UNIVERSAL
LIBRARY

OU_164233

UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 311 Accession No. 29198

Author L64F Lindquist.

Title A First Course in Statistics.

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

A FIRST COURSE IN STATISTICS

THEIR USE AND INTERPRETATION IN EDUCATION AND PSYCHOLOGY

E. F. LINDQUIST
Professor of Education
State University of Iowa

REVISED EDITION

HOUGHTON MIFFLIN COMPANY
TON · NEW YORK · CHICAGO · DALLAS
ATLANTA · SAN FRANCISCO
The Riverside Press Cambridge

COPYRIGHT, 1942

BY E. F. LINDQUIST

COPYRIGHT, 1938, BY E. F. LINDQUIST

ALL RIGHTS RESERVED INCLUDING THE RIGHT TO REPRODUCE
THIS BOOK OR PARTS THEREOF IN ANY FORM

The Riverside Press

CAMBRIDGE · MASSACHUSETTS

PRINTED IN THE U.S.A.

PREFACE

THIS text, plus the accompanying *Study Manual*,[1] may be considered as constituting a *method of teaching* a first course in statistics for students of education and psychology. The distinguishing features of these materials are as follows:

1. *They rely primarily upon the Socratic method to develop in the student a reasoned understanding of statistical techniques.*

Students in first courses in statistics in education and psychology have been prone to take a passive attitude in the learning process. Upon meeting concepts which they have not readily understood, they have often resorted to memorization of stereotyped interpretations and have not made an insistent and aggressive effort to discover underlying meanings. They have been required to spend so much time on the mechanics of the solution of computational problems that they have had little time left to think about the meaning of results obtained in practical situations. They have learned how to apply statistical techniques only in the very limited sense of knowing how to compute numerical results, but have not learned when and why these techniques should be applied in actual practice or how the results obtained should be interpreted. They have often completed the first course in statistics with little more than a stock of arbitrary rule-of-thumb procedures and stereotyped generalizations. Because of lack of understanding of basic principles, they have been helpless in the many situations to which these procedures and generalizations do not apply, or have tried, with false confidence and with unfortunate consequences, to apply them to situations for which they are not intended.

In this text and study manual, through extensive use of the Socratic method, an attempt is made to require the student to take a more active role in learning. Much of what has formerly been *presented to* him (for memorization) is here *drawn out* of him through leading questions and suggestive illustrations. The prob-

[1] *Study Manual for a First Course in Statistics. Revised Edition.*

lems and questions for discussion contained in the manual suggest an unusually large number and variety of concrete illustrative situations which may be employed by the student to demonstrate the uses and limitations of each statistical technique considered. These exercises are also intended to help him appreciate what are the most important mathematical properties and essential characteristics of each technique and what is the significance of these characteristics in the interpretation of results. It is left to the student himself, however, to develop these illustrations and to formulate in his own words the generalizations which they support.

2. *These materials stress as much as possible the uses and interpretation of statistics, and minimize as much as possible the mathematical theory of statistics and the mechanics of computation.*

Students in introductory courses in statistics in education and psychology seldom have the mathematical training essential to a ready understanding of the mathematical theory of statistics. The prominence given to mathematical derivations in many courses has, therefore, only contributed to the student's bewilderment and has kept him from devoting his time more profitably to those interpretative aspects of statistics which he can more readily understand. The frequent practice of requiring the student to solve a large number of computational problems has similarly detracted from the time available for consideration of interpretative aspects, and has neither contributed significantly to his understanding nor developed in him any skill in computation as such. In this text and manual, therefore, the mathematical and computational aspects of statistics will be given only the minimum consideration essential to an adequate treatment of the interpretative aspect.

The need for the greatest possible emphasis upon the interpretative aspects of statistics in introductory courses has been admirably stated by Professor Helen Walker, of Teachers College, Columbia University, as follows:[1]

[1] Walker, Helen M., "Problems in the Training of Research Workers." *Journal of Educational Research*, February, 1933.

It is relatively easy to conduct a course in either research methods or statistical methods in such a way that students emerge from it with a confident faith in their ability to discover truth by routine processes, a zeal for applying their new techniques to the first data they can secure, and complete lack of any comprehension of the great variety of ways in which it is possible to reach results that delude rather than enlighten. In the long run, such courses probably do more harm than good. Personally, the author does not believe in teaching, even in elementary classes, the application of a technique whose limitations cannot also be suggested.

