the Statistical Study of Biological Problems. Egon S. Pearson, Editor, University College, London.

Journal of the American Statistical Association. Lester S. Kellogg, Managing Editor, 1603 K Street, N.W., Washington 6, D. C.

Journal of Applied Psychology. Jack Dunlap, Editor, University of Rochester, Rochester, New York.

Journal of Educational Research. A. S. Barr, Chairman of Editorial Board, University of Wisconsin, Madison 6, Wisconsin.

Journal of the Royal Statistical Society. 4 Portugal Street, W.C. 2, London.

National Education Association, Publications. Washington, D. C.

Psychometrika: A Journal Devoted to the Development of Psychology as a Quantitative Rational Science. H. O. Gulliksen, Managing Editor, Princeton, New Jersey.

Public Opinion Quarterly. E. F. Goldman, Editor, Princeton University Press, Princeton, New Jersey.

U. S. Bureau of the Census, Publications. Washington, D. C.

U. S. Office of Education, Publications. Washington, D. C.

U. S. Public Health Service, Publications. Washington, D. C.

MATHEMATICAL STATISTICS AND ADVANCED STATISTICAL METHODS

Ezekiel, Mordecai, *Methods of Correlation Analysis*, John Wiley & Sons, New York, 2nd ed., 1941.

Fisher, R. A., *The Design of Experiments*, Oliver & Boyd, London, 2nd ed., 1937.

Fisher, R. A., *Statistical Methods for Research Workers*, Oliver & Boyd, London, 7th ed., 1938.

Kelley, T. L., *Statistical Method*, Macmillan, New York, 1923.

Kenney, J. F., *Mathematics of Statistics*, 2 vols., Van Nostrand, New York, 1939.

Peters, C. C., and Van Voorhis, W. R., *Statistical Procedures and Their Mathematical Bases*, McGraw-Hill, New York, 1940.

Smith, J. G., and Duncan, A. J., *Sampling Statistics and Applications*, McGraw-Hill, New York, 1945.

Thomson, G. H., *The Factorial Analysis of Human Ability*, Houghton Mifflin, Boston, 1939.

Yule, G. U., and Kendall, M. G., *An Introduction to the Theory of Statistics*, Griffin, London, 12th ed., 1940.

APPENDIX B

Tables of Statistical Functions

Table

I. Areas and Ordinates of the Normal Probability Curve 508

IA. Ordinate Values of the Normal Curve Expressed as Proportions of the Ordinate at the Mean 511

II. Probability Values of T for Normal Sampling Distributions of Large Sample Theory 512

III. Distribution of t for Small Samples 514

IV. Distribution of Chi-Square 515

V. Values of Functions of r 516

VI. Values of Fisher's z Function for Given Values of Pearson's r 518

VII. Values of Proportions p and q 519

Table I

AREAS AND ORDINATES OF THE NORMAL PROBABILITY CURVE

(In Terms of x/σ Units and a Total Area (a) Equal to 1.0)

Example: .4066 (or 40.66%) of the total area of the normal probability curve lies between the mean and a point 1.32 standard deviations units above or below the mean; i.e., $x/\sigma = 1.32$. The proportionate value of the ordinate, y, at x/σ of 1.32 is .1669.

$\frac{x}{\sigma}$	Area (a)	Ordinate y	$\frac{x}{\sigma}$	Area (a)	Ordinate y	$\frac{x}{\sigma}$	Area (a)	Ordinate y
.00	.0000	.3989	.40	.1554	.3683	.80	.2881	.2897
.01	.0040	.3989	.41	.1591	.3668	.81	.2910	.2874
.02	.0080	.3989	.42	.1628	.3653	.82	.2939	.2850
.03	.0120	.3988	.43	.1664	.3637	.83	.2967	.2827
.04	.0160	.3986	.44	.1700	.3621	.84	.2996	.2803
.05	.0199	.3984	.45	.1736	.3605	.85	.3023	.2780
.06	.0239	.3982	.46	.1772	.3589	.86	.3051	.2756
.07	.0279	.3980	.47	.1808	.3572	.87	.3078	.2732
.08	.0319	.3977	.48	.1844	.3555	.88	.3106	.2709
.09	.0359	.3973	.49	.1879	.3538	.89	.3133	.2685
.10	.0398	.3970	.50	.1915	.3521	.90	.3159	.2661
.11	.0438	.3965	.51	.1950	.3503	.91	.3186	.2637
.12	.0478	.3961	.52	.1985	.3485	.92	.3212	.2613
.13	.0517	.3956	.53	.2019	.3467	.93	.3238	.2589
.14	.0557	.3951	.54	.2054	.3448	.94	.3264	.2565
.15	.0596	.3945	.55	.2088	.3429	.95	.3289	.2541
.16	.0636	.3939	.56	.2123	.3410	.96	.3315	.2516
.17	.0675	.3932	.57	.2157	.3391	.97	.3340	.2492
.18	.0714	.3925	.58	.2190	.3372	.98	.3365	.2468
.19	.0754	.3918	.59	.2224	.3352	.99	.3389	.2444
.20	.0793	.3910	.60	.2258	.3332	1.00	.3413	.2420
.21	.0832	.3902	.61	.2291	.3312	1.01	.3438	.2396
.22	.0871	.3894	.62	.2324	.3292	1.02	.3461	.2371
.23	.0910	.3885	.63	.2357	.3271	1.03	.3485	.2347
.24	.0948	.3876	.64	.2389	.3251	1.04	.3508	.2323
.25	.0987	.3867	.65	.2422	.3230	1.05	.3531	.2299
.26	.1026	.3857	.66	.2454	.3209	1.06	.3554	.2275
.27	.1064	.3847	.67	.2486	.3187	1.07	.3577	.2251
.28	.1103	.3836	.68	.2518	.3166	1.08	.3599	.2227
.29	.1141	.3825	.69	.2549	.3144	1.09	.3621	.2203
.30	.1179	.3814	.70	.2580	.3123	1.10	.3643	.2179
.31	.1217	.3802	.71	.2612	.3101	1.11	.3665	.2155
.32	.1255	.3790	.72	.2642	.3079	1.12	.3686	.2131
.33	.1293	.3778	.73	.2673	.3056	1.13	.3708	.2107
.34	.1331	.3765	.74	.2704	.3034	1.14	.3729	.2083
.35	.1368	.3752	.75	.2734	.3011	1.15	.3749	.2059
.36	.1406	.3739	.76	.2764	.2989	1.16	.3770	.2036
.37	.1443	.3725	.77	.2794	.2966	1.17	.3790	.2012
.38	.1480	.3712	.78	.2823	.2943	1.18	.3810	.1989
.39	.1517	.3697	.79	.2852	.2920	1.19	.3830	.1965

Table I (*continued*)

$\frac{x}{\sigma}$	Area (a)	Ordinate y	$\frac{x}{\sigma}$	Area (a)	Ordinate y	$\frac{x}{\sigma}$	Area (a)	Ordinate y
1.20	.3849	.1942	1.70	.4554	.0940	2.20	.4861	.0355
1.21	.3869	.1919	1.71	.4564	.0925	2.21	.4864	.0347
1.22	.3888	.1895	1.72	.4573	.0909	2.22	.4868	.0339
1.23	.3907	.1872	1.73	.4582	.0893	2.23	.4871	.0332
1.24	.3925	.1849	1.74	.4591	.0878	2.24	.4875	.0325
1.25	.3944	.1826	1.75	.4599	.0863	2.25	.4878	.0317
1.26	.3962	.1804	1.76	.4608	.0848	2.26	.4881	.0310
1.27	.3980	.1781	1.77	.4616	.0833	2.27	.4884	.0303
1.28	.3997	.1758	1.78	.4625	.0818	2.28	.4887	.0297
1.29	.4015	.1736	1.79	.4633	.0804	2.29	.4890	.0290
1.30	.4032	.1714	1.80	.4641	.0790	2.30	.4893	.0283
1.31	.4049	.1691	1.81	.4649	.0775	2.31	.4896	.0277
1.32	.4066	.1669	1.82	.4656	.0761	2.32	.4898	.0270
1.33	.4082	.1647	1.83	.4664	.0748	2.33	.4901	.0264
1.34	.4099	.1626	1.84	.4671	.0734	2.34	.4904	.0258
1.35	.4115	.1604	1.85	.4678	.0721	2.35	.4906	.0252
1.36	.4131	.1582	1.86	.4686	.0707	2.36	.4909	.0246
1.37	.4147	.1561	1.87	.4693	.0694	2.37	.4911	.0241
1.38	.4162	.1539	1.88	.4700	.0681	2.38	.4913	.0235
1.39	.4177	.1518	1.89	.4706	.0669	2.39	.4916	.0229
1.40	.4192	.1497	1.90	.4713	.0656	2.40	.4918	.0224
1.41	.4207	.1476	1.91	.4719	.0644	2.41	.4920	.0219
1.42	.4222	.1456	1.92	.4726	.0632	2.42	.4922	.0213
1.43	.4236	.1435	1.93	.4732	.0620	2.43	.4925	.0208
1.44	.4251	.1415	1.94	.4738	.0608	2.44	.4927	.0203
1.45	.4265	.1394	1.95	.4744	.0596	2.45	.4929	.0198
1.46	.4279	.1374	1.96	.4750	.0584	2.46	.4931	.0194
1.47	.4292	.1354	1.97	.4756	.0573	2.47	.4932	.0189
1.48	.4306	.1334	1.98	.4762	.0562	2.48	.4934	.0184
1.49	.4319	.1315	1.99	.4767	.0551	2.49	.4936	.0180
1.50	.4332	.1295	2.00	.4772	.0540	2.50	.4938	.0175
1.51	.4345	.1276	2.01	.4778	.0529	2.51	.4940	.0171
1.52	.4357	.1257	2.02	.4783	.0519	2.52	.4941	.0167
1.53	.4370	.1238	2.03	.4788	.0508	2.53	.4943	.0163
1.54	.4382	.1219	2.04	.4793	.0498	2.54	.4945	.0158
1.55	.4394	.1200	2.05	.4798	.0488	2.55	.4946	.0154
1.56	.4406	.1182	2.06	.4803	.0478	2.56	.4948	.0151
1.57	.4418	.1163	2.07	.4808	.0468	2.57	.4949	.0147
1.58	.4430	.1145	2.08	.4812	.0459	2.58	.4951	.0143
1.59	.4441	.1127	2.09	.4817	.0449	2.59	.4952	.0139
1.60	.4452	.1109	2.10	.4821	.0440	2.60	.4953	.0136
1.61	.4463	.1092	2.11	.4826	.0431	2.61	.4955	.0132
1.62	.4474	.1074	2.12	.4830	.0422	2.62	.4956	.0129
1.63	.4484	.1057	2.13	.4834	.0413	2.63	.4957	.0126
1.64	.4495	.1040	2.14	.4838	.0404	2.64	.4959	.0122
1.65	.4505	.1023	2.15	.4842	.0396	2.65	.4960	.0119
1.66	.4515	.1006	2.16	.4846	.0387	2.66	.4961	.0116
1.67	.4525	.0989	2.17	.4850	.0379	2.67	.4962	.0113
1.68	.4535	.0973	2.18	.4854	.0371	2.68	.4963	.0110
1.69	.4545	.0957	2.19	.4857	.0363	2.69	.4964	.0107

Table I (*continued*)

$\frac{x}{\sigma}$	Area (a)	Ordinate y	$\frac{x}{\sigma}$	Area (a)	Ordinate y	$\frac{x}{\sigma}$	Area (a)	Ordinate y
2.70	.4965	.0104	2.80	.4974	.0079	2.90	.4981	.0060
2.71	.4966	.0101	2.81	.4975	.0077	2.91	.4982	.0058
2.72	.4967	.0099	2.82	.4976	.0075	2.92	.4982	.0056
2.73	.4968	.0096	2.83	.4977	.0073	2.93	.4983	.0055
2.74	.4969	.0093	2.84	.4977	.0071	2.94	.4984	.0053
2.75	.4970	.0091	2.85	.4978	.0069	2.95	.4984	.0051
2.76	.4971	.0088	2.86	.4979	.0067	2.96	.4985	.0050
2.77	.4972	.0086	2.87	.4980	.0065	2.97	.4985	.0048
2.78	.4973	.0084	2.88	.4980	.0063	2.98	.4986	.0047
2.79	.4974	.0081	2.89	.4981	.0061	2.99	.4986	.0046
						3.00	.49865	.0044
						3.50	.49977	.0009
						4.00	.49997	.0001
						4.50	.499997	.00002
						5.00	.4999997	.000002

Table IA

ORDINATE VALUES OF THE NORMAL, BELL-SHAPED PROBABILITY CURVE, EXPRESSED AS PROPORTIONS OF THE ORDINATE AT THE MEAN

Thus: The height of the mean ordinate is taken as 1.000. An ordinate point value 2σ above or below the mean is .135 as high as it is at the mean. The mean ordinate for a finite distribution is: $y_M = N_s/2.51\sigma$. See page 432.

$\frac{x}{\sigma}$	0	1	2	3	4	5	6	7	8	9
0.0	1.000	.999 +	.999 +	.999 +	.999 +	.999	.998	.998	.997	.996
0.1	.995	.994	.993	.992	.990	.989	.987	.986	.984	.982
0.2	.980	.978	.976	.974	.972	.969	.967	.964	.962	.959
0.3	.956	.953	.950	.947	.944	.941	.937	.934	.930	.927
0.4	.923	.918	.916	.912	.908	.904	.900	.895	.891	.887
0.5	.882	.878	.874	.869	.864	.860	.855	.850	.845	.841
0.6	.835	.830	.825	.820	.815	.810	.804	.799	.794	.788
0.7	.783	.777	.772	.766	.760	.755	.749	.743	.738	.732
0.8	.726	.720	.714	.709	.703	.697	.691	.685	.679	.673
0.9	.667	.661	.655	.649	.643	.637	.631	.625	.619	.613
1.0	.607	.600	.594	.588	.582	.576	.570	.564	.558	.552
1.1	.546	.540	.534	.528	.522	.516	.510	.504	.498	.493
1.2	.487	.481	.475	.469	.464	.458	.452	.446	.441	.435
1.3	.430	.424	.418	.413	.407	.402	.397	.391	.386	.381
1.4	.375	.370	.365	.360	.355	.350	.344	.339	.334	.330
1.5	.325	.320	.315	.310	.306	.301	.296	.292	.287	.283
1.6	.278	.274	.269	.265	.261	.256	.252	.248	.244	.240
1.7	.236	.232	.228	.224	.220	.216	.213	.209	.205	.201
1.8	.198	.194	.191	.187	.184	.181	.177	.174	.171	.168
1.9	.164	.161	.158	.155	.152	.149	.146	.144	.141	.138
2.0	.135	.133	.130	.127	.125	.122	.120	.117	.115	.113
2.1	.110	.108	.106	.103	.101	.099	.097	.095	.093	.091
2.2	.089	.087	.085	.083	.081	.080	.078	.076	.074	.073
2.3	.071	.069	.068	.066	.065	.063	.062	.060	.059	.058
2.4	.056	.055	.054	.052	.051	.050	.049	.047	.046	.045
2.5	.044	.043	.042	.041	.040	.039	.038	.037	.036	.035
2.6	.034	.033	.032	.031	.031	.030	.029	.028	.028	.027
2.7	.026	.025	.025	.024	.023	.023	.022	.022	.021	.020
2.8	.020	.019	.019	.018	.018	.017	.017	.016	.016	.015
2.9	.015	.014	.014	.014	.013	.013	.013	.012	.012	.011
3.0	.011									
4.0	.0003									
5.0	.00000									

Table II

PROBABILITY VALUES FOR T OF NORMAL SAMPLING DISTRIBUTIONS OF LARGE SAMPLE THEORY

Example: If T, the test ratio of a Test of Significance, $(s - h)/\sigma_{s}$, is 2.0, the P (probability) value is .0228 for a result equal to or larger than (or less than, depending on which tail of the sampling distribution is involved) the sample value of the statistic (s).

T	P	T	P	T	P	T	P
.00	.5000	.45	.3264	.90	.1841	1.35	.0885
.01	.4960	.46	.3228	.91	.1814	1.36	.0869
.02	.4920	.47	.3192	.92	.1788	1.37	.0853
.03	.4880	.48	.3156	.93	.1762	1.38	.0838
.04	.4840	.49	.3121	.94	.1736	1.39	.0823
.05	.4801	.50	.3085	.95	.1711	1.40	.0808
.06	.4761	.51	.3050	.96	.1685	1.41	.0793
.07	.4721	.52	.3015	.97	.1660	1.42	.0778
.08	.4681	.53	.2981	.98	.1635	1.43	.0764
.09	.4641	.54	.2946	.99	.1611	1.44	.0749
.10	.4602	.55	.2912	1.00	.1587	1.45	.0735
.11	.4562	.56	.2877	1.01	.1562	1.46	.0721
.12	.4522	.57	.2843	1.02	.1539	1.47	.0708
.13	.4483	.58	.2810	1.03	.1515	1.48	.0694
.14	.4443	.59	.2776	1.04	.1492	1.49	.0681
.15	.4404	.60	.2742	1.05	.1469	1.50	.0668
.16	.4364	.61	.2709	1.06	.1446	1.51	.0655
.17	.4325	.62	.2676	1.07	.1423	1.52	.0643
.18	.4286	.63	.2643	1.08	.1401	1.53	.0630
.19	.4246	.64	.2611	1.09	.1379	1.54	.0618
.20	.4207	.65	.2578	1.10	.1357	1.55	.0606
.21	.4168	.66	.2546	1.11	.1335	1.56	.0594
.22	.4129	.67	.2514	1.12	.1314	1.57	.0582
.23	.4090	.68	.2482	1.13	.1292	1.58	.0570
.24	.4052	.69	.2451	1.14	.1271	1.59	.0559
.25	.4013	.70	.2420	1.15	.1251	1.60	.0548
.26	.3974	.71	.2388	1.16	.1230	1.61	.0537
.27	.3936	.72	.2358	1.17	.1210	1.62	.0526
.28	.3897	.73	.2327	1.18	.1190	1.63	.0516
.29	.3859	.74	.2296	1.19	.1170	1.64	.0505
.30	.3821	.75	.2266	1.20	.1151	1.65	.0495
.31	.3783	.76	.2236	1.21	.1131	1.66	.0485
.32	.3745	.77	.2206	1.22	.1112	1.67	.0475
.33	.3707	.78	.2177	1.23	.1093	1.68	.0465
.34	.3669	.79	.2148	1.24	.1075	1.69	.0455
.35	.3632	.80	.2119	1.25	.1056	1.70	.0446
.36	.3594	.81	.2090	1.26	.1038	1.71	.0436
.37	.3557	.82	.2061	1.27	.1020	1.72	.0427
.38	.3520	.83	.2033	1.28	.1003	1.73	.0418
.39	.3483	.84	.2004	1.29	.0985	1.74	.0409
.40	.3446	.85	.1977	1.30	.0968	1.75	.0401
.41	.3409	.86	.1949	1.31	.0951	1.76	.0392
.42	.3372	.87	.1922	1.32	.0934	1.77	.0384
.43	.3336	.88	.1894	1.33	.0918	1.78	.0375
.44	.3300	.89	.1867	1.34	.0901	1.79	.0367

Table II (*continued*)

T	P	T	P	T	P	T	P
1.80	.0359	2.10	.0179	2.40	.0082	2.70	.0035
1.81	.0351	2.11	.0174	2.41	.0080	2.71	.0034
1.82	.0344	2.12	.0170	2.42	.0078	2.72	.0033
1.83	.0336	2.13	.0166	2.43	.0075	2.73	.0032
1.84	.0329	2.14	.0162	2.44	.0073	2.74	.0031
1.85	.0322	2.15	.0158	2.45	.0071	2.75	.0030
1.86	.0314	2.16	.0154	2.46	.0069	2.76	.0029
1.87	.0307	2.17	.0150	2.47	.0068	2.77	.0028
1.88	.0300	2.18	.0146	2.48	.0066	2.78	.0027
1.89	.0294	2.19	.0143	2.49	.0064	2.79	.0026
1.90	.0287	2.20	.0139	2.50	.0062	2.80	.0026
1.91	.0281	2.21	.0136	2.51	.0060	2.81	.0025
1.92	.0274	2.22	.0132	2.52	.0059	2.82	.0024
1.93	.0268	2.23	.0129	2.53	.0057	2.83	.0023
1.94	.0262	2.24	.0125	2.54	.0055	2.84	.0023
1.95	.0256	2.25	.0122	2.55	.0054	2.85	.0022
1.96	.0250	2.26	.0119	2.56	.0052	2.86	.0021
1.97	.0244	2.27	.0116	2.57	.0051	2.87	.0020
1.98	.0238	2.28	.0113	2.58	.0049	2.88	.0020
1.99	.0233	2.29	.0110	2.59	.0048	2.89	.0019
2.00	.0228	2.30	.0107	2.60	.0047	2.90	.0019
2.01	.0222	2.31	.0104	2.61	.0045	2.91	.0018
2.02	.0217	2.32	.0102	2.62	.0044	2.92	.0018
2.03	.0212	2.33	.0099	2.63	.0043	2.93	.0017
2.04	.0207	2.34	.0096	2.64	.0041	2.94	.0016
2.05	.0202	2.35	.0094	2.65	.0040	2.95	.0016
2.06	.0197	2.36	.0091	2.66	.0039	2.96	.0015
2.07	.0192	2.37	.0089	2.67	.0038	2.97	.0015
2.08	.0188	2.38	.0087	2.68	.0037	2.98	.0014
2.09	.0183	2.39	.0084	2.69	.0036	2.99	.0014
						3.00	.00135

Table III

DISTRIBUTION OF t FOR TESTS OF SIGNIFICANCE OF SMALL SAMPLES *

$(N_s - 1)$	Probability: P					
	.5	.1	.05	.02	.01	.001
1	1.000	6.314	12.706	31.821	63.657	636.619
2	.816	2.920	4.303	6.965	9.925	31.598
3	.765	2.353	3.182	4.541	5.841	12.941
4	.741	2.132	2.776	3.747	4.604	8.610
5	.727	2.015	2.571	3.365	4.032	6.859
6	.718	1.943	2.447	3.143	3.707	5.959
7	.711	1.895	2.365	2.998	3.499	5.405
8	.706	1.860	2.306	2.896	3.355	5.041
9	.703	1.833	2.262	2.821	3.250	4.781
10	.700	1.812	2.228	2.764	3.169	4.587
11	.697	1.796	2.201	2.718	3.106	4.437
12	.695	1.782	2.179	2.681	3.055	4.318
13	.694	1.771	2.160	2.650	3.012	4.221
14	.692	1.761	2.145	2.624	2.977	4.140
15	.691	1.753	2.131	2.602	2.947	4.073
16	.690	1.746	2.120	2.583	2.921	4.015
17	.689	1.740	2.110	2.567	2.898	3.965
18	.688	1.734	2.101	2.552	2.878	3.922
19	.688	1.729	2.093	2.539	2.861	3.883
20	.687	1.725	2.086	2.528	2.845	3.850
21	.686	1.721	2.080	2.518	2.831	3.819
22	.686	1.717	2.074	2.508	2.819	3.792
23	.685	1.714	2.069	2.500	2.807	3.767
24	.685	1.711	2.064	2.492	2.797	3.745
25	.684	1.708	2.060	2.485	2.787	3.725
26	.684	1.706	2.056	2.479	2.779	3.707
27	.684	1.703	2.052	2.473	2.771	3.690
28	.683	1.701	2.048	2.467	2.763	3.674
29	.683	1.699	2.045	2.462	2.756	3.659
30	.683	1.697	2.042	2.457	2.750	3.646
40	.681	1.684	2.021	2.423	2.704	3.551
60	.679	1.671	2.000	2.390	2.660	3.460
120	.677	1.658	1.980	2.358	2.617	3.373
∞	.674	1.645	1.960	2.326	2.576	3.291

* Table III is abridged from Table III of Fisher: *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver & Boyd, Ltd., Edinburgh, by permission of the Author and Publishers.