An increase in the extent to which educators think in terms of mass data, a growth in the ability to reason statistically, is of enormous value. An increase in the number of persons who compute partial coefficients of correlation with but little idea of their meaning, may be considered of no value at all. It may be a relatively easy task to induct an intelligent student into certain of the computational processes employed in research, to show him certain routines useful in experimentation and in the organization of mass data, but it is a much more difficult task to teach him to think straight, to know what assumptions are implicit in the formulas he employs, to know when those assumptions are inconsistent with the practical situation in which he is working, to draw only such conclusions as are logical, and to make only such generalizations as are justifiable. This difficulty is by no means lessened when such teaching must be conducted en masse and given to students who have carried over from their high-school days a dislike of numbers and an unpleasant emotional reaction to the use of algebraic symbolism.

The longer the writer teaches statistics and the more dissertations she attempts to direct, the more profoundly does she believe that the chief challenge to teachers of research methods is not to produce good computers and not to produce people who can juggle algebraic formulas or who can invent new terminology and new procedures, but to improve the quality of logic which goes into research.

3. *These materials are relatively restricted in number of techniques considered.*

In the belief that it is better for the student to acquire a thorough understanding of a few basic concepts and techniques than a superficial acquaintance with many, only the most fundamental and frequently used statistical techniques are considered in this

text. It has been the author's experience that to develop in the student a satisfactory generalized understanding of any statistical technique requires much more explanatory and illustrative material than has been included in most textbooks. The restriction in the scope of this text and manual permits, without undue demands upon the student's time, the presentation of an adequate amount of such material in relation to each technique considered. The author believes that, if the student develops a thorough understanding of the basic techniques included in this course, he will have little difficulty in interpreting for himself other more specialized and less frequently used techniques if and when the occasion to use them arises.

4. *These materials are designed particularly to develop in the student a critical attitude toward the use of statistical methods in education and psychology.*

Sound statistical judgment involves a keen appreciation of the inherent *limitations* of statistical techniques and of the original data to which they are applied. In the derivation of these techniques, assumptions are frequently made which cannot be satisfied completely in practical applications. The failure to satisfy these conditions necessitates many qualifications in the interpretation of the results obtained. These qualifications have frequently been ignored in the condensed treatments made necessary in many texts by the large number of techniques included. In this text and manual, major emphasis will be placed upon the limitations of statistical methods, upon the many prevalent misconceptions and fallacies in statistical thinking, and upon the many sources of error involved in the use of statistical techniques. By these means, these materials are intended to develop in the student a critical attitude and an appreciation of the fact that statistical methods are an aid to, not a substitute for, common sense.

These and other features of this first course in statistics are explained in greater detail and more definitely from the student's point of view in the introductory chapter. It is highly important

that the student be advised to consider these features carefully at the beginning of the course in order that he may appreciate fully the requirements made of him and may use the materials to the greatest advantage.

These materials have been gradually developed by the author through the use in his own classes of a series of mimeographed preliminary editions which have been successively revised and improved through experience. It is the author's opinion, based upon this experience, that materials of this type may be most effectively used if the course is conducted on a laboratory or work-period basis, in which the instructor dispenses almost entirely with formal lectures and allows the student to spend the major part of each class period in supervised work on the exercises in the study manual. It is, of course, essential that the student be given opportunity at frequent intervals to verify the results of his own reasoning. This may be most readily done through periodic class discussion in which the instructor presents and explains the correct solutions to the exercises after they have been independently attempted by all students. When thus used, these materials should prove adequate for a one-semester undergraduate or graduate course meeting three or four times per week.

The author is deeply indebted to many of his graduate students who, during the course of the successive revisions of experimental editions, offered valuable suggestions for the improvement of the material. He is indebted also to Professor P. J. Rulon, of Harvard University, who read the manuscript and whose criticisms were of great assistance in the final revision.

Grateful acknowledgment is made to Oliver & Boyd, Limited, Edinburgh, and to Professor R. A. Fisher, for permission to reprint the table on page 240.

E. F. LINDQUIST

CONTENTS

CHAP. PAGE

I. INTRODUCTION

The Purposes of Statistical Methods in Education and Psychology 1
The Major Aspects of Instruction in Statistics 3
Importance of the Interpretative Aspect 4
The Organization of Instructional Materials in This Course . . 6
How to Use These Materials in Study 9

II. THE FREQUENCY DISTRIBUTION

Introductory 11
Frequency Distributions of Integral Test Scores 16
Comments on Procedure Suggested for Test Scores 19
Variations in Procedure for Other Than Test Score Data . . . 22
Real Limits and the Meaning of Integral Measures 24

III. PERCENTILES

The Nature of the Measuring Scales on Educational and Psychological Tests 29
Ranks, Percentiles, Deciles, and Quartiles 31
The Computation of Percentile Ranks in Grouped Frequency Distributions 33
Computation of a Given Percentile 35
The Uses and Interpretation of Percentiles 37

IV. GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTIONS

Introductory 39
The Histogram 39
The Frequency Polygon 41
The Cumulative Frequency Curve or the Ogive 42
Supplementary Suggestions for the Construction of Histograms, Polygons, and Ogives 45
"Smoothing" Frequency Polygons and Ogives 46
The Form of a Frequency Distribution 48
The Uses and Interpretation of Histograms, Polygons, and Ogives 50

V. MEASURES OF CENTRAL TENDENCY

Introductory 51
The Arithmetic Mean 52
The "Short" Method of Computing the Mean 55
The Short Method Applied to the Frequency Distribution . . 57
The Median 60

CHAP. PAGE

The Mode 61
The Number of Significant Digits in the Mean 61
The Importance of "Errors" in Statistical Work 67
The Uses and Interpretation of the Measures of Central Tendency 68

VI. Measures of Variability

Introductory 69
Computation of the M.D. 71
Computation of the S.D. 75
Important Characteristics of the Various Measures of Variability 77
The Uses and Interpretation of Measures of Variability . . . 79

VII. The Normal Curve of Distribution

The Characteristics of the Normal Curve 81
Area Relationships under the Normal Curve 85
The Significance of the Normal Curve in Education and Psychology 93
"Fitting" a Normal Curve to a Frequency Polygon or Histogram 99

VIII. Sampling Error Theory

The General Nature of Sampling Studies 102
The Sampling Distribution of the Mean 104
The Standard Error of the Mean of a Random Sample . . . 106
Levels of Confidence 106
Establishing a "Confidence Interval" for the True Mean . . . 108
Testing an Exact Hypothesis about the True Mean 110
The Formula for Estimating the Standard Error of the Mean . 115
The Use of the Standard Error of the Mean with Large Samples 120
The Probable Error of the Mean 123
The Standard Errors of the Median, Q, and S.D. 124
The Standard Errors of Proportions and Percentages. 125
The Standard Error of a Difference. 129
The "Significance" of a Difference; Testing the Null Hypothesis . 130
The Standard Error of a Difference between the Means of Related Variables 134
Small Sample Theory: Establishing a Confidence Interval for the True Mean 136
Small Sample Theory: The Significance of a Difference in Means of Independent Samples 138
Limitations of Sampling Error Techniques Designed for Large Random Samples 139

IX. Standard Measures and Methods of Combining Test Scores

Standard Measures or z-Scores 145
Transforming Raw Scores into Their z-Score Equivalents . . . 147

CHAP. PAGE

T-Scores 149
Composite Measures 150

X. CORRELATION THEORY

The Meaning of Correlation 153
The Significance of Correlation 158
The Need for a Quantitative Measure of Relationship 160
The Selection of an Index of Relationship 161
The Mean z-Score Product 163
The Computation of r 167
Directions for Using the Correlation Chart 169
The Phenomenon of Regression 175
The Use of the Regression Equations in Prediction 182
The Raw Score Form of the Regression Equations 184
The Reliability of Prediction; The Standard Error of Estimate . 186
The Assumption of Rectilinearity 190
Sampling Errors in r 191
Influence of the Variability of Measures upon the Magnitude of r 195
The Meaning of a Given Value of r 198
Causal vs. Casual Relationship 203

XI. CORRELATION TECHNIQUES APPLIED IN THE EVALUATION OF TEST MATERIALS

Introductory 205
The Nature of Measurement in Education and Psychology . . 206
All Measurement Involves Sampling 208
All Mental Measures Are Uncertain as to Meaning 211
The Measurement of Errors in Measurement 213
Test Validity 213
Test Reliability 215
The Coefficient of Reliability 216
Ways of Estimating Coefficients of Reliability 218
The Reliability of a Single Score 220
The Significance of Measures of Reliability 223