Table IV

DISTRIBUTION OF CHI-SQUARE *

d.f.	Probability: P										
	.99	.95	.90	.50	.30	.20	.10	.05	.02	.01	.001
1	.00	.00	.02	.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	.02	.10	.21	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	.12	.35	.58	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	.30	.71	1.06	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	.55	1.14	1.61	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	.87	1.64	2.20	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	2.17	2.83	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.73	3.49	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	3.32	4.17	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.94	4.86	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	4.58	5.58	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	5.23	6.30	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	5.89	7.04	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	6.57	7.79	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	7.26	8.55	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	7.96	9.31	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.25
17	6.41	8.67	10.08	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.79
18	7.02	9.39	10.86	17.34	20.60	22.76	25.99	28.87	32.35	34.80	42.31
19	7.63	10.12	11.65	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	10.85	12.44	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	11.59	13.24	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	9.54	12.34	14.04	21.34	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	13.09	14.85	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	13.85	15.66	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	14.61	16.47	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	15.38	17.29	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	16.15	18.11	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	16.93	18.94	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	17.72	19.77	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	18.49	20.60	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

* Table IV is abridged from Table IV of Fisher: *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver & Boyd, Ltd., Edinburgh, by permission of the author and Publishers.

Table V

VALUES OF FUNCTIONS OF r*

r	$\sqrt{r}$	r^2	$\sqrt{r-r^2}$	$\sqrt{1-r}$	$1-r^2$	$\sqrt{1-r^2}$	$100(1-k)$	r
			$\sigma_{\infty \cdot 1}$	$\sigma_{(M)}$		k	% Eff.	
1.00	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	100.00	1.00
.99	.9950	.9801	.0995	.1000	.0199	.1411	85.89	.99
.98	.9899	.9604	.1400	.1414	.0396	.1990	80.10	.98
.97	.9849	.9409	.1706	.1732	.0591	.2431	75.69	.97
.96	.9798	.9216	.1960	.2000	.0784	.2800	72.00	.96
.95	.9747	.9025	.2179	.2236	.0975	.3122	68.78	.95
.94	.9695	.8836	.2375	.2449	.1164	.3412	65.88	.94
.93	.9644	.8649	.2551	.2646	.1351	.3676	63.24	.93
.92	.9592	.8464	.2713	.2828	.1536	.3919	60.81	.92
.91	.9539	.8281	.2862	.3000	.1719	.4146	58.54	.91
.90	.9487	.8100	.3000	.3162	.1900	.4359	56.41	.90
.89	.9434	.7921	.3129	.3317	.2079	.4560	54.40	.89
.88	.9381	.7744	.3250	.3464	.2256	.4750	52.50	.88
.87	.9327	.7569	.3363	.3606	.2431	.4931	50.69	.87
.86	.9274	.7396	.3470	.3742	.2604	.5103	48.97	.86
.85	.9220	.7225	.3571	.3873	.2775	.5268	47.32	.85
.84	.9165	.7056	.3666	.4000	.2944	.5426	45.74	.84
.83	.9110	.6889	.3756	.4123	.3111	.5578	44.22	.83
.82	.9055	.6724	.3842	.4243	.3276	.5724	42.76	.82
.81	.9000	.6561	.3923	.4359	.3439	.5864	41.36	.81
.80	.8944	.6400	.4000	.4472	.3600	.6000	40.00	.80
.79	.8888	.6241	.4073	.4583	.3759	.6131	38.69	.79
.78	.8832	.6084	.4142	.4690	.3916	.6258	37.42	.78
.77	.8775	.5929	.4208	.4796	.4071	.6380	36.20	.77
.76	.8718	.5776	.4271	.4899	.4224	.6499	35.01	.76
.75	.8660	.5625	.4330	.5000	.4375	.6614	33.86	.75
.74	.8602	.5476	.4386	.5099	.4524	.6726	32.74	.74
.73	.8544	.5329	.4440	.5196	.4671	.6834	31.66	.73
.72	.8485	.5184	.4490	.5292	.4816	.6940	30.60	.72
.71	.8426	.5041	.4538	.5385	.4959	.7042	29.58	.71
.70	.8367	.4900	.4583	.5477	.5100	.7141	28.59	.70
.69	.8307	.4761	.4625	.5568	.5239	.7238	27.62	.69
.68	.8246	.4624	.4665	.5657	.5376	.7332	26.68	.68
.67	.8185	.4489	.4702	.5745	.5511	.7424	25.76	.67
.66	.8124	.4356	.4737	.5831	.5644	.7513	24.87	.66
.65	.8062	.4225	.4770	.5916	.5775	.7599	24.01	.65
.64	.8000	.4096	.4800	.6000	.5904	.7684	23.16	.64
.63	.7937	.3969	.4828	.6083	.6031	.7766	22.34	.63
.62	.7874	.3844	.4854	.6164	.6156	.7846	21.54	.62
.61	.7810	.3721	.4877	.6245	.6279	.7924	20.76	.61
.60	.7746	.3600	.4899	.6325	.6400	.8000	20.00	.60
.59	.7681	.3481	.4918	.6403	.6519	.8074	19.26	.59
.58	.7616	.3364	.4936	.6481	.6636	.8146	18.54	.58
.57	.7550	.3249	.4951	.6557	.6751	.8216	17.84	.57
.56	.7483	.3136	.4964	.6633	.6864	.8285	17.15	.56
.55	.7416	.3025	.4975	.6708	.6975	.8352	16.48	.55
.54	.7348	.2916	.4984	.6782	.7084	.8417	15.83	.54
.53	.7280	.2809	.4991	.6856	.7191	.8480	15.20	.53
.52	.7211	.2704	.4996	.6928	.7296	.8542	14.58	.52
.51	.7141	.2601	.4999	.7000	.7399	.8602	13.98	.51
.50	.7071	.2500	.5000	.7071	.7500	.8660	13.40	.50

* From W. V. Bingham, *Aptitudes and Aptitude Testing*, Harper & Brothers, New York, 1937, Table XVIII.

Table V (*continued*)

r	$\sqrt{r}$	r^2	$\sqrt{r-r^2}$	$\sqrt{1-r}$	$1-r^2$	$\sqrt{1-r^2}$	$100(1-k)$	r
			$\sigma_{\infty \cdot 1}$	$\sigma_{(M)}$		k	% Eff.	
.50	.7071	.2500	.5000	.7071	.7500	.8660	13.40	.50
.49	.7000	.2401	.4999	.7141	.7599	.8717	12.83	.49
.48	.6928	.2304	.4996	.7211	.7696	.8773	12.27	.48
.47	.6856	.2209	.4991	.7280	.7791	.8827	11.73	.47
.46	.6782	.2116	.4984	.7348	.7884	.8879	11.21	.46
.45	.6708	.2025	.4975	.7416	.7975	.8930	10.70	.45
.44	.6633	.1936	.4964	.7483	.8064	.8980	10.20	.44
.43	.6557	.1849	.4951	.7550	.8151	.9028	9.72	.43
.42	.6481	.1764	.4936	.7616	.8236	.9075	9.25	.42
.41	.6403	.1681	.4918	.7681	.8319	.9121	8.79	.41
.40	.6325	.1600	.4899	.7746	.8400	.9165	8.35	.40
.39	.6245	.1521	.4877	.7810	.8479	.9208	7.92	.39
.38	.6164	.1444	.4854	.7874	.8556	.9250	7.50	.38
.37	.6083	.1369	.4828	.7937	.8631	.9290	7.10	.37
.36	.6000	.1296	.4800	.8000	.8704	.9330	6.70	.36
.35	.5916	.1225	.4770	.8062	.8775	.9367	6.33	.35
.34	.5831	.1156	.4737	.8124	.8844	.9404	5.96	.34
.33	.5745	.1089	.4702	.8185	.8911	.9440	5.60	.33
.32	.5657	.1024	.4665	.8246	.8976	.9474	5.25	.32
.31	.5568	.0961	.4625	.8307	.9039	.9507	4.93	.31
.30	.5477	.0900	.4583	.8367	.9100	.9539	4.61	.30
.29	.5385	.0841	.4538	.8426	.9159	.9570	4.30	.29
.28	.5292	.0784	.4490	.8485	.9216	.9600	4.00	.28
.27	.5196	.0729	.4440	.8544	.9271	.9629	3.71	.27
.26	.5099	.0676	.4386	.8602	.9324	.9656	3.44	.26
.25	.5000	.0625	.4330	.8660	.9375	.9682	3.18	.25
.24	.4899	.0576	.4271	.8718	.9424	.9708	2.92	.24
.23	.4796	.0529	.4208	.8775	.9471	.9732	2.68	.23
.22	.4690	.0484	.4142	.8832	.9516	.9755	2.45	.22
.21	.4583	.0441	.4073	.8888	.9559	.9777	2.23	.21
.20	.4472	.0400	.4000	.8944	.9600	.9798	2.02	.20
.19	.4359	.0361	.3923	.9000	.9639	.9818	1.82	.19
.18	.4243	.0324	.3842	.9055	.9676	.9837	1.63	.18
.17	.4123	.0289	.3756	.9110	.9711	.9854	1.46	.17
.16	.4000	.0256	.3666	.9165	.9744	.9871	1.29	.16
.15	.3873	.0225	.3571	.9220	.9775	.9887	1.13	.15
.14	.3742	.0196	.3470	.9274	.9804	.9902	.98	.14
.13	.3606	.0169	.3363	.9327	.9831	.9915	.85	.13
.12	.3464	.0144	.3250	.9381	.9856	.9928	.72	.12
.11	.3317	.0121	.3129	.9434	.9879	.9939	.61	.11
.10	.3162	.0100	.3000	.9487	.9900	.9950	.50	.10
.09	.3000	.0081	.2862	.9539	.9919	.9959	.41	.09
.08	.2828	.0064	.2713	.9592	.9936	.9968	.32	.08
.07	.2646	.0049	.2551	.9644	.9951	.9975	.25	.07
.06	.2449	.0036	.2375	.9695	.9964	.9982	.18	.06
.05	.2236	.0025	.2179	.9747	.9975	.9987	.13	.05
.04	.2000	.0016	.1960	.9798	.9984	.9992	.08	.04
.03	.1732	.0009	.1706	.9849	.9991	.9995	.05	.03
.02	.1414	.0004	.1400	.9899	.9996	.9998	.02	.02
.01	.1000	.0001	.0995	.9950	.9999	.9999	.01	.01
.00	.0000	.0000	.0000	1.0000	1.0000	1.0000	.00	.00

Table VI

VALUES OF FISHER'S z FUNCTION FOR GIVEN VALUES OF PEARSON'S *r**

r	z	r	z	r	z	r	z
.00	.00	.25	.26	.50	.55	.75	.97
.01	.01	.26	.27	.51	.56	.76	1.00
.02	.02	.27	.28	.52	.58	.77	1.02
.03	.03	.28	.29	.53	.59	.78	1.05
.04	.04	.29	.30	.54	.60	.79	1.07
.05	.05	.30	.31	.55	.62	.80	1.10
.06	.06	.31	.32	.56	.63	.81	1.13
.07	.07	.32	.33	.57	.65	.82	1.16
.08	.08	.33	.34	.58	.66	.83	1.19
.09	.09	.34	.35	.59	.68	.84	1.22
.10	.10	.35	.37	.60	.69	.85	1.26
.11	.11	.36	.38	.61	.71	.86	1.29
.12	.12	.37	.39	.62	.73	.87	1.33
.13	.13	.38	.40	.63	.74	.88	1.38
.14	.14	.39	.41	.64	.76	.89	1.42
.15	.15	.40	.42	.65	.78	.90	1.47
.16	.16	.41	.44	.66	.79	.91	1.53
.17	.17	.42	.45	.67	.81	.92	1.59
.18	.18	.43	.46	.68	.83	.93	1.66
.19	.19	.44	.47	.69	.85	.94	1.74
.20	.20	.45	.48	.70	.87	.95	1.83
.21	.21	.46	.50	.71	.89	.96	1.95
.22	.22	.47	.51	.72	.91	.97	2.09
.23	.23	.48	.52	.73	.93	.98	2.30
.24	.24	.49	.54	.74	.95	.99	2.65

* Table VI is adapted from Table VII of Fisher: *Statistical Tables for Biological, Agricultural and Medical Research*, Oliver & Boyd, Ltd., Edinburgh, by permission of the Author and Publishers.

Table VII

VALUES OF PROPORTIONS p AND q*

(Values Employed in the Determination of Biserial and Point-Biserial Correlations)

(1) p	(2) q	(3) pq	(4) $\frac{p}{y}$	(5) $\frac{pq}{y}$	(6) $\sqrt{pq}$	(7) $\sqrt{\frac{p}{q}}$
.01	.99	.0099	.3745	.3700	.0994	.1005
.02	.98	.0196	.4132	.3935	.1380	.1428
.03	.97	.0291	.4412	.4264	.1703	.1758
.04	.96	.0384	.4640	.4452	.1959	.2042
.05	.95	.0475	.4850	.4605	.2179	.2293
.06	.94	.0564	.5038	.4736	.2375	.2526
.07	.93	.0651	.5212	.4844	.2551	.2744
.08	.92	.0736	.5380	.4950	.2713	.2950
.09	.91	.0819	.5542	.5044	.2862	.3145
.10	.90	.0900	.5698	.5129	.3000	.3333
.11	.89	.0979	.5851	.5207	.3129	.3416
.12	.88	.1056	.6000	.5278	.3249	.3693
.13	.87	.1131	.6147	.5347	.3363	.3865
.14	.86	.1204	.6289	.5410	.3470	.4035
.15	.85	.1275	.6432	.5469	.3571	.4201
.16	.84	.1344	.6576	.5523	.3666	.4365
.17	.83	.1411	.6717	.5574	.3756	.4525
.18	.82	.1476	.6860	.5627	.3842	.4685
.19	.81	.1539	.7001	.5670	.3923	.4844
.20	.80	.1600	.7143	.5714	.4000	.5000
.21	.79	.1659	.7287	.5758	.4073	.5156
.22	.78	.1716	.7430	.5793	.4142	.5311
.23	.77	.1771	.7576	.5832	.4208	.5465
.24	.76	.1824	.7720	.5868	.4271	.5620
.25	.75	.1875	.7867	.5900	.4330	.5773
.26	.74	.1924	.8015	.5929	.4386	.5928
.27	.73	.1971	.8167	.5960	.4439	.6082
.28	.72	.2016	.8318	.5989	.4490	.6236
.29	.71	.2059	.8472	.6016	.4538	.6391
.30	.70	.2100	.8628	.6037	.4582	.6547
.31	.69	.2139	.8787	.6062	.4625	.6703
.32	.68	.2176	.8949	.6086	.4665	.6860
.33	.67	.2211	.9114	.6107	.4702	.7018
.34	.66	.2244	.9279	.6125	.4737	.7178
.35	.65	.2275	.9449	.6143	.4770	.7338
.36	.64	.2304	.9623	.6159	.4800	.7500
.37	.63	.2331	.9799	.6173	.4828	.7664
.38	.62	.2356	.9979	.6187	.4854	.7829
.39	.61	.2379	1.0164	.6200	.4877	.7996
.40	.60	.2400	1.0355	.6214	.4899	.8165
.41	.59	.2419	1.0548	.6222	.4918	.8336
.42	.58	.2436	1.0744	.6230	.4935	.8509
.43	.57	.2451	1.0947	.6241	.4951	.8686
.44	.56	.2464	1.1156	.6247	.4964	.8864
.45	.55	.2475	1.1369	.6254	.4975	.9045
.46	.54	.2484	1.1590	.6258	.4984	.9230
.47	.53	.2491	1.1815	.6262	.4991	.9417
.48	.52	.2496	1.2048	.6265	.4996	.9508
.49	.51	.2499	1.2287	.6266	.4999	.9802
.50	.50	.2500	1.2534	.6266	.5000	1.0000

* This table was developed by E. K. Taylor of the Adjutant General's Office, War Department, and is reproduced by permission.

APPENDIX C

Tables of Squares, Square Roots, Reciprocals, and Random Numbers

Table

I. Squares, Square Roots, and Reciprocals of Integers from 1 to 1000 522

II. A Table of Random Numbers 543

Table I

SQUARES, SQUARE ROOTS AND RECIPROCALS OF INTEGERS FROM 1 TO 1000

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
1	1	1.0000	1.000000	1.0000
2	4	1.4142	.500000	.7071
3	9	1.7321	.333333	.5774
4	16	2.0000	.250000	.5000
5	25	2.2361	.200000	.4472
6	36	2.4495	.166667	.4082
7	49	2.6458	.142857	.3780
8	64	2.8284	.125000	.3536
9	81	3.0000	.111111	.3333
10	100	3.1623	.100000	.3162
11	121	3.3166	.090909	.3015
12	144	3.4641	.083333	.2887
13	169	3.6056	.076923	.2774
14	196	3.7417	.071429	.2673
15	225	3.8730	.066667	.2582
16	256	4.0000	.062500	.2500
17	289	4.1231	.058824	.2425
18	324	4.2426	.055556	.2357
19	361	4.3589	.052632	.2294
20	400	4.4721	.050000	.2236
21	441	4.5826	.047619	.2182
22	484	4.6904	.045455	.2132
23	529	4.7958	.043478	.2085
24	576	4.8990	.041667	.2041
25	625	5.0000	.040000	.2000
26	676	5.0990	.038462	.1961
27	729	5.1962	.037037	.1925
28	784	5.2915	.035714	.1890
29	841	5.3852	.034483	.1857
30	900	5.4772	.033333	.1826
31	961	5.5678	.032258	.1796
32	1024	5.6569	.031250	.1768
33	1089	5.7446	.030303	.1741
34	1156	5.8310	.029412	.1715
35	1225	5.9161	.028571	.1690
36	1296	6.0000	.027778	.1667
37	1369	6.0828	.027027	.1644
38	1444	6.1644	.026316	.1622
39	1521	6.2450	.025641	.1601
40	1600	6.3246	.025000	.1581
41	1681	6.4031	.024390	.1562
42	1764	6.4807	.023810	.1543
43	1849	6.5574	.023256	.1525
44	1936	6.6332	.022727	.1508
45	2025	6.7082	.022222	.1491
46	2116	6.7823	.021739	.1474
47	2209	6.8557	.021277	.1459
48	2304	6.9282	.020833	.1443
49	2401	7.0000	.020408	.1429
50	2500	7.0711	.020000	.1414

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
51	2601	7.1414	.019608	.1400
52	2704	7.2111	.019231	.1387
53	2809	7.2801	.018868	.1374
54	2916	7.3485	.018519	.1361
55	3025	7.4162	.018182	.1348
56	3136	7.4833	.017857	.1336
57	3249	7.5498	.017544	.1325
58	3364	7.6158	.017241	.1313
59	3481	7.6811	.016949	.1302
60	3600	7.7460	.016667	.1291
61	3721	7.8102	.016393	.1280
62	3844	7.8740	.016129	.1270
63	3969	7.9373	.015873	.1260
64	4096	8.0000	.015625	.1250
65	4225	8.0623	.015385	.1240
66	4356	8.1240	.015152	.1231
67	4489	8.1854	.014925	.1222
68	4624	8.2462	.014706	.1213
69	4761	8.3066	.014493	.1204
70	4900	8.3666	.014286	.1195
71	5041	8.4261	.014085	.1187
72	5184	8.4853	.013889	.1179
73	5329	8.5440	.013699	.1170
74	5476	8.6023	.013514	.1162
75	5625	8.6603	.013333	.1155
76	5776	8.7178	.013158	.1147
77	5929	8.7750	.012987	.1140
78	6084	8.8318	.012821	.1132
79	6241	8.8882	.012658	.1125
80	6400	8.9443	.012500	.1118
81	6561	9.0000	.012346	.1111
82	6724	9.0554	.012195	.1104
83	6889	9.1104	.012048	.1098
84	7056	9.1652	.011905	.1091
85	7225	9.2195	.011765	.1085
86	7396	9.2736	.011628	.1078
87	7569	9.3274	.011494	.1072
88	7744	9.3808	.011364	.1066
89	7921	9.4340	.011236	.1060
90	8100	9.4868	.011111	.1054
91	8281	9.5394	.010989	.1048
92	8464	9.5917	.010870	.1043
93	8649	9.6437	.010753	.1037
94	8836	9.6954	.010638	.1031
95	9025	9.7468	.010526	.1026
96	9216	9.7980	.010417	.1021
97	9409	9.8489	.010309	.1015
98	9604	9.8995	.010204	.1010
99	9801	9.9499	.010101	.1005
100	10000	10.0000	.010000	.1000

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
101	10201	10.0499	.009901	.0995
102	10404	10.0995	.009804	.0990
103	10609	10.1489	.009709	.0985
104	10816	10.1980	.009615	.0981
105	11025	10.2470	.009524	.0976
106	11236	10.2956	.009434	.0971
107	11449	10.3441	.009346	.0967
108	11664	10.3923	.009259	.0962
109	11881	10.4403	.009174	.0958
110	12100	10.4881	.009091	.0953
111	12321	10.5357	.009009	.0949
112	12544	10.5830	.008929	.0945
113	12769	10.6301	.008850	.0941
114	12996	10.6771	.008772	.0937
115	13225	10.7238	.008696	.0933
116	13456	10.7703	.008621	.0928
117	13689	10.8167	.008547	.0925
118	13924	10.8628	.008475	.0921
119	14161	10.9087	.008403	.0917
120	14400	10.9545	.008333	.0913
121	14641	11.0000	.008264	.0909
122	14884	11.0454	.008197	.0905
123	15129	11.0905	.008130	.0902
124	15376	11.1355	.008065	.0898
125	15625	11.1803	.008000	.0894
126	15876	11.2250	.007937	.0891
127	16129	11.2694	.007874	.0887
128	16384	11.3137	.007813	.0884
129	16641	11.3578	.007752	.0880
130	16900	11.4018	.007692	.0877
131	17161	11.4455	.007634	.0874
132	17424	11.4891	.007576	.0870
133	17689	11.5326	.007519	.0867
134	17956	11.5758	.007463	.0864
135	18225	11.6190	.007407	.0861
136	18496	11.6619	.007353	.0857
137	18769	11.7047	.007299	.0854
138	19044	11.7473	.007246	.0851
139	19321	11.7898	.007194	.0848
140	19600	11.8322	.007143	.0845
141	19881	11.8743	.007092	.0842
142	20164	11.9164	.007042	.0839
143	20449	11.9583	.006993	.0836
144	20736	12.0000	.006944	.0833
145	21025	12.0416	.006897	.0830
146	21316	12.0830	.006849	.0828
147	21609	12.1244	.006803	.0825
148	21904	12.1655	.006757	.0822
149	22201	12.2066	.006711	.0819
150	22500	12.2474	.006667	.0816

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
151	22801	12.2882	.006623	.0814
152	23104	12.3288	.006579	.0811
153	23409	12.3693	.006536	.0808
154	23716	12.4097	.006494	.0806
155	24025	12.4499	.006452	.0803
156	24336	12.4900	.006410	.0801
157	24649	12.5300	.006369	.0798
158	24964	12.5698	.006329	.0796
159	25281	12.6095	.006289	.0793
160	25600	12.6491	.006250	.0791
161	25921	12.6886	.006211	.0788
162	26244	12.7279	.006173	.0786
163	26569	12.7671	.006135	.0783
164	26896	12.8062	.006098	.0781
165	27225	12.8452	.006061	.0778
166	27556	12.8841	.006024	.0776
167	27889	12.9228	.005988	.0774
168	28224	12.9615	.005952	.0772
169	28561	13.0000	.005917	.0769
170	28900	13.0384	.005882	.0767
171	29241	13.0767	.005848	.0765
172	29584	13.1149	.005814	.0762
173	29929	13.1529	.005780	.0760
174	30276	13.1909	.005747	.0758
175	30625	13.2288	.005714	.0756
176	30976	13.2665	.005682	.0754
177	31329	13.3041	.005650	.0752
178	31684	13.3417	.005618	.0750
179	32041	13.3791	.005587	.0747
180	32400	13.4164	.005556	.0745
181	32761	13.4536	.005525	.0743
182	33124	13.4907	.005495	.0741
183	33489	13.5277	.005464	.0739
184	33856	13.5647	.005435	.0737
185	34225	13.6015	.005405	.0735
186	34596	13.6382	.005376	.0733
187	34969	13.6748	.005348	.0731
188	35344	13.7113	.005319	.0729
189	35721	13.7477	.005291	.0727
190	36100	13.7840	.005263	.0725
191	36481	13.8203	.005236	.0724
192	36864	13.8564	.005208	.0722
193	37249	13.8924	.005181	.0720
194	37636	13.9284	.005155	.0718
195	38025	13.9642	.005128	.0716
196	38416	14.0000	.005102	.0714
197	38809	14.0357	.005076	.0712
198	39204	14.0712	.005051	.0711
199	39601	14.1067	.005025	.0709
200	40000	14.1421	.005000	.0707

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
201	40401	14.1774	.004975	.0705
202	40804	14.2127	.004950	.0704
203	41209	14.2478	.004926	.0702
204	41616	14.2829	.004902	.0700
205	42025	14.3178	.004878	.0698
206	42436	14.3527	.004854	.0697
207	42849	14.3875	.004831	.0695
208	43264	14.4222	.004808	.0693
209	43681	14.4568	.004785	.0692
210	44100	14.4914	.004762	.0690
211	44521	14.5258	.004739	.0688
212	44944	14.5602	.004717	.0687
213	45369	14.5945	.004695	.0685
214	45796	14.6287	.004673	.0684
215	46225	14.6629	.004651	.0682
216	46656	14.6969	.004630	.0680
217	47089	14.7309	.004608	.0679
218	47524	14.7648	.004587	.0677
219	47961	14.7986	.004566	.0676
220	48400	14.8324	.004545	.0674
221	48841	14.8661	.004525	.0673
222	49284	14.8997	.004505	.0671
223	49729	14.9332	.004484	.0670
224	50176	14.9666	.004464	.0668
225	50625	15.0000	.004444	.0667
226	51076	15.0333	.004425	.0665
227	51529	15.0665	.004405	.0664
228	51984	15.0997	.004386	.0662
229	52441	15.1327	.004367	.0661
230	52900	15.1658	.004348	.0659
231	53361	15.1987	.004329	.0658
232	53824	15.2315	.004310	.0657
233	54289	15.2643	.004292	.0655
234	54756	15.2971	.004274	.0654
235	55225	15.3297	.004255	.0652
236	55696	15.3623	.004237	.0651
237	56169	15.3948	.004219	.0650
238	56644	15.4272	.004202	.0648
239	57121	15.4596	.004184	.0647
240	57600	15.4919	.004167	.0645
241	58081	15.5242	.004149	.0644
242	58564	15.5563	.004132	.0643
243	59049	15.5885	.004115	.0642
244	59536	15.6205	.004098	.0640
245	60025	15.6525	.004082	.0639
246	60516	15.6844	.004065	.0638
247	61009	15.7162	.004049	.0636
248	61504	15.7480	.004032	.0635
249	62001	15.7797	.004016	.0634
250	62500	15.8114	.004000	.0632

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
251	63001	15.8430	.003984	.0631
252	63504	15.8745	.003968	.0630
253	64009	15.9060	.003953	.0629
254	64516	15.9374	.003937	.0627
255	65025	15.9687	.003922	.0626
256	65536	16.0000	.003906	.0625
257	66049	16.0312	.003891	.0624
258	66564	16.0624	.003876	.0623
259	67081	16.0935	.003861	.0621
260	67600	16.1245	.003846	.0620
261	68121	16.1555	.003831	.0619
262	68644	16.1864	.003817	.0618
263	69169	16.2173	.003802	.0617
264	69696	16.2481	.003788	.0615
265	70225	16.2788	.003774	.0614
266	70756	16.3095	.003759	.0613
267	71289	16.3401	.003745	.0612
268	71824	16.3707	.003731	.0611
269	72361	16.4012	.003717	.0610
270	72900	16.4317	.003704	.0609
271	73441	16.4621	.003690	.0607
272	73984	16.4924	.003676	.0606
273	74529	16.5227	.003663	.0605
274	75076	16.5529	.003650	.0604
275	75625	16.5831	.003636	.0603
276	76176	16.6132	.003623	.0602
277	76729	16.6433	.003610	.0601
278	77284	16.6733	.003597	.0600
279	77841	16.7033	.003584	.0599
280	78400	16.7332	.003571	.0598
281	78961	16.7631	.003559	.0597
282	79524	16.7929	.003546	.0595
283	80089	16.8226	.003534	.0594
284	80656	16.8523	.003521	.0593
285	81225	16.8819	.003509	.0592
286	81796	16.9115	.003497	.0591
287	82369	16.9411	.003484	.0590
288	82944	16.9706	.003472	.0589
289	83521	17.0000	.003460	.0588
290	84100	17.0294	.003448	.0587
291	84681	17.0587	.003436	.0586
292	85264	17.0880	.003425	.0585
293	85849	17.1172	.003413	.0584
294	86436	17.1464	.003401	.0583
295	87025	17.1756	.003390	.0582
296	87616	17.2047	.003378	.0581
297	88209	17.2337	.003367	.0580
298	88804	17.2627	.003356	.0579
299	89401	17.2916	.003344	.0578
300	90000	17.3205	.003333	.0577

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
301	90601	17.3494	.003322	.0576
302	91204	17.3781	.003311	.0575
303	91809	17.4069	.003300	.0574
304	92416	17.4356	.003289	.0574
305	93025	17.4642	.003279	.0573
306	93636	17.4929	.003268	.0572
307	94249	17.5214	.003257	.0571
308	94864	17.5499	.003247	.0570
309	95481	17.5784	.003236	.0569
310	96100	17.6068	.003226	.0568
311	96721	17.6352	.003215	.0567
312	97344	17.6635	.003205	.0566
313	97969	17.6918	.003195	.0565
314	98596	17.7200	.003185	.0564
315	99225	17.7482	.003175	.0563
316	99856	17.7764	.003165	.0563
317	100489	17.8045	.003155	.0562
318	101124	17.8326	.003145	.0561
319	101761	17.8606	.003135	.0560
320	102400	17.8885	.003125	0559
321	103041	17.9165	.003115	.0558
322	103684	17.9444	.003106	.0557
323	104329	17.9722	.003096	.0556
324	104976	18.0000	.003086	.0556
325	105625	18.0278	.003077	.0555
326	106276	18.0555	.003067	.0554
327	106929	18.0831	.003058	.0553
328	107584	18.1108	.003049	.0552
329	108241	18.1384	.003040	.0551
330	108900	18.1659	.003030	.0550
331	109561	18.1934	.003021	.0550
332	110224	18.2209	.003012	.0549
333	110889	18.2483	.003003	.0548
334	111556	18.2757	.002994	.0547
335	112225	18.3030	.002985	.0546
336	112896	18.3303	.002976	.0546
337	113569	18.3576	.002967	.0545
338	114244	18.3848	.002959	.0544
339	114921	18.4120	.002950	.0543
340	115600	18.4391	.002941	.0542
341	116281	18.4662	.002933	.0542
342	116964	18.4932	.002924	.0541
343	117649	18.5203	.002915	.0540
344	118336	18.5472	.002907	.0539
345	119025	18.5742	.002899	.0538
346	119716	18.6011	.002890	.0538
347	120409	18.6279	.002882	.0537
348	121104	18.6548	.002874	.0536
349	121801	18.6815	.002865	.0535
350	122500	18.7083	.002857	.0535

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
351	123201	18.7350	.002849	.0534
352	123904	18.7617	.002841	.0533
353	124609	18.7883	.002833	.0532
354	125316	18.8149	.002825	.0531
355	126025	18.8414	.002817	.0531
356	126736	18.8680	.002809	.0530
357	127449	18.8944	.002801	.0529
358	128164	18.9209	.002793	.0529
359	128881	18.9473	.002786	.0528
360	129600	18.9737	.002778	.0527
361	130321	19.0000	.002770	.0526
362	131044	19.0263	.002762	.0526
363	131769	19.0526	.002755	.0525
364	132496	19.0788	.002747	.0524
365	133225	19.1050	.002740	.0523
366	133956	19.1311	.002732	.0523
367	134689	19.1572	.002725	.0522
368	135424	19.1833	.002717	.0521
369	136161	19.2094	.002710	.0521
370	136900	19.2354	.002703	.0520
371	137641	19.2614	.002695	.0519
372	138384	19.2873	.002688	.0518
373	139129	19.3132	.002681	.0518
374	139876	19.3391	.002674	.0517
375	140625	19.3649	.002667	.0516
376	141376	19.3907	.002660	.0516
377	142129	19.4165	.002653	.0515
378	142884	19.4422	.002646	.0514
379	143641	19.4679	.002639	.0514
380	144400	19.4936	.002632	.0513
381	145161	19.5192	.002625	.0512
382	145924	19.5448	.002618	.0512
383	146689	19.5704	.002611	.0511
384	147456	19.5959	.002604	.0510
385	148225	19.6214	.002597	.0510
386	148996	19.6469	.002591	.0509
387	149769	19.6723	.002584	.0508
388	150544	19.6977	.002577	.0508
389	151321	19.7231	.002571	.0507
390	152100	19.7484	.002564	.0506
391	152881	19.7737	.002558	.0506
392	153664	19.7990	.002551	.0505
393	154449	19.8242	.002545	.0504
394	155236	19.8494	.002538	.0504
395	156025	19.8746	.002532	.0503
396	156816	19.8997	.002525	.0503
397	157609	19.9249	.002519	.0502
398	158404	19.9499	.002513	.0501
399	159201	19.9750	.002506	.0501
400	160000	20.0000	.002500	.0500

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
401	160801	20.0250	.002494	.0499
402	161604	20.0499	.002488	.0499
403	162409	20.0749	.002481	.0498
404	163216	20.0998	.002475	.0498
405	164025	20.1246	.002469	.0497
406	164836	20.1494	.002463	.0496
407	165649	20.1742	.002457	.0496
408	166464	20.1990	.002451	.0495
409	167281	20.2237	.002445	.0494
410	168100	20.2485	.002439	.0494
411	168921	20.2731	.002433	.0493
412	169744	20.2978	.002427	.0493
413	170569	20.3224	.002421	.0492
414	171396	20.3470	.002415	.0491
415	172225	20.3715	.002410	.0491
416	173056	20.3961	.002404	.0490
417	173889	20.4206	.002398	.0490
418	174724	20.4450	.002392	.0489
419	175561	20.4695	.002387	.0489
420	176400	20.4939	.002381	.0488
421	177241	20.5183	.002375	.0487
422	178084	20.5426	.002370	.0487
423	178929	20.5670	.002364	.0486
424	179776	20.5913	.002358	.0486
425	180625	20.6155	.002353	.0485
426	181476	20.6398	.002347	.0485
427	182329	20.6640	.002342	.0484
428	183184	20.6882	.002336	.0483
429	184041	20.7123	.002331	.0483
430	184900	20.7364	.002326	.0482
431	185761	20.7605	.002320	.0482
432	186624	20.7846	.002315	.0481
433	187489	20.8087	.002309	.0481
434	188356	20.8327	.002304	.0480
435	189225	20.8567	.002299	.0479
436	190096	20.8806	.002294	.0479
437	190969	20.9045	.002288	.0478
438	191844	20.9284	.002283	.0478
439	192721	20.9523	.002278	.0477
440	193600	20.9762	.002273	.0477
441	194481	21.0000	.002268	.0476
442	195364	21.0238	.002262	.0476
443	196249	21.0476	.002257	.0475
444	197136	21.0713	.002252	.0475
445	198025	21.0950	.002247	.0474
446	198916	21.1187	.002242	.0474
447	199809	21.1424	.002237	.0473
448	200704	21.1660	.002232	.0472
449	201601	21.1896	.002227	.0472
450	202500	21.2132	.002222	.0471

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
451	203401	21.2368	.002217	.0471
452	204304	21.2603	.002212	.0470
453	205209	21.2838	.002208	.0470
454	206116	21.3073	.002203	.0469
455	207025	21.3307	.002198	.0469
456	207936	21.3542	.002193	.0468
457	208849	21.3776	.002188	.0468
458	209764	21.4009	.002183	.0467
459	210681	21.4243	.002179	.0467
460	211600	21.4476	.002174	.0466
461	212521	21.4709	.002169	.0466
462	213444	21.4942	.002165	.0465
463	214369	21.5174	.002160	.0465
464	215296	21.5407	.002155	.0464
465	216225	21.5639	.002151	.0464
466	217156	21.5870	.002146	.0463
467	218089	21.6102	.002141	.0463
468	219024	21.6333	.002137	.0462
469	219961	21.6564	.002132	.0462
470	220900	21.6795	.002128	.0461
471	221841	21.7025	.002123	.0461
472	222784	21.7256	.002119	.0460
473	223729	21.7486	.002114	.0460
474	224676	21.7715	.002110	.0459
475	225625	21.7945	.002105	.0459
476	226576	21.8174	.002101	.0458
477	227529	21.8403	.002096	.0458
478	228484	21.8632	.002092	.0457
479	229441	21.8861	.002088	.0457
480	230400	21.9089	.002083	.0456
481	231361	21.9317	.002079	.0456
482	232324	21.9545	.002075	.0455
483	233289	21.9773	.002070	.0455
484	234256	22.0000	.002066	.0455
485	235225	22.0227	.002062	.0454
486	236196	22.0454	.002058	.0454
487	237169	22.0681	.002053	.0453
488	238144	22.0907	.002049	.0453
489	239121	22.1133	.002045	.0452
490	240100	22.1359	.002041	.0452
491	241081	22.1585	.002037	.0451
492	242064	22.1811	.002033	.0451
493	243049	22.2036	.002028	.0450
494	244036	22.2261	.002024	.0450
495	245025	22.2486	.002020	.0449
496	246016	22.2711	.002016	.0448
497	247009	22.2935	.002012	.0449
498	248004	22.3159	.002008	.0449
499	249001	22.3383	.002004	.0448
500	250000	22.3607	.002000	.0447

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
501	251001	22.3830	.001996	.0447
502	252004	22.4054	.001992	.0446
503	253009	22.4277	.001988	.0446
504	254016	22.4499	.001984	.0445
505	255025	22.4722	.001980	.0445
506	256036	22.4944	.001976	.0445
507	257049	22.5167	.001972	.0444
508	258064	22.5389	.001969	.0444
509	259081	22.5610	.001965	.0443
510	260100	22.5832	.001961	.0443
511	261121	22.6053	.001957	.0442
512	262144	22.6274	.001953	.0442
513	263169	22.6495	.001949	.0442
514	264196	22.6716	.001946	.0441
515	265225	22.6936	.001942	.0441
516	266256	22.7156	.001938	.0440
517	267289	22.7376	.001934	.0440
518	268324	22.7596	.001931	.0439
519	269361	22.7816	.001927	.0439
520	270400	22.8035	.001923	.0439
521	271441	22.8254	.001919	.0438
522	272484	22.8473	.001916	.0438
523	273529	22.8692	.001912	.0437
524	274576	22.8910	.001908	.0437
525	275625	22.9129	.001905	.0436
526	276676	22.9347	.001901	.0436
527	277729	22.9565	.001898	.0436
528	278784	22.9783	.001894	.0435
529	279841	23.0000	.001890	.0435
530	280900	23.0217	.001887	.0434
531	281961	23.0434	.001883	.0434
532	283024	23.0651	.001880	.0434
533	284089	23.0868	.001876	.0433
534	285156	23.1084	.001873	.0433
535	286225	23.1301	.001869	.0432
536	287296	23.1517	.001866	.0432
537	288369	23.1733	.001862	.0432
538	289444	23.1948	.001859	.0431
539	290521	23.2164	.001855	.0431
540	291600	23.2379	.001852	.0430
541	292681	23.2594	.001848	.0430
542	293764	23.2809	.001845	.0430
543	294849	23.3024	.001842	.0429
544	295936	23.3238	.001838	.0429
545	297025	23.3452	.001835	.0428
546	298116	23.3666	.001832	.0428
547	299209	23.3880	.001828	.0428
548	300304	23.4094	.001825	.0427
549	301401	23.4307	.001821	.0427
550	302500	23.4521	.001818	.0426

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
551	303601	23.4734	.001815	.0426
552	304704	23.4947	.001812	.0426
553	305809	23.5160	.001808	.0425
554	306916	23.5372	.001805	.0425
555	308025	23.5584	.001802	.0424
556	309136	23.5797	.001799	.0424
557	310249	23.6008	.001795	.0424
558	311364	23.6220	.001792	.0423
559	312481	23.6432	.001789	.0423
560	313600	23.6643	.001786	.0423
561	314721	23.6854	.001783	.0422
562	315844	23.7065	.001779	.0422
563	316969	23.7276	.001776	.0421
564	318096	23.7487	.001773	.0421
565	319225	23.7697	.001770	.0421
566	320356	23.7908	.001767	.0420
567	321489	23.8118	.001764	.0420
568	322624	23.8328	.001761	.0420
569	323761	23.8537	.001757	.0419
570	324900	23.8747	.001754	.0419
571	326041	23.8956	.001751	.0418
572	327184	23.9165	.001748	.0418
573	328329	23.9374	.001745	.0418
574	329476	23.9583	.001742	.0417
575	330625	23.9792	.001739	.0417
576	331776	24.0000	.001736	.0417
577	332929	24.0208	.001733	.0416
578	334084	24.0416	.001730	.0416
579	335241	24.0624	.001727	.0416
580	336400	24.0832	.001724	.0415
581	337561	24.1039	.001721	.0415
582	338724	24.1247	.001718	.0415
583	339889	24.1454	.001715	.0414
584	341056	24.1661	.001712	.0414
585	342225	24.1868	.001709	.0413
586	343396	24.2074	.001706	.0413
587	344569	24.2281	.001704	.0413
588	345744	24.2487	.001701	.0412
589	346921	24.2693	.001698	.0412
590	348100	24.2899	.001695	.0412
591	349281	24.3105	.001692	.0411
592	350464	24.3311	.001689	.0411
593	351649	24.3516	.001686	.0411
594	352836	24.3721	.001684	.0410
595	354025	24.3926	.001681	.0410
596	355216	24.4131	.001678	.0410
597	356409	24.4336	.001675	.0409
598	357604	24.4540	.001672	.0409
599	358801	24.4745	.001669	.0409
600	360000	24.4949	.001667	.0408

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
601	361201	24.5153	.001664	.0408
602	362404	24.5357	.001661	.0408
603	363609	24.5561	.001658	.0407
604	364816	24.5764	.001656	.0407
605	366025	24.5967	.001653	.0407
606	367236	24.6171	.001650	.0406
607	368449	24.6374	.001647	.0406
608	369664	24.6577	.001645	.0406
609	370881	24.6779	.001642	.0405
610	372100	24.6982	.001639	.0405
611	373321	24.7184	.001637	.0405
612	374544	24.7386	.001634	.0404
613	375769	24.7588	.001631	.0404
614	376996	24.7790	.001629	.0404
615	378225	24.7992	.001626	.0403
616	379456	24.8193	.001623	.0403
617	380689	24.8395	.001621	.0403
618	381924	24.8596	.001618	.0402
619	383161	24.8797	.001616	.0402
620	384400	24.8998	.001613	.0402
621	385641	24.9199	.001610	.0401
622	386884	24.9399	.001608	.0401
623	388129	24.9600	.001605	.0401
624	389376	24.9800	.001603	.0400
625	390625	25.0000	.001600	.0400
626	391876	25.0200	.001597	.0400
627	393129	25.0400	.001595	.0399
628	394384	25.0599	.001592	.0399
629	395641	25.0799	.001590	.0399
630	396900	25.0998	.001587	.0398
631	398161	25.1197	.001585	.0398
632	399424	25.1396	.001582	.0398
633	400689	25.1595	.001580	.0397
634	401956	25.1794	.001577	.0397
635	403225	25.1992	.001575	.0397
636	404496	25.2190	.001572	.0397
637	405769	25.2389	.001570	.0396
638	407044	25.2587	.001567	.0396
639	408321	25.2784	.001565	.0396
640	409600	25.2982	.001563	.0395
641	410881	25.3180	.001560	.0395
642	412164	25.3377	.001558	.0395
643	413449	25.3574	.001555	.0394
644	414736	25.3772	.001553	.0394
645	416025	25.3969	.001550	.0394
646	417316	25.4165	.001548	.0393
647	418609	25.4362	.001546	.0393
648	419904	25.4558	.001543	.0393
649	421201	25.4755	.001541	.0393
650	422500	25.4951	.001538	.0392

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
651	423801	25.5147	.001536	.0392
652	425104	25.5343	.001534	.0392
653	426409	25.5539	.001531	.0391
654	427716	25.5734	.001529	.0391
655	429025	25.5930	.001527	.0391
656	430336	25.6125	.001524	.0390
657	431649	25.6320	.001522	.0390
658	432964	25.6515	.001520	.0390
659	434281	25.6710	.001517	.0390
660	435600	25.6905	.001515	.0389
661	436921	25.7099	.001513	.0389
662	438244	25.7294	.001511	.0389
663	439569	25.7488	.001508	.0388
664	440896	25.7682	.001506	.0388
665	442225	25.7876	.001504	.0388
666	443556	25.8070	.001502	.0387
667	444889	25.8263	.001499	.0387
668	446224	25.8457	.001497	.0387
669	447561	25.8650	.001495	.0387
670	448900	25.8844	.001493	.0386
671	450241	25.9037	.001490	.0386
672	451584	25.9230	.001488	.0386
673	452929	25.9422	.001486	.0385
674	454276	25.9615	.001484	.0385
675	455625	25.9808	.001481	.0385
676	456976	26.0000	.001479	.0385
677	458329	26.0192	.001477	.0384
678	459684	26.0384	.001475	.0384
679	461041	26.0576	.001473	.0384
680	462400	26.0768	.001471	.0383
681	463761	26.0960	.001468	.0383
682	465124	26.1151	.001466	.0383
683	466489	26.1343	.001464	.0383
684	467856	26.1534	.001462	.0382
685	469225	26.1725	.001460	.0382
686	470596	26.1916	.001458	.0382
687	471969	26.2107	.001456	.0382
688	473344	26.2298	.001453	.0381
689	474721	26.2488	.001451	.0381
690	476100	26.2679	.001449	.0381
691	477481	26.2869	.001447	.0380
692	478864	26.3059	.001445	.0380
693	480249	26.3249	.001443	.0380
694	481636	26.3439	.001441	.0380
695	483025	26.3629	.001439	.0379
696	484416	26.3818	.001437	.0379
697	485809	26.4008	.001435	.0379
698	487204	26.4197	.001433	.0379
699	488601	26.4386	.001431	.0378
700	490000	26.4575	.001429	.0378

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
701	491401	26.4764	.001427	.0378
702	492804	26.4953	.001425	.0377
703	494209	26.5141	.001422	.0377
704	495616	26.5330	.001420	.0377
705	497025	26.5518	.001418	.0377
706	498436	26.5707	.001416	.0376
707	499849	26.5895	.001414	.0376
708	501264	26.6083	.001412	.0376
709	502681	26.6271	.001410	.0376
710	504100	26.6458	.001408	.0375
711	505521	26.6646	.001406	.0375
712	506944	26.6833	.001404	.0375
713	508369	26.7021	.001403	.0375
714	509796	26.7208	.001401	.0374
715	511225	26.7395	.001399	.0374
716	512656	26.7582	.001397	.0374
717	514089	26.7769	.001395	.0373
718	515524	26.7955	.001393	.0373
719	516961	26.8142	.001391	.0373
720	518400	26.8328	.001389	.0373
721	519841	26.8514	.001387	.0372
722	521284	26.8701	.001385	.0372
723	522729	26.8887	.001383	.0372
724	524176	26.9072	.001381	.0372
725	525625	26.9258	.001379	.0371
726	527076	26.9444	.001377	.0371
727	528529	26.9629	.001376	.0371
728	529984	26.9815	.001374	.0371
729	531441	27.0000	.001372	.0370
730	532900	27.0185	.001370	.0370
731	534361	27.0370	.001368	.0370
732	535824	27.0555	.001366	.0370
733	537289	27.0740	.001364	.0369
734	538756	27.0924	.001362	.0369
735	540225	27.1109	.001361	.0369
736	541696	27.1293	.001359	.0369
737	543169	27.1477	.001357	.0368
738	544644	27.1662	.001355	.0368
739	546121	27.1846	.001353	.0368
740	547600	27.2029	.001351	.0368
741	549081	27.2213	.001350	.0367
742	550564	27.2397	.001348	.0367
743	552049	27.2580	.001346	.0367
744	553536	27.2764	.001344	.0367
745	555025	27.2947	.001342	.0366
746	556516	27.3130	.001340	.0369
747	558009	27.3313	.001339	.0366
748	559504	27.3496	.001337	.0366
749	561001	27.3679	.001335	.0365
750	562500	27.3861	.001333	.0365

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
751	564001	27.4044	.001332	.0365
752	565504	27.4226	.001330	.0365
753	567009	27.4408	.001328	.0364
754	568516	27.4591	.001326	.0364
755	570025	27.4773	.001325	.0364
756	571536	27.4955	.001323	.0364
757	573049	27.5136	.001321	.0363
758	574564	27.5318	.001319	.0363
759	576081	27.5500	.001318	.0363
760	577600	27.5681	.001316	.0363
761	579121	27.5862	.001314	.0363
762	580644	27.6043	.001312	.0362
763	582169	27.6225	.001311	.0362
764	583696	27.6405	.001309	.0362
765	585225	27.6586	.001307	.0362
766	586756	27.6767	.001305	.0361
767	588289	27.6948	.001304	.0361
768	589824	27.7128	.001302	.0361
769	591361	27.7308	.001300	.0361
770	592900	27.7489	.001299	.0360
771	594441	27.7669	.001297	.0360
772	595984	27.7849	.001295	.0360
773	597529	27.8029	.001294	.0360
774	599076	27.8209	.001292	.0359
775	600625	27.8388	.001290	.0359
776	602176	27.8568	.001289	.0359
777	603729	27.8747	.001287	.0359
778	605284	27.8927	.001285	.0359
779	606841	27.9106	.001284	.0358
780	608400	27.9285	.001282	.0358
781	609961	27.9464	.001280	.0358
782	611524	27.9643	.001279	.0358
783	613089	27.9821	.001277	.0357
784	614656	28.0000	.001276	.0357
785	616225	28.0179	.001274	.0357
786	617796	28.0357	.001272	.0357
787	619369	28.0535	.001271	.0356
788	620944	28.0713	.001269	.0356
789	622521	28.0891	.001267	.0356
790	624100	28.1069	.001266	.0356
791	625681	28.1247	.001264	.0356
792	627264	28.1425	.001263	.0355
793	628849	28.1603	.001261	.0355
794	630436	28.1780	.001259	.0355
795	632025	28.1957	.001258	.0355
796	633616	28.2135	.001256	.0354
797	635209	28.2312	.001255	.0354
798	636804	28.2489	.001253	.0354
799	638401	28.2666	.001252	.0354
800	640000	28.2843	.001250	.0354

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
801	641601	28.3019	.001248	.0353
802	643204	28.3196	.001247	.0353
803	644809	28.3373	.001245	.0353
804	646416	28.3549	.001244	.0353
805	648025	28.3725	.001242	.0352
806	649636	28.3901	.001241	.0352
807	651249	28.4077	.001239	.0352
808	652864	28.4253	.001238	.0352
809	654481	28.4429	.001236	.0352
810	656100	28.4605	.001235	.0351
811	657721	28.4781	.001233	.0351
812	659344	28.4956	.001232	.0351
813	660969	28.5132	.001230	.0351
814	662596	28.5307	.001229	.0351
815	664225	28.5482	.001227	.0350
816	665856	28.5657	.001225	.0350
817	667489	28.5832	.001224	.0350
818	669124	28.6007	.001222	.0350
819	670761	28.6182	.001221	.0349
820	672400	28.6356	.001220	.0349
821	674041	28.6531	.001218	.0349
822	675684	28.6705	.001217	.0349
823	677329	28.6880	.001215	.0349
824	678976	28.7054	.001214	.0348
825	680625	28.7228	.001212	.0348
826	682276	28.7402	.001211	.0348
827	683929	28.7576	.001209	.0348
828	685584	28.7750	.001208	.0348
829	687241	28.7924	.001206	.0347
830	688900	28.8097	.001205	.0347
831	690561	28.8271	.001203	.0347
832	692224	28.8444	.001202	.0347
833	693889	28.8617	.001200	.0346
834	695556	28.8791	.001199	.0346
835	697225	28.8964	.001198	.0346
836	698896	28.9137	.001196	.0346
837	700569	28.9310	.001195	.0346
838	702244	28.9482	.001193	.0345
839	703921	28.9655	.001192	.0345
840	705600	28.9828	.001190	.0345
841	707281	29.0000	.001189	.0345
842	708964	29.0172	.001188	.0345
843	710649	29.0345	.001186	.0344
844	712336	29.0517	.001185	.0344
845	714025	29.0689	.001183	.0344
846	715716	29.0861	.001182	.0344
847	717409	29.1033	.001181	.0344
848	719104	29.1204	.001179	.0343
849	720801	29.1376	.001178	.0343
850	722500	29.1548	.001176	.0343

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
851	724201	29.1719	.001175	.0343
852	725904	29.1890	.001174	.0343
853	727609	29.2062	.001172	.0342
854	729316	29.2233	.001171	.0342
855	731025	29.2404	.001170	.0342
856	732736	29.2575	.001168	.0342
857	734449	29.2746	.001167	.0342
858	736164	29.2916	.001166	.0341
859	737881	29.3087	.001164	.0341
860	739600	29.3258	.001163	.0341
861	741321	29.3428	.001161	.0341
862	743044	29.3598	.001160	.0341
863	744769	29.3769	.001159	.0340
864	746496	29.3939	.001157	.0340
865	748225	29.4109	.001156	.0340
866	749956	29.4279	.001155	.0340
867	751689	29.4449	.001153	.0340
868	753424	29.4618	.001152	.0339
869	755161	29.4788	.001151	.0339
870	756900	29.4958	.001149	.0339
871	758641	29.5127	.001148	.0339
872	760384	29.5296	.001147	.0339
873	762129	29.5466	.001145	.0338
874	763876	29.5635	.001144	.0338
875	765625	29.5804	.001143	.0338
876	767376	29.5973	.001142	.0338
877	769129	29.6142	.001140	.0338
878	770884	29.6311	.001139	.0337
879	772641	29.6479	.001138	.0337
880	774400	29.6648	.001136	.0337
881	776161	29.6816	.001135	.0337
882	777924	29.6985	.001134	.0337
883	779689	29.7153	.001133	.0337
884	781456	29.7321	.001131	.0336
885	783225	29.7489	.001130	.0336
886	784996	29.7658	.001129	.0336
887	786769	29.7825	.001127	.0336
888	788544	29.7993	.001126	.0336
889	790321	29.8161	.001125	.0335
890	792100	29.8329	.001124	.0335
891	793881	29.8496	.001122	.0335
892	795664	29.8664	.001121	.0335
893	797449	29.8831	.001120	.0335
894	799236	29.8998	.001119	.0334
895	801025	29.9166	.001117	.0334
896	802816	29.9333	.001116	.0334
897	804609	29.9500	.001115	.0334
898	806404	29.9666	.001114	.0334
899	808201	29.9833	.001112	.0334
900	810000	30.0000	.001111	.0333

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
901	811801	30.0167	.001110	.0333
902	813604	30.0333	.001109	.0333
903	815409	30.0500	.001107	.0333
904	817216	30.0666	.001106	.0333
905	819025	30.0832	.001105	.0332
906	820836	30.0998	.001104	.0332
907	822649	30.1164	.001103	.0332
908	824464	30.1330	.001101	.0332
909	826281	30.1496	.001100	.0332
910	828100	30.1662	.001099	.0331
911	829921	30.1828	.001098	.0331
912	831744	30.1993	.001096	.0331
913	833569	30.2159	.001095	.0331
914	835396	30.2324	.001094	.0331
915	837225	30.2490	.001093	.0331
916	839056	30.2655	.001092	.0330
917	840889	30.2820	.001091	.0330
918	842724	30.2985	.001089	.0330
919	844561	30.3150	.001088	.0330
920	846400	30.3315	.001087	.0330
921	848241	30.3480	.001086	.0330
922	850084	30.3645	.001085	.0329
923	851929	30.3809	.001083	.0329
924	853776	30.3974	.001082	.0329
925	855625	30.4138	.001081	.0329
926	857476	30.4302	.001080	.0329
927	859329	30.4467	.001079	.0328
928	861184	30.4631	.001078	.0328
929	863041	30.4795	.001076	.0328
930	864900	30.4959	.001075	.0328
931	866761	30.5123	.001074	.0328
932	868624	30.5287	.001073	.0328
933	870489	30.5450	.001072	.0327
934	872356	30.5614	.001071	.0327
935	874225	30.5778	.001070	.0327
936	876096	30.5941	.001068	.0327
937	877969	30.6105	.001067	.0327
938	879844	30.6268	.001066	.0327
939	881721	30.6431	.001065	.0326
940	883600	30.6594	.001064	.0326
941	885481	30.6757	.001063	.0326
942	887364	30.6920	.001062	.0326
943	889249	30.7083	.001060	.0326
944	891136	30.7246	.001059	.0325
945	893025	30.7409	.001058	.0325
946	894916	30.7571	.001057	.0325
947	896809	30.7734	.001056	.0325
948	898704	30.7896	.001055	.0325
949	900601	30.8058	.001054	.0325
950	902500	30.8221	.001053	.0324

Table I (*continued*)

n	n^2	$\sqrt{n}$	$\frac{1}{n}$	$\frac{1}{\sqrt{n}}$
951	904401	30.8383	.001052	.0324
952	906304	30.8545	.001050	.0324
953	908209	30.8707	.001049	.0324
954	910116	30.8869	.001048	.0324
955	912025	30.9031	.001047	.0324
956	913936	30.9192	.001046	.0323
957	915849	30.9354	.001045	.0323
958	917764	30.9516	.001044	.0323
959	919681	30.9677	.001043	.0323
960	921600	30.9839	.001042	.0323
961	923521	31.0000	.001041	.0323
962	925444	31.0161	.001040	.0322
963	927369	31.0322	.001038	.0322
964	929296	31.0483	.001037	.0322
965	931225	31.0644	.001036	.0322
966	933156	31.0805	.001035	.0322
967	935089	31.0966	.001034	.0322
968	937024	31.1127	.001033	.0321
969	938961	31.1288	.001032	.0321
970	940900	31.1448	.001031	.0321
971	942841	31.1609	.001030	.0321
972	944784	31.1769	.001029	.0321
973	946729	31.1929	.001028	.0321
974	948676	31.2090	.001027	.0320
975	950625	31.2250	.001026	.0320
976	952576	31.2410	.001025	.0320
977	954529	31.2570	.001024	.0320
978	956484	31.2730	.001022	.0320
979	958441	31.2890	.001021	.0320
980	960400	31.3050	.001020	.0319
981	962361	31.3209	.001019	.0319
982	964324	31.3369	.001018	.0319
983	966289	31.3528	.001017	.0319
984	968256	31.3688	.001016	.0319
985	970225	31.3847	.001015	0319
986	972196	31.4006	.001014	.0318
987	974169	31.4166	.001013	.0318
988	976144	31.4325	.001012	.0318
989	978121	31.4484	.001011	.0318
990	980100	31.4643	.001010	.0318
991	982081	31.4802	.001009	.0318
992	984064	31.4960	.001008	.0318
993	986049	31.5119	.001007	.0317
994	988036	31.5278	.001006	.0317
995	990025	31.5436	.001005	.0317
996	992016	31.5595	.001004	.0317
997	994009	31.5753	.001003	.0317
998	996004	31.5911	.001002	.0317
999	998001	31.6070	.001001	.0316
1000	1000000	31.6228	.001000	.0316

Table II

A TABLE OF RANDOM NUMBERS *

Locating the Starting Point of a Series of Random Numbers

The procedure ordinarily employed to locate the first number of a series is simply to close one's eyes, place one's finger or a pencil on the table, and take the number thus pointed to as the first one of the series.† Once the starting point is located, it makes no difference to the random character of the series whether successive digits of the series are obtained by going across the row, up or down the column, obliquely, or in any other direction. If two- or three-place numbers are needed, they can be readily obtained by combining adjoining digits of two or three columns or rows.

---

* From J. G. Peatman and Roy Schafer, "A Table of Random Numbers from Selective Service Numbers," *Journal of Psychology*, 14:295–305, 1942.

† A more systematic method for locating the initial number of a series may be employed if an investigator wishes to indulge in the following "game": Place a pencil or finger on the page without looking at the table. Combine the digit thus obtained with the one immediately above to give a two-place number. If this two-place number is less than 32, use it to locate the column. If it is greater than 32, combine the initial digit with those around it in clockwise or counterclockwise order, continuing until a two-place number of 32 or less is obtained. Repeat the procedure in order to locate the row, having in mind that there are 50 rows so that the two-place numbers greater than 50 cannot be used.

To illustrate: With eyes closed, we locate a digit on the table, and we find it to be Digit 4 of Column 15, Row 25. The digit immediately above this is 9. This, therefore, gives a two-place number equal to 94—too large to locate the column. Proceeding in a clockwise direction, we find the next digit to be 3. However, 34 is still too large. The next one is 9; 94 again is too large. The next one is 2; 24 then is a two-place number giving us the location of the column, namely, the 24th column.

Again placing the finger on the page with eyes closed, we locate Number 5 of Column 7, Row 24. The digit immediately above is 5. Number 55 is too large for the location of the row. Proceeding this time counterclockwise, we find the digit to the left to be 1. This gives a two-place number equal to 15, and the 15th row is chosen. The initial digit for beginning a series of random numbers is thus located as in the 24th column and 15th row. This is Number 2.

Table II (*continued*)

A TABLE OF RANDOM NUMBERS

Row	Column number 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	2	7	8	9	4	0	7	2	3	2	5	4	2	6	7	1
2	2	2	6	0	4	1	7	7	3	8	7	3	6	7	9	4
3	9	1	6	6	3	9	4	9	1	0	5	1	5	2	2	7
4	7	0	5	5	9	2	7	5	7	8	0	8	8	5	0	6
5	4	7	3	6	6	3	9	8	2	1	7	9	7	6	4	2
6	8	2	0	2	8	7	7	6	0	2	2	3	1	1	1	6
7	0	8	7	5	3	3	6	4	2	6	8	3	1	6	5	0
8	9	4	1	9	0	8	4	6	6	8	6	3	3	2	2	3
9	5	0	0	6	7	4	0	0	0	1	9	5	9	9	1	8
10	1	9	5	4	1	5	2	6	2	9	4	1	1	5	8	4
11	5	6	4	4	1	8	7	2	8	3	6	1	5	9	8	6
12	7	9	2	5	1	9	7	9	3	1	8	6	8	7	7	6
13	3	3	3	5	9	5	1	4	0	8	2	5	6	3	5	4
14	1	9	0	4	0	0	9	9	5	7	4	1	5	9	4	7
15	5	4	4	7	2	0	3	7	9	1	0	9	6	2	9	7
16	2	9	8	2	5	5	9	3	2	0	4	9	0	6	4	4
17	9	7	6	2	6	7	7	3	3	3	1	7	5	0	9	6
18	5	8	2	4	3	3	0	8	5	3	5	7	5	8	3	5
19	4	3	4	9	5	0	3	6	2	9	7	4	6	2	5	6
20	1	1	9	8	4	8	0	6	7	0	9	7	9	6	9	9
21	6	9	1	8	3	3	7	5	9	6	6	7	7	6	0	4
22	7	0	0	3	8	1	3	4	7	9	5	2	6	9	9	7
23	3	7	2	0	8	1	5	6	9	0	1	7	8	9	6	6
24	2	7	0	0	0	6	5	0	6	5	6	0	3	2	9	3
25	3	0	7	0	7	8	4	9	4	2	8	2	4	7	4	9
26	6	2	9	3	3	1	7	7	5	2	2	3	4	6	4	2
27	5	4	9	2	1	4	8	5	7	0	9	6	4	7	2	1
28	0	3	7	0	1	7	3	8	0	3	6	2	3	1	0	9
29	9	3	6	6	2	2	0	9	7	2	3	9	2	8	7	3
30	2	9	5	6	9	9	5	6	9	8	2	8	0	0	4	4
31	8	5	7	2	9	2	6	5	9	3	9	7	1	8	3	5
32	8	4	5	7	7	9	9	5	1	4	5	5	0	9	5	3
33	8	7	9	8	1	8	4	1	4	3	7	7	0	9	1	9
34	7	3	2	5	1	8	6	3	2	8	5	8	6	9	3	4
35	8	9	9	0	1	8	8	8	9	5	7	5	0	4	1	1
36	0	2	9	7	8	8	1	7	6	1	6	7	6	4	2	5
37	0	5	2	3	2	3	8	1	8	8	1	6	2	3	0	7
38	2	2	6	8	1	6	9	6	2	6	7	9	1	7	8	0
39	0	7	8	4	9	5	8	8	0	7	2	1	8	1	7	5
40	4	8	0	7	0	5	9	9	4	9	6	9	8	2	0	6
41	9	2	0	1	6	7	2	8	3	9	8	8	3	4	7	8
42	0	8	8	3	4	0	9	2	2	8	1	5	0	4	8	2
43	2	0	6	9	7	5	2	8	2	5	5	4	0	7	7	1
44	3	1	8	6	8	3	5	6	3	2	7	4	1	8	9	4
45	0	0	8	6	1	7	5	0	8	5	6	5	0	8	2	7
46	3	3	2	9	4	2	5	3	3	8	2	4	2	6	2	5
47	8	4	7	4	0	4	5	1	2	1	0	4	2	5	7	7
48	0	2	4	3	0	2	0	7	2	8	8	0	8	4	1	6
49	4	6	5	6	3	0	4	5	2	0	1	5	2	7	9	5
50	3	4	8	3	4	5	8	7	5	9	7	1	6	3	9	9

Table II (*continued*)

Column number																
17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	Row
6	8	5	9	1	3	5	4	0	3	6	6	7	6	5	1	1
2	1	3	8	9	0	3	4	9	0	2	6	3	0	9	8	2
5	2	5	3	4	1	3	9	5	8	1	3	8	2	9	2	3
0	5	9	0	5	7	4	5	2	0	6	1	6	4	2	0	4
4	9	6	0	3	6	3	5	3	9	9	1	8	5	1	3	5
4	8	5	2	2	3	4	2	2	6	5	2	2	4	9	6	6
0	5	5	7	8	1	0	1	2	9	1	4	3	4	7	6	7
7	4	7	5	1	5	7	6	3	7	9	4	5	5	3	5	8
1	4	7	4	9	8	7	2	4	3	0	8	6	4	2	7	9
4	4	6	1	8	7	8	6	4	8	7	4	4	0	5	8	10
2	2	9	1	9	0	4	8	1	0	1	3	5	3	4	4	11
6	5	0	3	8	1	1	2	4	7	8	9	1	7	5	2	12
6	5	7	2	6	7	8	9	9	9	8	0	9	1	5	3	13
6	4	8	2	6	4	4	1	8	8	1	5	4	3	8	0	14
4	7	6	1	1	6	1	2	2	9	5	8	4	4	8	6	15
2	1	5	7	3	6	5	5	4	5	7	9	6	6	4	0	16
1	1	3	9	2	1	1	0	0	1	3	7	7	3	7	3	17
9	3	4	5	4	6	3	9	2	7	1	1	4	9	1	3	18
9	8	3	6	1	4	0	3	5	9	7	1	8	0	6	9	19
4	0	6	0	0	5	9	6	5	1	4	2	0	4	1	9	20
5	3	4	5	7	3	0	6	1	0	3	0	0	3	5	0	21
3	2	5	0	2	3	5	3	9	7	4	8	9	4	1	5	22
6	0	7	8	1	9	6	7	4	8	9	6	3	6	5	1	23
1	7	2	2	8	4	9	0	4	3	2	4	5	5	1	2	24
6	0	4	3	8	1	7	7	0	9	8	4	6	3	1	2	25
2	4	7	5	4	4	4	1	7	1	6	7	1	2	6	8	26
8	9	7	6	1	3	3	4	6	6	5	9	0	7	0	3	27
5	5	2	5	9	2	0	2	8	7	7	2	0	2	7	2	28
1	0	7	0	8	9	3	8	8	5	3	1	3	1	0	9	29
8	8	5	7	2	1	3	4	9	5	2	6	8	3	6	6	30
6	6	1	2	1	5	5	5	6	1	7	1	5	7	5	9	31
1	3	9	3	7	8	1	4	0	5	4	1	5	4	4	0	32
4	6	1	3	8	6	5	9	2	2	8	1	6	9	0	1	33
5	2	6	1	9	0	6	9	0	5	4	6	8	0	3	2	34
6	0	3	1	3	0	3	5	8	9	2	7	8	8	7	1	35
0	5	8	3	2	4	7	7	2	2	6	2	6	8	6	0	36
3	0	1	2	6	2	6	8	3	7	4	4	3	8	9	9	37
2	4	8	0	4	7	3	3	8	4	4	8	4	3	3	8	38
3	0	7	4	1	0	3	2	0	1	2	8	6	5	9	4	39
4	0	7	8	1	1	4	2	1	6	7	0	7	3	1	2	40
4	0	5	1	6	8	7	8	3	5	4	5	0	4	0	6	41
6	2	9	2	1	9	8	5	3	1	0	7	8	5	3	9	42
7	8	6	8	5	1	3	7	8	2	7	1	9	3	6	3	43
5	6	8	0	6	4	6	4	1	0	9	1	9	8	1	4	44
1	1	6	3	4	6	0	0	9	4	7	9	2	4	8	7	45
2	9	0	1	3	7	6	5	9	1	4	6	0	1	0	0	46
9	4	6	5	8	3	3	8	1	0	3	7	7	7	8	6	47
0	2	3	5	9	7	5	1	3	6	3	2	8	7	5	8	48
3	0	2	2	1	6	1	1	0	0	9	1	6	1	7	7	49
0	9	4	2	5	8	9	5	3	3	3	6	4	5	2	0	50

Glossary of Symbols

(*NOTE: The chief symbols in this book and the page on which they are first used are listed below.*)

a area from mean of normal distribution, 264
A Yule's Coefficient of Association, 91
A.D. average deviation, 168

c correction (in computation of mean and σ from a guessed mean), 159
C Contingency Coefficient, 94
C_C any centile point value (usually written with numerical subscript, thus: C_1, C_2, etc.), 138
χ^2 chi-square, 426

d differences, especially between paired deviations, 248
D differences between ranks or paired measures in original score form, 248
D a decile point value or a decile interval (usually written with numerical subscript, thus, D_1, D_2, etc.), 130
D range, 130
d.f. degrees of freedom, 428
DK's Don't Knows, 28 ff.

E index of predictive efficiency, 459

f frequencies (number of cases in a class interval or in a group of data), 113
f_h hypothetical frequencies, 95

G.M. guessed mean, 157

h hypothetical value of a parameter (usually a subscript), 323

i size of a class interval, 137
I.Q. intelligence quotient, 54

k coefficient of alienation, 452
Ku kurtosis, 392

L ratio between hypothetical length and actual length of a test, 475

M arithmetic mean, 150
Mdn. median, 139
Mo mode, 151

N number of cases or frequencies in a group of data, 51
n_c number of cases in a column, 84
n_i number of cases or frequencies in an interval, 135–136
n_r number of cases in a row, 84
N_s number of cases in a sample, 323
N_u number of cases in a universe, 323

p proportion, 43
P probability value, 329
P.E. probable error, 183
ϕ phi coefficient of correlation, 92
ϕ_r phi correlation coefficient for dichotomized variables, 93

q proportionate remainder of $1.0 - p$, 329
Q a quartile point value or a quartile interval (usually identified with appropriate subscripts, thus: Q_1, Q_3); also used to symbolize the quartile deviation, 128
Q_1 to Q_3 inter-quartile range, 128
Q.D. quartile deviation, 140
Qn a quintile point value or a quintile interval (usually identified with appropriate subscripts, thus: Qn_1, Qn_2, etc.), 130

r Pearson's product-moment correlation coefficient, 197
R multiple correlation coefficient, 482
ρ, *rho* Spearman's rank-difference correlation coefficient, 254
r_{bi} biserial correlation coefficient, 259
r_t tetrachoric correlation coefficient, 276
r_{tri} triserial correlation coefficient, 272
r^2 coefficient of determination, 489

s statistic; a sample value (usually a subscript), 323
S arithmetic summation of a series of measures, 169
S Standard score, 54
σ standard deviation, 150

σ (with the subscript of a statistic) standard error, 325
Σ algebraic summation of a series of measures, 151
Sk skewness, 390

t test ratio of Test of Significance in small sample theory, 397
T test ratio of Test of Significance, 343
T tercile point value or a tercile interval (usually identified with appropriate subscripts, thus: T_1, T_3), 130
T_1 to T_3 inter-tercile range, 128
T.D. tercile deviation, 141

u universe (as subscript), 323

v any variable, 243
V Pearson's Coefficient of Relative Variation, 171
Vn vigintile point value or a vigintile interval (usually identified by appropriate subscripts, thus: Vn_1, Vn_2, etc.), 130

x deviate value of X from mean; also a variable, 151
x' deviations in unit interval terms, 157
X original score, 166

y deviate value of Y from mean; also a variable, 151; also ordinate, 260

z z scores (deviations in units of the standard deviation), 177
z Fisher's **z** transformation function for r, 386

Glossary of Principal Formulas

(In the case of alternative formulas, more than one page number is given unless they appear on the same page.)

Alienation, coefficient of, 452

$$k = \frac{\sigma_y\sqrt{1 - r_{xy}^2}}{\sigma_y} = \sqrt{1 - r_{xy}^2}$$

Arithmetic mean, *see* Mean

Association, Coefficient of, 92

$$A = \frac{ad - bc}{ad + bc}$$

Average deviation

grouped data, 170

$$A.D. = \frac{Sf(x)}{N}$$

ungrouped data, 169

$$A.D. = \frac{S(X - M)}{N} \text{ or } \frac{S(x)}{N}$$

standard error of, 380

$$\sigma_{AD} = \frac{0.603\sigma}{\sqrt{N_s}} \text{ or in terms of } A.D. \quad \sigma_{AD} = \frac{0.756AD}{\sqrt{N_s}}$$

Binomial for any power of n, 338

$$(p + q)^n = p^n + \frac{n}{1}p^{(n-1)}q + \frac{n(n-1)}{1 \cdot 2}p^{(n-2)}q^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3}p^{(n-3)}q^3 + \cdots + q^n$$

Biserial r, 260

$$r_{bi} = \left(\frac{M_h - M_l}{\sigma_t}\right)\left(\frac{p_h q}{y}\right)$$

alternative form, Dunlap's formula, 260

$$r_{bi} = \left(\frac{M_h - M_l}{\sigma_t}\right)\left(\frac{p_h}{y}\right)$$

point-biserial r, 271

$$r_{pt\text{-}bi} = \left(\frac{M_P - M_Q}{\sigma_t}\right)\sqrt{pq} \quad \text{or} \quad r_{pt\text{-}bi} = \left(\frac{M_P - M_T}{\sigma_t}\right)\sqrt{\frac{p}{q}}$$

standard error of, 388, 389

$$\sigma_{r_{bi}} = \frac{\frac{\sqrt{pq}}{y} - r_{bi}^2}{\sqrt{N_s}} \quad \text{or for null hypothesis} \quad \sigma_{r_{bi}} = \frac{\frac{\sqrt{pq}}{y}}{\sqrt{N_s}} = \frac{\sqrt{pq}}{y\sqrt{N_s}}$$

Centile, 136

$$C_C = X_l + \left(\frac{pN - f_b}{f_i}\right)i$$

standard error of, 381, 382

$$\sigma_{C_c} = \frac{\sigma}{y}\sqrt{\frac{pq}{N_s}} \quad \text{or in terms of } Q \quad \sigma_{C_c} = \frac{1.483Q}{y}\sqrt{\frac{pq}{N_s}}$$

Chi-square, 426

$$\chi^2 = \sum\left[\frac{(f_s - f_h)^2}{f_h}\right]$$

for Test of Independence, Pearson's short-cut formula, 441

$$\chi^2 = \frac{N_s(ad - bc)^2}{(a + b)(c + d)(b + d)(a + c)}$$

in terms of phi, 443

$$\chi^2 = N_s\phi^2$$

Test of Significance, when $d.f. > 30$, 430

$$T = \frac{s - h}{\sigma_s} = \frac{(\sqrt{2\chi^2} - \sqrt{2(d.f.) - 1}) - 0}{1.0}$$

Coefficient of Relative Variation, 171

$$V = \frac{100\sigma}{M}$$

standard error of, 418

$$\sigma_V = \frac{V}{\sqrt{2N_s}}\sqrt{1 + 2\left(\frac{V}{100}\right)^2}$$

standard error of a difference between Coefficients of Relative Variation, 418

$$\sigma_{(V_x - V_y)} = \sqrt{\sigma_{V_x}^2 + \sigma_{V_y}^2 - 2r_{V_xV_y}\sigma_{V_x}\sigma_{V_y}}$$

Contingency coefficient of correlation

from chi-square, 443

$$C = \sqrt{\chi^2/(N_s + \chi^2)}$$

Pearson's coefficient of mean square contingency, 95

$$C = \sqrt{\frac{S - N}{S}}$$

Correlation, *see* Association; Biserial r; Contingency; Phi; Product-moment r; etc.

between correlation coefficients (one array in common), 421

$$r_{(r_{12}r_{13})} = \frac{r_{23} - r_{12}r_{13}(1 - r_{23}^2 - r_{12}^2 - r_{13}^2 + 2r_{23}r_{12}r_{13})}{2(1 - r_{12}^2)(1 - r_{13}^2)}$$

D range, 140

$$D = C_{90} - C_{10}$$

standard error of, 383

$$\sigma_D = \frac{2.279\sigma}{\sqrt{N_s}} \quad \text{or in terms of } Q \quad \sigma_D = \frac{3.380Q}{\sqrt{N_s}} \quad \text{or in terms of } D \quad \sigma_D = \frac{.889D}{\sqrt{N_s}}$$

Degrees of freedom for a Test of Independence, 441

$$d.f. = (A_c - 1)(B_c - 1)$$

Efficiency index, *see* Index of predictive efficiency

Fisher's z transformation function for r, see z

Frequencies, number at mean of normal distribution, 432

$$f_{y_M} = \frac{N_s}{\sigma'\sqrt{2\pi}} = \frac{N_s}{2.51\sigma'}$$

Frequency, standard error of, 375

$$\sigma_f = \sqrt{N_s pq}$$

Hypothetical frequencies for any cell of fourfold table, 439

$$f_h = \left(\frac{n_r}{N_s}\right)\left(\frac{n_c}{N_s}\right)N_s = \frac{n_r n_c}{N_s}$$

Index of predictive efficiency, 460

$$E = 100\%(1 - \sqrt{1 - r_{xy}^2}) = 100\%(1 - k)$$

Kurtosis, 392

$$Ku = \frac{(C_{75} - C_{25})/2}{C_{90} - C_{10}} = \frac{Q.D.}{D}$$

standard error of, 392

$$\sigma_{Ku} = \frac{.27779}{\sqrt{N_s}}$$

Mean, arithmetic

grouped data, 155

$$M = \frac{\Sigma(fX_{mp})}{N} \quad \text{or} \quad \frac{\Sigma(fX)}{N}$$

grouped data, with guessed mean, 159

$$M = G.M. + i\frac{\Sigma(fx')}{N} \quad \text{or} \quad G.M. + ic$$

ungrouped data, 154

$$M = \frac{\Sigma X}{N}$$

standard error of, 376

$$\sigma_M = \frac{\sigma}{\sqrt{N_s - 1}}, \text{ or for large samples, } \frac{\sigma}{\sqrt{N_s}}$$

standard error of a difference between means of correlated samples, 413

$$\sigma_{(M_C - M_E)} = \sqrt{\sigma_{M_C}^2 + \sigma_{M_E}^2 - 2r_{M_C M_E}\sigma_{M_C}\sigma_{M_E}} = \sqrt{\sigma_{M_C}^2 + \sigma_{M_E}^2 - 2r_{CE}\sigma_{M_C}\sigma_{M_E}}$$

standard error of a difference between means, when variance of samples is same, 411

$$\sigma_{(M_1 - M_2)} = \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

standard error of a difference between means, when variance of sampling distribution is based on average of both samples, 411

$$\sigma_{(M_1 - M_2)} = \sqrt{\left(\frac{\sigma}{\sqrt{N_1}}\right)^2 + \left(\frac{\sigma}{\sqrt{N_2}}\right)^2}$$

Mean frequency of p events in binomial distribution, 333

$$M_f = N_s p$$

Mean, of a series of ranks, 1 to n, 258

$$M_{\text{ranks}} = \frac{n+1}{2}$$

of two or more groups combined, 417

$$M_c = \frac{N_1M_1 + N_2M_2 + \cdots + N_nM_n}{N_1 + N_2 + \cdots + N_n}$$

Median, 139

$$Mdn = X_\ell + \left(\frac{N/2 - f_b}{f_i}\right)i$$

standard error of, 382

$$\sigma_{Mdn} = \frac{1.253\sigma}{\sqrt{N_s}} \quad \text{or in terms of } Q \quad \sigma_{Mdn} = \frac{1.858Q}{\sqrt{N_s}}$$

Mode of a binomial distribution, 351

Mo = the integer value between $N_sp - q$ and $N_sp + q$

Multiple correlation coefficient (two variables with a third), 483

$$R_{c \cdot xy} = \sqrt{\frac{r_{cx}^2 + r_{cy}^2 - 2r_{cx}r_{cy}r_{xy}}{1 - r_{xy}^2}}$$

regression equation (three-variable problem), 484

$$\bar{z}_c = \frac{r_{cx} - r_{cy}r_{xy}}{1 - r_{xy}^2} z_x + \frac{r_{cy} - r_{cx}r_{xy}}{1 - r_{xy}^2} z_y$$

Normal probability function in terms of σ, 184

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{\frac{-x^2}{2\sigma^2}}$$

Original score values

from Standard scores, 187

$$X = M_x - 5.0\sigma_x + S\sigma_x$$

from z scores, 215

$$X = z_x\sigma_x + M_x$$

standard error of, *see* Standard error of a measure

Partial correlation coefficient (three-variable problem), 486

$$r_{cy \cdot x} = \frac{r_{cy} - r_{cx}r_{xy}}{\sqrt{1 - r_{cx}^2}\sqrt{1 - r_{xy}^2}}$$

Pearson r, *see* Product-moment coefficient of correlation

Percentage

standard error of, 373

$$\sigma_{\%} = 100\sqrt{\frac{pq}{N_s}}$$

standard error of a difference between percentages, correlated samples, 407

$$\sigma_{(\%_x-\%_y)} = 100\sqrt{\sigma_{p_x}^2 + \sigma_{p_y}^2 - 2r_{p_xp_y}\sigma_{p_x}\sigma_{p_y}} = 100\sqrt{\frac{p_xq_x}{N_x} + \frac{p_yq_y}{N_y} - 2r_{xy}\sqrt{\frac{p_xq_xp_yq_y}{N_xN_y}}}$$

standard error of a difference between percentages, non-correlated samples, 404

$$\sigma_{(\%_x-\%_y)} = 100\sqrt{\frac{p_xq_x}{N_x} + \frac{p_yq_y}{N_y}}$$

Phi coefficient of correlation

for dichotomized non-variable attributes, 92

$$\phi = \frac{bc - ad}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

for dichotomized variate with true dichotomy, 93

$$\phi = \frac{\phi}{.798}$$

for dichotomized variates, 93

$$\phi_r = \frac{\phi}{.637}$$

in terms of chi-square, 443

$$\phi = \sqrt{\chi^2/N}.$$

Point-biserial r, *see* Biserial r

Probability

of joint occurrence of independent events, 336

$$P_{(a\cdot b\cdot c\ldots n)} = P_a \cdot P_b \cdot P_c \cdot \ldots \cdot P_n$$

of occurrence of disjunctive events, 336

$$P_{(a+b+c+\ldots+n)} = P_a + P_b + P_c + \ldots + P_n$$

ratio, 329

$$P = \frac{p}{p+q}$$

Probable error of any statistic for normal sampling distributions of large sample theory, 393

$$P.E._s = .6745\sigma_s$$

See also pp. 394–395 for *P.E.* formulas of commonly used statistics

Product-moment coefficient of correlation (Pearson r)

grouped data, with guessed means, 229

$$r_{xy} = \frac{\frac{\Sigma(fx'y')}{N} - \left(\frac{\Sigma fx'}{N}\right)\left(\frac{\Sigma fy'}{N}\right)}{\sqrt{\frac{\Sigma(fx'^2)}{N} - \left(\frac{\Sigma fx'}{N}\right)^2}\sqrt{\frac{\Sigma(fy'^2)}{N} - \left(\frac{\Sigma fy'}{N}\right)^2}} = \frac{\frac{\Sigma(fx'y')}{N} - c_x c_y}{\sqrt{\frac{\Sigma(fx'^2)}{N} - c_x^2}\sqrt{\frac{\Sigma(fy'^2)}{N} - c_y^2}}$$

ungrouped data, deviations, 226, 227

$$r_{xy} = \frac{\frac{\Sigma(xy)}{N}}{\sigma_x \sigma_y} = \frac{\Sigma(xy)}{N\sigma_x\sigma_y} = \frac{\Sigma(xy)}{N\sqrt{\frac{\Sigma x^2}{N}}\sqrt{\frac{\Sigma y^2}{N}}} = \frac{\Sigma(xy)}{\sqrt{\Sigma x^2}\sqrt{\Sigma y^2}}$$

ungrouped data, original scores, 237, 238

$$r_{xy} = \frac{\frac{\Sigma(XY)}{N} - \left(\frac{\Sigma X}{N}\right)\left(\frac{\Sigma Y}{N}\right)}{\sqrt{\frac{\Sigma(X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2}\sqrt{\frac{\Sigma(Y^2)}{N} - \left(\frac{\Sigma Y}{N}\right)^2}} = \frac{\frac{\Sigma(XY)}{N} - M_x M_y}{\sqrt{\frac{\Sigma(X^2)}{N} - (M_x)^2}\sqrt{\frac{\Sigma(Y^2)}{N} - (M_y)^2}}$$

z score form, 224

$$r_{xy} = \frac{\Sigma(z_x z_y)}{N}$$

standard error of, 384

$$\sigma_r = \frac{1 - r_h^2}{\sqrt{N_s}}$$

standard error of a difference between product-moment r's, dependent samples with no array in common, 420

$$\sigma_{(r_{12}-r_{13})} = \sqrt{\sigma_{r_{12}}^2 + \sigma_{r_{13}}^2 - 2r_{r_{12}r_{13}}\sigma_{r_{12}}\sigma_{r_{13}}}$$

standard error of a difference between product-moment r's, dependent samples with one array in common, 420

$$\sigma_{(r_{12}-r_{34})} = \sqrt{\sigma_{r_{12}}^2 + \sigma_{r_{34}}^2 - 2r_{r_{12}r_{34}}\sigma_{r_{12}}\sigma_{r_{34}}}$$

standard error of a difference between product-moment r's, non-correlated samples, 419

$$\sigma_{(r_{12}-r_{34})} = \sqrt{\sigma_{r_{12}}^2 + \sigma_{r_{34}}^2} = \sqrt{\left(\frac{1 - r_{12}^2}{\sqrt{N_{12}}}\right)^2 + \left(\frac{1 - r_{34}^2}{\sqrt{N_{34}}}\right)^2}$$

Prophecy formula, 475

$$r_L = \frac{Lr_{xx'}}{1 + (L - 1)r_{xx'}}$$

Proportion
standard error of, 375

$$\sigma_p = \sqrt{\frac{pq}{N_s}}$$

standard error of a difference between proportions, non-correlated samples, 404

$$\sigma_{(p_x - p_y)} = \sqrt{{\sigma_{p_x}}^2 + {\sigma_{p_y}}^2} = \sqrt{\frac{p_x q_x}{N_x} + \frac{p_y q_y}{N_y}}$$

Quadriserial correlation coefficient, 273

$$r_{quad} = \frac{y_h M_h + (y_d - y_h)M_d + (y_b - y_d)M_b - y_b M_l}{\sigma_t \left[\frac{{y_h}^2}{p_h} + \frac{(y_d - y_h)^2}{p_d} + \frac{(y_b - y_d)^2}{p_b} + \frac{{y_b}^2}{p_l}\right]}$$

Quartile, standard error of Q_1 and Q_3, 382

$$\left.\begin{matrix}\sigma_{Q_1} = \sigma_{C_{25}} \\ \text{or} \\ \sigma_{Q_3} = \sigma_{C_{75}}\end{matrix}\right\} = \frac{1.362\sigma}{\sqrt{N_s}} \quad \text{or in terms of } Q \quad \left.\begin{matrix}\sigma_{Q_1} \\ \text{or} \\ \sigma_{Q_3}\end{matrix}\right\} = \frac{2.020Q}{\sqrt{N_s}}$$

Quartile deviation, 140

$$Q.D. = \frac{Q_3 - Q_1}{2} = \frac{C_{75} - C_{25}}{2}$$

standard error of, 383

$$\sigma_Q = \frac{.787\sigma}{\sqrt{N_s}} \quad \text{or in terms of } Q \quad \sigma_Q = \frac{1.167Q}{\sqrt{N_s}}$$

Quintiserial correlation coefficient, 275

$$r_{quint} = \frac{y_h M_h + (y_d - y_h)M_d + (y_c - y_d)M_c + (y_b - y_c)M_b - y_b M_l}{\sigma_t \left[\frac{{y_h}^2}{p_h} + \frac{(y_d - y_h)^2}{p_d} + \frac{(y_c - y_d)^2}{p_c} + \frac{(y_b - y_c)^2}{p_b} + \frac{{y_b}^2}{p_l}\right]}$$

r, *see* Product-moment coefficient of correlation
r by method of differences, 248

$$r = \frac{{\sigma_x}^2 + {\sigma_y}^2 - {\sigma_d}^2}{2\sigma_x\sigma_y} \quad \text{or when } M\text{'s and } \sigma\text{'s are equal} \quad r = 1 - \frac{\Sigma(D^2)}{2N{\sigma_x}^2}$$

r by method of sums, 247

$$r = \frac{{\sigma_s}^2 - {\sigma_x}^2 - {\sigma_y}^2}{2\sigma_x\sigma_y}$$

Rank-difference correlation coefficient, 255

$$\rho = 1 - \frac{6\Sigma(D^2)}{N(N^2 - 1)}$$

standard error of, 388

$$\sigma_\rho = \frac{(1 - \rho_h^2)}{\sqrt{N_s}} \quad \text{or for null hypothesis} \quad \sigma_\rho = \frac{1}{\sqrt{N_s}}$$

Regression coefficient

$\bar{x}$ on y, 223

$$b_{xy} = r_{xy}\frac{\sigma_x}{\sigma_y}$$

$\bar{y}$ on x, 223

$$b_{yx} = r_{yx}\frac{\sigma_y}{\sigma_x}$$

$\bar{z}_x$ on z_y, 223

$$\beta_{xy} = r_{xy}$$

$\bar{z}_y$ on z_x, 223

$$\beta_{yx} = r_{yx}$$

Regression equation

$\bar{x}$ on y, 224

$$\bar{x} = r_{xy}\sigma_x\frac{y}{\sigma_y} = r_{xy}\frac{\sigma_x}{\sigma_y}y$$

$\overline{X}$ on Y, 221

$$\overline{X} = r_{xy}\frac{\sigma_x}{\sigma_y}(Y - M_y) + M_x$$

$\bar{y}$ on x, 223

$$\bar{y} = r_{yx}\sigma_y\frac{x}{\sigma_x} = r_{yx}\frac{\sigma_y}{\sigma_x}x$$

$\overline{Y}$ on X, 219

$$\overline{Y} = r_{yx}\frac{\sigma_y}{\sigma_x}(X - M_x) + M_y$$

$\bar{z}_x$ on z_y, 222

$$\bar{z}_x = r_{xy}z_y$$

$\bar{z}_y$ on z_x, 222

$$\bar{z}_y = r_{yx}z_x$$

Reliability coefficient

effect of increasing variability of universe, 477

$$r_{xx_L} = \frac{\sigma_{x_L}^2 - \sigma_{x_s}^2(1 - r_{xx})}{\sigma_{x_L}^2}$$

Spearman-Brown prophecy formula for reliability of test as whole, 475

$$r_{zz'(2L)} = \frac{2r_{\frac{x}{2}\frac{x'}{2}}}{1 + r_{\frac{x}{2}\frac{x'}{2}}}$$

Serial correlation, *see* Biserial *r;* Quadriserial *r;* Quintiserial *r;* Triserial *r*

Skewness, 390

$$Sk = \frac{C_{10} + C_{90}}{2} - mdn$$

standard error of, 391

$$\sigma_{Sk} = \frac{.5185(C_{90} - C_{10})}{\sqrt{N_s}} = \frac{.5185D}{\sqrt{N_s}}$$

Standard deviation for distributions

grouped data, 163

$$\sigma = \sqrt{\frac{\Sigma f(x^2)}{N}}$$

grouped data, with guessed mean, 165

$$\sigma = i\sqrt{\frac{\Sigma f(x'^2)}{N} - \left(\frac{\Sigma f(x')}{N}\right)^2} \quad \text{or} \quad i\sqrt{\frac{\Sigma f(x'^2)}{N} - c^2}$$

ungrouped data, 162

$$\sigma = \sqrt{\frac{\Sigma(x^2)}{N}}$$

ungrouped data, with guessed mean equal to zero, 243

$$\sigma = i\sqrt{\frac{\Sigma(X^2)}{N} - \left(\frac{\Sigma X}{N}\right)^2}$$

with Sheppard's correction for broad classes, 168

$$\sigma_{corrected} = i\sqrt{\sigma_{u.d.}^2 - .0833}$$

standard error of, 380

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2(N_s)}} = \frac{0.707}{\sqrt{N_s}}\sigma \quad \text{or} \quad .707\,\sigma_M$$

standard error of a difference between standard deviations, 415

$$\sigma_{(\sigma_x - \sigma_y)} = \sqrt{\sigma_{\sigma_x}^2 + \sigma_{\sigma_y}^2 - 2r_{\sigma_x\sigma_y}\sigma_{\sigma_x}\sigma_{\sigma_y}} = \sqrt{\sigma_{\sigma_x}^2 + \sigma_{\sigma_y}^2 - 2r_{xy}^2\sigma_{\sigma_x}\sigma_{\sigma_y}}$$

Standard deviation for special situations

average of two or more distributions, with deviations taken from respective means, 411

$$\sigma = \sqrt{\frac{\Sigma x_1^2 + \Sigma x_2^2 + \cdots + \Sigma x_n^2}{N_1 + N_2 + \cdots + N_n}}$$

of a series of ranks, 1 to n, 258

$$\sigma = \sqrt{(x_1^2 + x_2^2 + x_3^2 \cdots + x_n^2)/N}$$

of n ranks, 258

$$\sigma = \sqrt{\frac{N^2 - 1}{12}}$$

of the frequency of p events in a binomial distribution, 334

$$\sigma = \sqrt{N_s pq}$$

of two or more combined distributions, with deviations taken from weighted mean of the combination, 417

$$\sigma = \sqrt{\frac{N_1\sigma_1^2 + N_2\sigma_2^2 + \cdots + N_n\sigma_n^2 + N_1(M_1 - M_c)^2 + N_2(M_2 - M_c)^2 + \cdots + N_n(M_n - M_c)^2}{N_1 + N_2 + \cdots + N_n}}$$

Standard error, *see* under various statistics: Average deviation; Biserial r; Centile; Coefficient of Relative Variation; D range; Kurtosis, etc.

Standard error of a difference between two statistics, 401

$$\sigma_{(s_x - s_y)} = \sqrt{\sigma_{s_x}^2 + \sigma_{s_y}^2 - 2r_{s_x s_y}\sigma_{s_x}\sigma_{s_y}}$$

for non-correlated samples, 402

$$\sigma_{(s_x - s_y)} = \sqrt{\sigma_{s_x}^2 + \sigma_{s_y}^2}$$

Standard error of a measure (test score), 378

$$\sigma_X = \sigma_x\sqrt{1 - r_{xx'}}$$

Standard error of estimate

for a multiple correlation coefficient, 485

$$\sigma_{R_{c\cdot 1,2,3,4\cdots n}} = \sigma_c\sqrt{1 - R_{c\cdot 1,2,3,4\cdots n}^2}$$

for mean of one variable predicted from mean of correlated variable, 460

$$\sigma_{est_{M_y}} = \frac{\sigma_y}{\sqrt{N_s - 1}}\sqrt{1 - r_{xy}^2}$$

of x on y, 451

$$\sigma_{est_x} = \sigma_x\sqrt{1 - r_{xy}^2}$$

of y on x, 451

$$\sigma_{est_y} = \sigma_y\sqrt{1 - r_{xy}^2}$$

Standard score, 185

$$S = 5.0 + z_x = 5.0 + \frac{X - M_x}{\sigma_x}$$

Tercile deviation, 141

$$T.D. = \frac{T_2 - T_1}{2} = \frac{C_{67} - C_{33}}{2}$$

standard error of, 383

$$\sigma_{T.D.} = \frac{.648\sigma}{\sqrt{N_s}} \quad \text{or in terms of } Q \quad \sigma_{T.D.} = \frac{.961Q}{\sqrt{N_s}} \quad \text{or in terms of } T.D.$$

$$\sigma_{T\,D.} = \frac{1.501T.D.}{\sqrt{N_s}}$$

Test of Significance, 343

$$T = \frac{\text{(sample measure)} - \text{(parameter measure)}}{\text{standard error of the measure}} = \frac{s - h}{\sigma_s}$$

Test of Significance for a difference between two statistics, 403

$$T = \frac{\text{sample difference} - \text{parameter difference of zero}}{\text{standard error of difference}} = \frac{(s_x - s_y) - 0_h}{\sigma_{(s_x - s_y)}}$$

Tetrachoric coefficient of correlation, 277

$$r_t = \frac{bc - ad}{y_1 y_2 N^2} - \frac{z_1 z_2}{2} r_t^2$$

standard error of, for null hypothesis, 389

$$\sigma_{r_t} = \frac{\sqrt{pp'qq'}}{yy'\sqrt{N_s}}$$

Triserial coefficient of correlation, 272

$$r_{tri} = \frac{y_h M_h + (y_c - y_h)M_c - y_c M_l}{\sigma_t \left[\frac{y_h^2}{p_h} + \frac{(y_c - y_h)^2}{p_c} + \frac{y_c^2}{p_l}\right]}$$

Validity coefficient, effect of increasing variability of universe, 480

$$r_{cx_L} = \sqrt{\frac{\sigma_{x_L}^2 - \sigma_{x_s}^2(1 - r_{cx_s}^2)}{\sigma_{x_L}^2}}$$

Value of statistic needed for rejection of a hypothesis, 375

$$p_s = \sigma_p T + p_h$$

z score, 177

$$z_x = \frac{X - M_x}{\sigma_x}$$

z transformation function for r, 386

$$z = \tfrac{1}{2}[\log_e (1 + r) - \log_e (1 - r)]$$

standard error of, 387

$$\sigma_z = \frac{1}{\sqrt{N_s - 3}}$$

standard error of a difference between **z**'s, non-correlated samples, 420

$$\sigma_{(z_{12}-z_{34})} = \sqrt{\sigma_{z_{12}}^2 + \sigma_{z_{34}}^2} = \sqrt{\frac{1}{N_{12} - 3} + \frac{1}{N_{34} - 3}}$$

Index

Abilities, and aptitudes, organization of, 489 ff.
  vs. aptitudes, 479
Academic success predicted from two variables, 482–483
Accidental samples, 316
Accuracy of prediction, 446, 451 ff.
Achievement test variables, 498 ff.
Actuarial analysis, 290
Actuarial prediction, 8
Addition theorem of probability, 336
Adequacy, in sampling, 294, 313–314, 353, 360, 371
  of psychological tests, 314
Aerial photography in sampling, 309
Age, as a control factor in sampling, 300 ff.
  as an index of maturity, 485
  of college freshmen, 126
Age differences in variability, 415–418
Agricultural census, 309
Alienation, coefficient of, 452 ff., 486 n., 490
Allport, G. W., 9 n.
Alternate-forms method of test reliability, 471, 473–474
Ambiguous trichotomy, 28
Amen, E. W., 23–24, 59–61, 431
American Institute of Public Opinion, 287, 315
Amount-limit test, 466
Analytical statistics, 10–11, 283 ff.
Andrew, D. M., 466 n.
Approximate measures, 16–17
Aptitudes, definition of, 469, 479
Areal sampling, 306–309
Areal units in sampling, 298
Areas of normal probability curve, 179, 264, 508–511
Arithmetic ability, 486
Arithmetic mean, *see* Mean
Army Alpha, 190–192, 380, 464, 473 n.
Army Beta, 464
Array, 101–103
Association, Coefficient of, 91 ff.
Asymmetrical distributions, 349 ff.
Attitude scales, 27–28
Attributes, classification and enumeration of, 19 ff.
Average deviation, 150, 168–170
  P.E. of, 394
  standard error of, 380
Average of standard deviations for two or more samples, 411, 417
Averaging ranks, 256
  inter-correlation coefficients, 476

Bar graphs, 59–64, 71
Bar trend graphs, 61–64
Barlow's tables, 121
Barometer, reliability and validity of, 465 ff.
Barr, A. S., 506
Bell, E. P., 4 n.
Bell, H. M., 473
Bell-shaped distribution, 129, 174 ff., 431 ff.
Belt graphs, 64–67
Bennett, G. K., 187, 190, 191
Bennett Mechanical Comprehension Test, 186 ff., 400
Bernoulli, Jacques, 3, 4
Bernoulli, Nicolas, 3
Bernreuter Personality Inventory, 190–192
Bernreuter scores, 100 ff., 160

Biased samples, 292–294, 314
Bilateral symmetry, 152, 175
Binet I.Q., 378
Binet intelligence test, 472
Binet mental age, 54
Bingham, W. V., 185 n., 516
Binomial, and normal probability curve, 331–332, 340 ff.
for any power of n, 338
when $p \neq q$, 350–353
Binomial expansion, when $n = 2$, 332–335; when $n = 3$, 332–338; when $n = 6$ and 12, 339–340
Biserial coefficient of correlation, 259 ff., 338
Biserial r, and test item analysis, 259, 481
P.E. of, 395
standard error of, 388–389
Bi-variate data, 205 ff.
Body measurements, 493 ff.
cluster analysis of, 493–498
Bowditch, A. P., 14
Box method of tallying, 36, 112
Brigham, C. C., 380
British Institute of Public Opinion, 287
Broadcasting Magazine, 147
Brown, W., 474 n.
Bryan, A. I., 503 n.
Bucknell College, 498
Buell, Bradley, 73 n.
Bureau, of Agricultural Economics, 307
of the Census, 307
Buros, O. K., 506

Cantril, Hadley, 287 n., 309 n., 315
Castenada, Carlos, 5
Categorical data, 13–14, 19 ff., 424
comparison of, 43 ff.
correlation of, 80 ff.
cross-tabulation of, 80 ff.
graphic methods for, 58 ff.
Categories, mutually exclusive, 21
Cattell, J. M., 316
Causal relations, 422, 489
Causality and statistical relations, 330
Census, 4–5, 11
data, 13, 285
enumeration districts, 307–308
vs. sample, 11, 285, 319
Centile, definition, 127–128
Centile graph, 131–133, 143, 147
Centile intervals, 127 ff.
Centile measure, of kurtosis, 392–393
of skewness, 390–392
Centile measures, 139 ff.
summary of, 130
Centile method, 127 ff., 381
Centile point values, 127 ff.
Centile values, checking of, 136–138
computation of, 134 ff.
Centiles, 127 ff.
determined by graphic method, 131 ff.
P.E. of, 394–395
standard error of, 381 ff.
Tests of Significance for, 381–383
Central tendency, 131, 139, 152, 154
C.g.s. system, 470
Chance errors, 286, 288, 293, 313, 332
Chance expectancy for cross-tabulated attributes, 438 ff.
Chance factors, 356
Chance hypothesis, 431
See also Null hypothesis
Chappell, M. N., 296 n.
Character, of sample results, 353
of samples vs. size of samples, 314–316
Charts, bar type, 71
circular, 67–68
class interval limits, 109
correlation, 81 ff., 207, 214, 231
pictorial, 72–78
purpose of, 58
See also Graphs
Chesire, L., 275 n., 408 n., 476 n., 506
Chi-square, 373, 424 ff.
and Contingency Coefficient, 443
and test item analysis, 482
as Test of Significance, 430
probability, 429, 515
Test of Significance, for correlation between dichotomized attributes, 437–441; for independence of two attributes, 437 ff.; for trichotomy, 431; for variable distributions, 431–437
vs. centile analysis of skewness and kurtosis, 437
Circular charts, 67–68
City College of New York opinion poll, 34 ff.
Class interval, 103–111
mathematical limits of, 105–108
mid-value of, 108–109
scale limits of, 109–110
size of, 104–105
Classes and subclasses, 22 ff.
Classification, 19 ff.
of judgments, attitudes, and opinions, 25 ff.
rules of, 24–25
Clerical efficiency predicted from two variables, 484
Clerical occupations, 478
Clerical proficiency test, 261 ff.
Clerical success, 469
Cluster analysis, 464
and factor analysis, 489 ff.
of body measurements, 493–498
of psychological variables, 498–503

Coefficient, alienation, 452 ff., 486 n., 490
Association, 91–92
correlation, *see* Correlation
determination, 489
mean square contingency, 86, 90, 94, 253 n.
non-determination, 490
Relative Variation, 171–172, 418–419
reliability, 470 ff.
Risk, 365
validity, 480
Cohen, M. R., 25 n., 34 n.
Coleman, J. H., 319 n.
Collective, 285
Columbia Broadcasting System, 80, 421
Combinations in sampling, 332 ff.
Common determiners, 493
Common factors, 491, 498
Communality of functions, 498
Component factors, 489 ff.
Confidence criteria, 364 ff., 372
and likelihood, 360 ff.
for significance of a difference, 403–404
Confidence levels, 364 ff., 429 ff.
0.1% level, 366, 429
1% level, 403
2% level, 365
5% level, 365–366
in terms of *P.E.*, 396–397
in terms of *T* ratios, 366 ff.
Confidence limits, 371 ff.
centiles, 381–382
chi-square, 429 ff.
I.Q. scores, 378–379
means, 377
percentages, 374
predictive estimates, 462
reliability of a sample *r*, 387–388
standard deviations, 379
testing continuum of hypotheses, 368 ff.
Conklin, E. G., 22 n.
Conrad, H. S., 464 n.
Constant errors, 293
and size of samples, 297
Consumer expenditures in N. Y. State, 67
Consumer's brand preference, 426–427
Contingency Coefficient, 86, 90, 94–98
from chi-square, 443
maximum values, 97
Continuous series, 15–16, 127
Continuum, 16
of hypotheses, 368 ff.
Control group, 407
Controlled experimentation and sampling, 319–321
Controls in sampling, 299 ff., 320
Converting *rho* to *r*, 388
Copy testing, 321
Cornfield, J., 306 n.
Correlation, 6–8, 24, 80 ff., 195 ff., 253 ff., 437 ff., 445 ff.
and causal relationships, 422, 489
and chi-square, 437 ff.
and heterogeneity in age, 486–487
between: correlation coefficients, 421; dichotomized attributes, 90 ff., 276 ff., 437–441; two proportions, 407
by method, of differences, 248–249; of sums, 247–248, 475
Contingency Coefficient, 94–98
cross-tabulation essential to, 80
distribution, 207
in descriptive statistics, 80 ff.
in sampling statistics, 80
index of test reliability, 470 ff.
methods for evaluation of psychological tests, 464 ff.
non-variable attributes, 84–86, 91–92, 437 ff.
of categorical data, 80 ff.
of dichotomized variables, 92–94, 258 ff., 276 ff.
of intelligence test scores and grades, 457–459
of "memory factors," 472
of polytomous attributes, 86 ff.
of ranks, 253–258
phi coefficient, 92–94, 253 n.
predictive meaning, 445 ff.
product-moment method, 91, 195 ff., 445 ff.
product-moment *r* and ϕ, 93
rank-difference method, 254 ff.
rank-product method, 254
rank-sum method, 254
spurious, 205, 487
standard error of, *see* Standard error
surfaces, 446
Correlation chart, 81 ff., 206 ff., 214, 231
as geometric field, 81–82, 199
means and standard deviations from, 235
Correlation coefficients, standard errors of, 384 ff.
Correlation profile, analysis, 492 ff.
for achievement test variables, 500
for body measurements variables, 495
Correlation tally, 111–112, 207
Correlational frequency, 203–206
Co-variability, 195 ff.
Co-variation, 8
Critical ratio, 366
Critical scores, 469
Crossley, Inc., 147
Cross-section vs. representative samples, 312
Cross-tabulation, 37
essential to correlation, 80 ff.
of bi-variate data, 197 ff.
of categorical data, 80

Cumulative frequency distribution, 121–122
Cureton, E. E., 261 n., 272, 492 n.
Curve of error, 4, 285, 326
  See also Normal probability curve

D range, 130, 140
  P.E. of, 395
  standard error of, 383
Data, 13–15
  of categories, 19 ff.; of variables, 99 ff.
Deciles, 128–130
  P.E. of, 395
  standard error of D_1 and D_9, 383
Degrees of freedom, 373, 399 n., 428 ff., 438
  for test of independence, 441
Deming, W. E., 293, 294 n., 296 n., 307 n.
de Moivre, 4
Dependent samples, 319, 402
Descriptive statistics, 4 ff.
  summary of methods, 283–284
Determination, coefficient of, 489–490
Deviation, 176
  mean, *see* Average deviation
  measures of, 140–141, 161 ff.
Dewey, T. E., 405–406
Dichotomization, of bi-variates, 276 ff.
  of height-weight measures, 276
Dichotomized attributes, 437–441
Dichotomized variables, and serial correlation, 258 ff.
  correlation of, 92–94
Dichotomous classification, 19–20
  and division, 20 ff.
Dichotomy, 90
  on normal, bell-type distribution, 265
Differences between statistics, Tests of Significance for, 401 ff.
Differentiation, quantitative, 33
Digit-span test, 227–228, 248–219, 472
Diminishing returns in sampling, 300, 305, 320
Discontinuous series, 15–16
Discrete data, 16
Discrete sampling distributions, 333 ff.
Dispersion, 100, 140
Distance, measurement of, 470
Distribution, bell-shaped, 129, 174 ff., 431 ff.
  J-type, 129, 175 n.
  normal probability, *see* Normal probability curve
  of chi-square, 427 ff., 515
  of frequencies, 112 ff.
  of sample results, 323–324 (*see also* Sampling distribution)
  of *t*, 398, 514
  rectangular, 129
  skewed, 349–353
  U-type, 175 n.

Division, 19 ff.
  by exact criteria, 21
D.K.'s, 29 ff., 87–88, 405
  Type I, 30
  Type II, 30
Doubtful inferences, 365
Dreyfuss, M., 34 n.
Du Bois, P. H., 254 n., 256 n.
Duncan, A. J., 350 n., 506
Dunlap, J. W., 260, 506
Dunlap's formula for biserial correlation, 260

Eaton, R. M., 25 n.
Edgerton, H. G., 506
Efficiency, of predicted estimates better than a guess, 455
  of prediction, 446, 451 ff.
Empirical appraisal of a test, 478
Empiricism and research, 360 ff.
Enumeration, 19 ff.
  of attributes. 19
  vs. measurement, 33–34
Equated groups in sampling, 318–321
Equiprobability, 331–332
Equivalence of psychological tests, 473
Error, and precision in sampling, 355
  of estimate, 451 ff.
  of measurement, 285, 326
  sources of, in sampling and measurement, 294
    See also Probable error; Standard error
Errors, constant vs. chance, 293
  in stratified sampling, 311–312
  of observation, 285, 356, 465
  of prediction, 314
  of sampling, 285–288
    See also Chance errors
Evaluation of psychological tests, 464 ff.
Exact measures, 16–17
Experimental group, 407
Experimental method, and sampling, 321–322
  of equated groups in sampling, 318–321
  of matched groups, 407
Experimental science, 360–361
Extra-chance factors, 384
Ezekiel, Mordecai, 506

Factor analysis, 464, 489 ff.
Fermat, 3, 4
Fiducial limits, 371
Findex system, 39–41
Finite populations, 289
Fisher, R. A., 12, 286 n., 295 n., 348, 371, 386, 397 n., 401 n., 410 n., 424 n., 425, 427 n., 430 n., 506, 514 n., 515 n., 518 n.

Fisher's null hypothesis, 410–412
Fisher's *t* statistic, 348–349, 355, 397–399, 514
Fisher's **z** function, 384, 386–387, 400, 420, 476
and Pearson's *r*, values of, 386, 518
P.E. of, 395
standard error of, 387, 420
Flanagan, J. C., 481 n., 499 n.
Form of sampling distributions, 331 ff.
Formulas, glossary of, 551–563
Fortune polls, 404–405
Fortune Survey, 75, 78
Fourfold table, 80 ff., 276 ff., 438–440
Frequencies, and chi-square, 424 ff.
at mean of normal distribution, 432
Frequency, *P.E.* of, 394
score value of, 135
standard error of, 353, 375
Tests of Significance for, 375–376
Frequency distribution, 103 ff.
cumulative, 121 ff.
Frequency polygon, 117–118
vs. histogram, advantages, 118–120
Frequency theory of probability, 328
Functional validity, 465, 478
Functions of *r*, values of, 516–517
Fundamentum divisionis, 25

Gallup, George, 78, 287, 288, 292, 310, 318, 373
Gallup poll, 78, 286 ff., 315, 317, 373–374, 400
Galton, Sir Francis, 6, 8, 14, 91, 196, 206, 209 n.
Garrett, H. E., 503 n.
Gauss, 4
General science factors, 501
Geometric field and correlation, 81, 82, 199
Girschick, M. A., 493 n.
Glossary, of formulas, 551–563
of symbols, 547–549
Godfrey, E. H., 4
Goldman, E. F., 506
Gosset, W. S., 348
Grade scores, 126, 447 ff.
Graphic methods for categorical data, 48–78
Graphs, bar, 59–64, 71
belt, 64–67
binomial distributions, 334, 345, 350
centile, 132, 143, 147
chi-square distributions, 427
correlation profile, 495, 500
cumulative frequency, 123, 124
error of estimate, 455–457
frequency polygon, 117, 118, 120, 121
histogram, 115, 116, 119
individual psychograph, 190–191
J-type, 129
line, 117–120
normal probability curve, 7, 129, 174, 180, 265, 345, 362, 369, 396, 406
normal curve fitted to sample distribution, 433
percentage cumulative distribution, 123
percentage frequency distribution, 121
pictorial, 78
predictive estimates, 449, 450
profile, 190–191
psychograph, 190–191
rectangular distribution, 129
relation of *P.E.* to σ, 396
sampling distributions, 334, 345, 350, 362, 369, 406
scatter of correlation frequencies, 198, 201–203, 445
variability in sample results, 358
See also Charts
Group factors, 491, 498
Guessed mean, 157
Guilford, J. P., 481 n.
Gulliksen, H. O., 272 n., 506

Halley, 5
Hallonquist, Tore, 80 n., 119 n., 407 n., 421 n.
Hand-sorting, 37–38
Hansen, M. H., 296 n., 307 n.
Height measurements of infants, 204
Heterogeneity, in age, 486, 493
in sampling, 293, 315
of matched samples, 414
Hidden factors in correlation, 487
Higgons, R. A., 120 n., 198 n.
Histogram, 114–117
advantages over frequency polygon, 118–120
Homo sapiens, 23
Homogeneity in sampling, 292–293
Homoscedasticity, 452 n.
Hooper, C. E., 296
Hoover, Herbert, 306
Hotelling, H., 491, 492
House-to-house interviewing, 297
Hull, Clark, 236 n., 459
Human traits, organization of, 489
Huygens, Christian, 3
Hypotheses, 325, 360 ff., 424
about distributions of frequencies, 424 ff.
and Tests of Significance, 360 ff.
of "no difference," 401
of zero difference, 410
See also Null hypothesis
Hypothetical frequencies, for normal distributions, 432 ff.
for test of independence, 439 ff.

I.B.M. card, 38, 236
Identification in measurement, 33
Ignorant samples, 316
Independence values, 95–96, 439 ff.
Independent samples, 319, 402
Index numbers, 52–55
  of predictive efficiency, 459–460
  of reliability, 465
Individual differences, 34 ff.
Inertia of large numbers, principle of, 297
Infants' height-weight measurements, 204
Infants' sitting ages, 120 ff.
Infinite populations, 289
Initial sampling units, 298
Insignificant differences, 403
Institute of Public Administration, 65, 66, 69–71
Intelligence, 470, 479
  and *G*, 491
Intelligence Quotient, 54
Intelligence test scores, 100, 126, 190–192, 378–379, 447 ff.
Inter-correlation, coefficients, 238–239
  of factors in sampling, 304
Inter-quartile range, 128–130
Inter-tercile range, 128–130, 141
Interest, 469
Interest ratings, Strong, 100
Internal controls in sampling, 306, 311
Interviewing, 297
Intra-group differences in sampling, 317–318
Invariant relationship, 195, 210
I.Q., 376–379, 492
I.Q. index, 54
Item inter-correlation, 471
  and test reliability, 476
Item reliability and validity, 481

J-type distribution, 129, 175 n.
Jaspen, Nathan, 272, 273, 275
Jessen, R. J., 309 n.
Judgments, attitudes, and opinions, 25 ff.

Kelley, T. L., 107, 390, 452, 491, 506
Kellogg, L. S., 506
Kendall, M. G., 15 n., 91 n., 295 n., 424 n., 506
Kenney, J. F., 506
King, A. J., 309 n.
Klineberg, O., 409–411
Koren, John, 4 n.
Kuder, G. F., 473 n.
Kurtosis, 348, 372 n., 390, 392–393, 431 ff.
  standard error of, 392
Kurtz, A. K., 506

Landon, 292
Laplace, 4
Large sample theory, 343, 355–356, 367
  Tests of Significance for, 360 ff.
  vs. small sample theory, 324
Large samples vs. small samples, 349
Larrabee, H. A., 360 n.
Laws of chance, 331
Lazarsfeld, Paul, 29 n., 31, 407 n.
Learning curves, 58
Least squares, method of, 225
Length factors, 498
Leptokurtic sampling distributions, 347–348, 353, 355
Leptokurtosis, 348, 393
Lerrigo, Ruth, 73 n.
Likelihood, 329–330
  and confidence criteria, 360 ff.
Likely hypotheses, 329, 360 ff., 397, 454
Likely results in sampling, 340, 358–359, 366
Likert, Rensis, 309
Limits, class intervals, 103 ff.
  tenable hypotheses, 368–369
  untenable hypotheses, 368–369
Line graph, 116–120
Linear correlation, *see* Product-moment *r*
Linear regression lines for bi-variate distributions, 208 ff.
Link, H. C., 444 n.
Linnaeus, 23
Literary Digest poll, 292, 306
Literature factors, 501
Locke, N. M., 227 n., 472 n.
Logical division, rules, 24–25
Longstaff, H. P., 466 n.
Lottery methods in sampling, 295

McNemar, Quinn, 316 n., 414 n.
Machine method for product-moment *r*, 236
Machine tabulation, 38–39
Mail-ballot poll, 306
Map charts, 69, 70
Maps, 68–72
Market research, 17, 29 ff., 86 ff., 296, 306 ff., 361 ff.
Master sample, technique of, 309–310
Matched groups, 412
  experimental method of, 407
Matched samples, 319–321, 407
Matching pairs, 407
Mathematical factors, 501
Mathematical limits of class intervals, 105 ff.
Mean, 7, 54, 150 ff.
  as fulcrum, 175
  as measure of central tendency, 152, 154
  as typical measure, 152
  as point of reference, 175
  correction for, 158–159
  for data grouped into class intervals, 154 ff.

for ungrouped data, 153–154
from guessed mean, 155–160
of series of ranks, 258
probable error of, 394
reliability of, 377
standard error of, 376
Test of Significance for, 376–377
Mean deviation, 168
See also Average deviation
Mean differences, 409 ff.
Mean frequency in binomial distribution, 333
Means, and standard deviations from correlation chart, 235
of independent samples, 409 ff.
Measure of scatter, 452 ff.
P.E. of, 394
standard error of, 378
Measurement, enumeration vs., 33–34
errors of, 285, 326
Mechanical Comprehension test, 186–188, 400
Median, 128, 131, 135, 139
P.E. of, 394
standard error of, 382
Median inter-correlation coefficient, 476
Mental age scores, 380
Merrill, M. A., 378 n., 379, 414 n.
Mesokurtosis, 348, 353, 392–393
Mid-case, 139 n.
Mid-values of class intervals, 108–109
Minnesota Vocational Test for Clerical Workers, 466 ff., 478
Modal frequencies, 351–352
Mode, 175
of binomial distribution, 351
Moments, 150 n.
Multiple choice method, 26
Multiple correlation, 481, 482–485, 489
Multiple-factor theories, 491
Multiple regression equation, 484

Nagel, Ernest, 25 n., 33 n., 34 n.
Name-checking, 466
Necessary inference, 361
Negative correlation, 201
and prediction, 450–451
Negative numbers in psychological scales, 114
New York *Daily News* poll, 315
New York *Herald Tribune*, 404 n.
New York State Teachers Association, 64, 67
New York *Times*, 373 n.
New York *Times Magazine*, 76
New York *World-Telegram*, 78
Non-chance factors, 489
Non-correlated samples, 404
Non-determination, coefficient of, 490
Non-linear correlation, 203
Non-variable attributes, 19 ff., 283, 425
comparison of, 43 ff.
correlation of, 84–86, 91–92, 437 ff.
Normal bell-shaped distribution, *see* Normal probability curve
Normal correlation surfaces, 446
Normal curve, asymptotic character of, 176
formula for, 184
implications of, 174 ff.
of error, 285, 326
Normal distribution, measures of variability for, 184
Normal probability, and binomial distributions, 340–341
and skewed sampling distributions, 349 ff.
Normal probability curve, 4, 6, 7, 14, 129, 174 ff., 183–184, 264 ff., 285, 331 ff., 431 ff., 508–511
fitted to sample distribution, 432–433
See also Graphs, normal probability curve
Normal probability distribution, 347–348, 396, 508–511
Normal sampling distributions, 333 ff.
Norms for Bennett Mechanical Comprehension test, 186 ff.
Northrup, M. S., 308 n.
Null hypothesis, 384–385, 388, 401, 403, 410–412, 425, 431, 437, 442–443
Number-checking, 466

O'Brien, R., 493 n.
Observation, errors of, 285, 356, 465
Obtained measures, 322
Occupational categories, 64
Odds-even method of reliability, 471, 474–476
Office of Public Opinion Research, 287 n., 315
Ogive, 123–124
Operational unities, 492
Operational validity, 465, 478
Opinion poll, 34 ff.
Opinion questionnaire, 421
Ordered series, 14
Ordinates of normal probability curve, 264, 432 ff., 508–511
Organization, and interrelation of psychological functions, 489 ff.
of human traits, 489
Original score value from a *z*-score, 215

Paired associates in correlation, 205–206
Parameter, 10, 322, 342 ff.
Parameter differences, 401
of zero, 403 ff.
Partial correlation, 481, 485–487, 493
Partial investigations in sampling, 316–317

Partial samples, 320
Pascal, 3, 4
Paterson, D. G., 466 n.
Payne, S. L., 308 n.
Pearson, E. S., 506
Pearson, Karl, 86, 91, 93, 94, 171, 197, 225, 279, 284, 424, 428 n., 440, 506
Pearson Coefficient of Relative Variation, 171–172, 418–419
Pearson product-moment r, 91, 195 ff., 445 ff.
Pearson r, by method of differences, 248–249
by method of sums, 247–248
Pearson short-cut computation of χ^2, 440–441
Peatman, J. G., 80 n., 105 n., 119 n., 130 n., 177 n., 198 n., 227 n., 321 n., 407 n., 421 n., 445 n., 464 n., 472 n., 543 n.
Per capita cost of education, 52–53
Per capita income, 63–64
Per capita indices, 52–53
Percentage value of a frequency, 121
Percentages, 43–46, 49–52
confusion in use of, 55–58
cumulative frequency distribution, 122–126
differences, 404 ff.
errors in averaging, 57–58
frequencies, 120
frequency distribution, 120–121
P.E. of, 394
sampling distributions, 362
standard error of, 373
Tests of Significance for, 373–374
Percentiles, 127
See also Centiles
Perl, R. E., 503 n.
Permutations, 332, 337
Personality differences, 23
Peters, C. C., 225 n., 247 n., 341 n., 356 n., 383 n., 401 n., 482 n., 506
Phi coefficient of correlation, 92–94, 253
and test item analysis, 482
P.E. of, 395
standard error of, 389
Philip, M., 225 n., 339 n.
Philip II, King of Spain, 5
Physical dimensions, 496
Pictograph Corporation, 76
Pictorial charts, 72–78
Pie diagrams, 67–68
Platykurtosis, 348
Point binomial, 331 ff.
Point-biserial correlation, 270–272
Points of inflection and σ, 175–176
Polytomous attributes, correlation of, 86 ff.
Polytomous classification, 19–21
Population, 285
Populations, finite and infinite, 289
Potts-Bennett Tests, 190–192
Practical English usage factors, 501
Precision, 353 ff.
and reliability, 355–359
and size of samples, 313, 353 ff.
function of $\sqrt{N}$, 354–355
in sampling, 313–314
measured by standard error, 353
Predictions as average estimates, 447 ff.
Predictive efficiency, in correlation, 451 ff.
index of, 459–460
of battery of tests, 462, 482 ff.
of combined tests, 482–483
Predictive meaning of correlation, 445 ff.
Primary control factor in all sampling, 311
Primary mental factors, 492
Princeton University Office of Public Opinion Research, 287 n., 315
Probability, 3, 4, 328 ff.
and Tests of Significance, 360 ff.
definition of, 328
implications of, *P.E.*, 396
of chi-square, 427 ff.
of result, 372
product and addition theorems of, 336–337
theory of, 328 ff.
Probability curve, *see* Normal probability curve
Probability distributions, 179, 334–335, 347, 396
Probability estimates, and likelihood, 329–330, 360 ff.
for normal distributions, 341 ff.
Probability ratio, 329
Probability values, chi-square, 429, 515
normal, bell-shaped distribution, 179, 508–511
proportions, 356
t of small samples, 398–399, 514
T, 512–513
Probable error, 183, 326, 356
and standard error, 396
and Tests of Significance, 393–397
of: a centile, 394; an arithmetic mean, 394; an average deviation, 394; biserial r, 395; D range, 395; estimate, 451 n.; Fisher's z function, 395; frequency, 394; mean, 394; measure, 394; median, 394; percentage, 394; product-moment r, 395; proportion, 394; quartile deviation, 394; rho, 395; statistic, 326, 393; tercile deviation, 394–395
Product deviations, 233–234
Product-moment correlation, 195 ff., 445
and phi, 92–94
and rho, 254 ff.
and serial correlation, 258 ff.
by method, of differences, 248–249; of sums, 247–248
computation of, 225 ff.

confidence limits for reliability, 387
estimation of, 208 ff.
from grouped data, 229–235
from ungrouped data, 226–229
functions of, 516–517
machine method for, 236 ff.
P.E. of, 395
practical meaning of, 445 ff.
predictive implications of, 445 ff.
sampling distributions of, 324
special methods for, 253 ff.
standard error of, 384
Tests of Significance for, 384 ff.
Product theorem of probability, 336
Proficiency, 479
Profile analysis, 492 ff.
Profile chart, 188–193, 190–191
Program Analyzer, 80, 119 n., 407 n., 421
Prophecy formula, 474–475
Proportion, *P.E.* of, 394
standard error of, 375
Proportions, 43 ff.
Tests of Significance for, 375
values for p and q, 267, 519
Psychograph, 188–193
Psychological Corporation, 87, 187, 190–191
Psychological functions, organization and interrelation, 489 ff.
Psychological tests, 464 ff.
and reliability, 314, 466–468, 470 ff.
and validity, 314, 468 ff., 478 ff.
Psychological variables, cluster analysis, 498–503
Public Affairs Committee, Inc., 73, 75, 77
Public opinion research, 286–288, 290, 292, 298, 300, 306 ff., 315
Punch card, 12, 38 ff.

Quadratic equation, 276–278
Quadriserial r, 272–273
Quantitative differences, 13–14
Quartile deviation, 140–141, 382
P.E. of, 394
standard error of, 383
Quartiles, 128–130
P.E. of, 394
standard error of, 382
Quetelet, 3, 4, 5–6, 14, 17, 196
Quintiles, 128–130
Quintiserial r, 272, 275
Quota method of sampling, 310–311

r, functions of, 516–517
values of, for k, 453, 516–517; for Fisher's z function, 386, 518
See also Correlation; Product-moment correlation
R, multiple correlation, 482–485
Radio research, 119, 296, 407, 421
Radio Station WOR, 147
Random numbers, 295–296, 543–545
Random samples, 294 ff., 310 ff.
Random sampling, *see* Sampling
Randomization, primary control factor in sampling, 299 ff., 311
principle of, 294–299
Range, 99–101
Rank-difference method of correlation, 254–258
Rank-product method of correlation, 254
Rank-sum method of correlation, 254
Ranking test scores, 256
Ratios, 43–55
Raw data, 33
Reaction-time, 466, 470
Reciprocal of N, 121
Reciprocals, table of, 522–541
Rectangular distribution, 129
Reduction, of data, 11–13, 19
of sampling error by stratification, 301 ff.
Refined data, 33
Regression, 209 n., 450–451
Regression coefficients, 222–223
Regression equations, $\bar{x}$ on y, 224, 447
$\bar{X}$ on Y, 221, 447
$\bar{y}$ on x, 223, 447
$\bar{Y}$ on X, 219, 447
$\bar{z}_x$ on z_y, 220–222
$\bar{z}_y$ on z_x, 218–219
Regression line, 208 ff., 446
for $\bar{}_y$ on z_x, 217–218
Relations between measures of variability, 184
Relationships, 8–9, *see* Correlation
Relative precision in sampling, 355–359
Relative variability, 171, 418
Reliability, 160 n., 465 ff.
and precision in sampling, 355–359, 369
effect of restricted range of ability, 477
of a mean, 376
of a sample r, 387
of a standard deviation, 379–380
of a statistic, 371
of a test, 314, 464 ff.; by alternate-forms method, 473–474; by item-correlation, 476; by retest method, 471–473; by split-half method, 474–476
of intelligence test scores, 378–379
of test items, 481
of test scores, 378–379, 464 ff.
Reliability coefficient, 467
Representative samples vs. typical cross-section, 312
Representativeness in sampling, 312
and precision, 313–314
Restricted universes in sampling, 316–317

Rho correlation coefficient, 254–258
P.E. of, 395
relation of, to *r*, 257–258
standard error of, 388
Richardson, M. W., 271 n.
Risk, Coefficient of, 365
Roosevelt, F. D., 292, 306, 405–406
Roper, Elmo, 75, 78, 288, 404–405
Rounding off numbers, 47–49

Saffir, M., 275, 408 n., 476 n., 506
Sample, 11, 17, 34
Sample frequencies, 425 ff.
Sample types, 290 ff.
Samples, 283 ff.
biased, 292–294, 314
character of, 353; vs. size, 314–316
dependent, 319, 402
ignorant, 316
independent, 319, 402
matched, 319–321, 407
random, 294 ff., 310 ff.
representative, 291–292, 312
simple, 311
Sampling, accidental, 316
adequacy, 313–314
and experimental method, 321–322
and test norms, 307
as research technique, 286 ff.
bias in, 292–294, 314
controlled, 311
in U. S. Census of 1940, 296
internal controls, 306
inter-relation of stratifying factors, 302–303
intra-group differences, 317–318
methods of, 290 ff.
normal probability curve, 324, 331, 341 ff.
partial investigations, 316–317
precision, 313–314, 353 ff.
primary control factor, 299
randomization, 294 ff., 310 ff.
repeated, in market research, 370
restricted universe, 316
skewed distributions, 350–353
small sample vs. large sample theory, 324
stratified-quota method, 310–311
stratified-random, 299 ff.
techniques, 283 ff.
theory and cluster analysis, 491
unit, 297–299; initial vs. basic, 298–299
variations in prediction, 454
vs. census, 285, 319
Sampling distribution, 323–324, 331 ff., 350 ff., 360 ff., 372
leptokurtic, 353
of a difference between two percentages, 406
of chi-square, 427 ff.
Sampling errors, 285–288, 311–312
reduced by stratification, 301 ff.
Sampling statistics, 9 ff., 283 ff.
and experimental science, 360–361
Scale, 15
of test difficulty, 188
Scatter, measure of, 451 ff.
of correlation frequencies, 445
Scattergram, 197 ff.
Schafer, Roy, 543 n.
Schedule, 11
of information, 34–35
Scholastic aptitude, 482 ff.
School expenditures, 251–252
School Life, 251 n.
Schools as sampling units, 298–299
Scientific method, 24–25, 285–286
Score, standard error of, 378–379, 467
Score value of frequency, 135
Scores, Standard, 54, 185–186
z, 177–178
See also Test scores
Secret-ballot technique in sampling, 309
Segmented variables, 258 ff.
Selective Service System, 295
Self-correlation, 495
See also Reliability of a test
Serial correlation, 258 ff.
Series, 15–16
Sex as control factor in sampling, 300 ff.
Sex differences in variability, 414
Sex ratio, 14, 55
Shen, Eugene, 414 n.
Sheppard, W. F., 167, 200 n.
Sheppard's correction for σ, 167–168, 208
Short-cut methods, correlation, 229 ff., 253 ff.
mean, 155 ff.
standard deviation, 163 ff.
Sigma, 162
See also Standard deviation
Significance, and null hypothesis, 385
of a difference, confidence criteria for, 403–404
of sample results, 363 ff.
Simple enumeration, method of, 16
Simple samples, 311
Simple sampling, 294
Sitting ages of infants, 120 ff.
Size of samples, precision (reliability) of, 353 ff.
Skewed sampling distributions, 349 ff.
and Standard scores, 188
Skewness, 139, 159, 372, 390–392, 431–432
and kurtosis, chi-square vs. centile analysis, 437
in sampling distributions, 324
standard error of, 391

Small sample theory, 347 ff., 355–356
vs. large sample theory, 324
Small samples, Tests of Significance for, 397–399
vs. large, 349
Smith, B. B., 295 n.
Smith, J. G., 350 n., 506
Smith, V. G., 53 n., 142 n.
Social intelligence, 482
Social statistics, 3
Spearman, Charles, 253, 279, 474 n., 490, 491, 498
Spearman-Brown prophecy formula, 474–476
Spearman's general factor, *G*, 490–491
Spearman's rank-difference method of correlation, 254–258
Spearman's two-factor theory, 490–491
Specific factors, 490, 498, 501
Sperry Gyroscope Co., 319 n.
Split-half method of test reliability, 471, 474–476
Split-half reliability by differences, 475–476
Split-run copy testing, 321
Spurious correlation, 205, 487
Square roots, table of, 522–541
Squares, table of, 522–541
Stalnaker, J. M., 271 n.
Standard deviation, 54, 150, 160 ff., 243–244
for ungrouped data, 161–162
from correlation chart, 235
of a frequency in a binomial distribution, 334
of a series of ranks, 258
of sampling distributions, 325, *see also* Standard error
of two or more combined groups, 417
P.E. of, 394
reliability of, 379–380
Sheppard's correction for, 167–168
short method of computation, 163–167
standard error of, 380
Test of Significance for, 379–380
Standard error, 342
an average deviation, 380–381
biserial *r*, 388–389
centile, 381–383
Coefficient of Association, 389
Coefficient of Relative Variation, 418
correlation coefficient, 384, 388–389
D range, 383
D_1 and D_9, 383
difference between: any two statistics, 401 ff.; coefficients of relative variation, 418; correlation coefficients, 419 ff.; means, 409 ff.; percentages, 404 ff.; proportions, 404 ff.; standard deviations, 414 ff.; z functions, 420–421
estimate, 446 ff., 451 ff.; for Fisher's z function, 387; for *R*, 484; of the mean, 460; of $\bar{x}$ on *y*, 451; of $\bar{y}$ on *x*, 451
frequency, 353, 375
kurtosis, 392
mean, 376
measure, 378–379, 467
median, 382
percentage, 357–358, 373
product-moment correlation coefficient, 384
proportion, 354, 356, 375
quartile, 382
quartile deviation, 383
rank-difference correlation, 388
skewness, 391
standard deviation, 380
statistic, 325, 353
tercile, 381
tercile deviation, 383
test score, 378–379, 467
tetrachoric correlation coefficient, 389
Standard measures, centile implications of, 178–182
Standard score, 54, 185–186
and skewed distributions, 188 n.
norms, 186–188
profile chart, 188 ff.
Stanford Binet, 54, 376–379, 473 n.
Stanton, Frank, 407 n.
Statistic, 10, 322–323
value needed for rejection of hypothesis, 375
Statistical data, 13 ff.
Statistical frequencies, 36, 424
Statistical hypotheses, 325, 360 ff., 371–372, 424
and parameter values, 368
Statistical inference, 9, 328 ff.
Statistical population, 288
Statistical probability, 328 ff.
Statistical terminology, 9 ff., 285 ff., 322 ff.
Statistical Tests of Significance, 360 ff., 401 ff., 424 ff.
Statistical universe, 285, 288–290, 303, 316–317, 322–323, 371
Statistical variable, definition, 15–16
Statistics, 322
actuarial nature, 8–9
definitions, 9–11
Stature, 497
Steinzor, Bernard, 153
Stephan, F. F., 296
Stereotypes, 24
Stewart, M. S., 75 n.
Stock, J. S., 309 n.
Straight-line function, 209

Strata controls in sampling, 299 ff.
Stratification, 34 ff.
  classes and subclasses, 22 ff.
  reduction of sampling errors by, 301 ff.
  secondary control factor in sampling, 299 ff., 311
Stratified matrix, 36
Stratified samples, 299 ff.
Stratified sampling, error in, 311–312
  and representativeness, 312
Stratified-random sampling, 299 ff., 353
Stratifying factors, in sampling, 300
  inter-relation of, 302–305
Strong, E. K., 26
Strong Interest Ratings, 100
Strong's Interest Inventory, 26
Student, 348, 356 n.
Sub-samples, 298 ff., 407
Sub-universes in sampling, 305–306
Symbols, glossary of, 547–549

T, test ratio, 363 ff., 372 ff.
T ratio, 342 ff., 397
  as confidence criteria, 366 ff.
  in terms of *P.E.*, 396–397
t statistic, 348–349, 355, 397–399
t values for small sample theory, 397, 514
T values for large sample theory, 512–513
Tabulation, 19 ff., 111 ff.
Tally, 111–112
  box method, 36, 112
  correlation, 111 n., 207
Tally sheet, 36
Taylor, E. K., 267 n., 519 n.
Teachers' salaries, comparison of, 142 ff.
Tentative hypotheses, 397
Tentative inferences, 365
Tercile deviation, 141
  P.E. of, 394–395
  standard error of, 383
Terciles, 128–130
  standard error of, 381
Terman, L. M., 378 n., 379
Terminology for sampling statistics, 322 ff.
Test batteries and predictive efficiency, 462, 482 ff.
Test battery validity, 480
Test equivalence, 473
Test evaluation, 464 ff.
Test item analysis, 268, 481–482
  and biserial *r*, 259, 481
  chi-square, 482
  phi correlation, 482
  tetrachoric *r*, 482
Test items, 259, 481–482
  reliability of, 481
Test norms, 186 ff.
  and sampling, 307
Test of clerical proficiency, 261 ff.
Test ratio (*T*), 342 ff., 363 ff., 372 ff., 397–398
Test reliability, 465 ff.
  and standard error of test score, 467
  by method: of alternate forms, 473–474; of item-intercorrelation, 476; of split halves, 474–475; of test-retest, 471–473
  determination of, 470 ff.
  effect of range of ability on, 476
Test-retest measure of reliability, 471–473
Test scores, 17, 100, 186–193, 261 ff., 466
  reliability of, 378–379
  standard error of, 360 ff., 378, 467
Test-tube sample, 294, 300 ff.
Tests of Significance, 288, 348–349, 360 ff., 372 ff., 401 ff.
  and *P.E.*, 393–397
  for continuum of hypotheses, 368 ff.
  for correlation coefficients other than *r*, 388–390
  for difference between: any two statistics, 401–404; arithmetic means, 409; coefficients of relative variation, 418–419; percentages (proportions) 404–409; product-moment coefficients of correlation, 419–422; standard deviations, 414–418
  for form of variate distribution, 431 ff.
  for frequencies, 375–376
  for kurtosis, 392–393
  for large sample theory, 397–399
  for mean, 376–377
  for percentages, 373–374
  for predictive estimates, 461–462
  for product moment *r*, 384 ff.
  for proportions, 375
  for skewness, 390–392
  for small sample theory, 397–399
  for standard deviations, 379–380
  for test scores, 378–379, 467–468
  for trichotomy, 431
  for variable distributions, 431–437
  logic of, 371 ff.
Tetrachoric correlation, 275–279, 389, 408, 476
  and test item analysis, 482
Thomson, G. H., 491, 506
Thorndike, E. L., 206, 491
Thorndike Intelligence Test Scores, 457
Thurstone, L. L., 275, 408 n., 476 n., 491, 492, 506
Thurstone's *Computing Diagrams*, 276–279, 408, 476
Time, measurement of, 470
Time series, 58, 64, 119
Tippett, L. H. C., 295 n.
Torricelli barometer, 465

Traits, organization of, 489 ff.
Trichotomy, 22, 26, 86 ff., 272 ff., 431
  and correlation, 86 ff., 272 ff.
Triserial *r*, 272–273
True measures, 322
Tryon, R. C., 491 n., 492 ff.
Tryon's method of correlation profile analysis, 492 ff.
Turnbull, W., 309 n.
Two-factor theory, 490

U-shaped distribution, 175
Uni-modal distributions, 152, 175, 431–432
Unit of sampling, 297–299
Universe, 285, 288–290, 303, 316–317, 322–323, 371
Universes, actual, 289–290
  hypothetical, 289–290
Unlikely hypotheses, 329, 360 ff., 397, 454
Unlikely results in sampling, 340, 358–359, 366
Upper critical score, 469
U. S. Census, 55, 66, 296
U. S. Department of Treasury, 74

Validity, functional, 465, 478
  of test, 314, 464 ff.
  of test batteries, 480
  of test items, 481
  operational, 465, 478
Validity coefficient, effect of increase in variability, 480
  of a test, 263
Validity criteria, 479
Van Voorhis, W. R., 225 n., 247 n., 341 n., 356 n., 383 n., 401 n., 482 n., 506
Variability, differences, 414–419
  in man as sampling unit, 319
  in sample results, 358
  in terms of *P.E.*, 326, 393–397
  measures of, 140–141, 160 ff., 184
  of normal probability distribution, 341 ff.
  of sample results, 362–363
  of sampling distributions, 324, 354–356
Variable, 7, 52
  definition of, 15–16
Variable attributes, 99 ff., 283, 425
Variable data, 13–14, 27–28
Variance, 162, 244, 490
  of paired differences, 248; sums, 248
Variates, 7, 14
Variation, Coefficient of Relative, 171–172, 418–419
Vigintiles, 128–130
Vital statistics, 3–5
Vocabulary test item, 268
Volume factors, 498
von Mises, R., 328 n.

Walker, H. M., 3 n., 4 n., 428 n.
Webb, J. N., 308 n.
Wechsler, David, 415 n., 416
Wechsler-Bellevue Scale, 415–419, 423, 479
Wechsler Information test, 415–419
Weight, 497
Weight measurements of infants, 204
Weighted mean of two or more groups combined, 417
Weighting tests, 484–485
Wundt, W., 316

Yates, F., 295 n., 427 n., 506
Yule, G. U., 15 n., 91, 92, 424 n., 506

z function, 384, 386–387, 400, 420, 476, 518
  standard error of, 387
z score correlation chart, 213 ff.
z scores, 177 ff.
  and centile values, 178–182
zero on psychological scales, 114
Zubin, J., 321 n.