APPENDIX 229

INDEX 241

LIST OF FIGURES

PAGE

Fig. 1. Frequency distribution and histogram of weights of a group of boys 40

Fig. 2. Frequency polygon of distribution of weights of a group of boys 41

Fig. 3. Superimposed frequency polygon and histogram of same distribution 42

Fig. 4. Cumulative frequency curve or ogive of weights of a group of boys 43

Fig. 5. Smoothed frequency curve of weights of a group of boys . . 47

Fig. 6. Smoothed ogive of distribution of weights of a group of boys . 48

Fig. 7. Typical forms of frequency distributions 49

Fig. 8. Ordinates under the normal curve 82

Fig. 9. Normal curves of varying ratios of height to "width" . . 83

Fig. 10. Normal curve "fitted" to a frequency polygon 85

Fig. 11. Area relationships under the normal curve 86

Fig. 12. Age distributions of male population of the United States in 1930 94

Fig. 13. Normal curve "fitted" to a histogram 101

Fig. 14. Relation of standard error of the mean to the size of a random sample 117

Fig. 15. Illustrating the 2 per cent confidence interval for the true mean 121

Fig. 16. Scatter-diagram of reading and arithmetic test scores . . . 155

Fig. 17. Illustrating the use of the correlation chart in computing the correlation coefficient . . . . . *inside back cover*

Fig. 18. Illustrating the phenomenon of regression in terms of the distributions of height and weight for a given sample . . . 177

Fig. 19. Illustrating the phenomenon of regression in terms of the scatter-diagram of height and weight measures for a given sample 180

Fig. 20. Showing influence of range of talent upon r 197

Fig. 21. Improvement in accuracy of prediction for increasing values of r 202

CHAPTER I

INTRODUCTION

The Purposes of Statistical Methods in Education and Psychology

STATISTICAL methods are the mathematical techniques used to facilitate the interpretation of numerical data secured from groups of individuals (or groups of observations of a single individual). In education and psychology, the individuals constituting these groups may be human beings variously classified (such as school pupils and teachers or subjects in the psychology laboratory), or they may be administrative units (school classes, school systems, school boards), political divisions (school districts, cities, counties, states), social or religious groups, homes, school buildings — in fact, any entities for which numerical data may be collected. The data gathered may be scores on educational or psychological tests, direct measures of physical traits, enrollment and attendance figures, fiscal data (salaries, incomes, expenditures), census enumerations, school marks, ratings, ages — or any other descriptive facts which may be expressed in numbers.

It is manifest that the student and research-worker in education or psychology, the school administrator, and the classroom teacher all have frequent occasion to interpret masses of data of the types just suggested. It should also be readily apparent that very little meaning can be derived from such data in the unordered form in which they are originally collected. Until they have been compactly and systematically arranged, and until their description has been condensed into a few derived measures which can be conveniently handled, such data cannot be adequately interpreted for any large group or meaningfully compared for different groups.

The statistical methods which will be considered in this introductory course may be classified into three sets of techniques, according to the major purposes that they are intended to serve.

One set of techniques will enable the student to organize group data, to describe and interpret these data in terms of derived measures of central tendency (averages), of variability, and of other characteristics of the group, and to portray these data in graphical form for more convenient interpretation or more ready assimilation.

A second set of techniques will enable the student to describe quantitatively the limits within which he may safely generalize about large groups or populations on the basis of facts derived from relatively small groups or *samples* selected at random from these populations. Nearly all research studies in education or psychology are of the type known as sampling studies. In these, relatively small groups of individuals are observed, investigated, or experimented with for what may be learned in general about all individuals of the same type or from the same population. In any such study, there is always the possibility that the sample of individuals used may not be truly representative of the whole population, since chance or factors beyond the investigator's control will always determine to some extent which individuals will constitute the sample employed. Hence, any fact derived from a sample must always be considered as only an *approximation* to the corresponding "true" fact, that is, to the fact which would have been obtained had the entire population been studied. Under certain conditions of sampling, statistical techniques (sampling error formulas) may be applied to determine quantitatively how *nearly* these obtained facts are likely to approximate the true facts. Proper use of these techniques will help guard against the dangerous tendency to jump to conclusions based on too few observations and will enable the investigator to qualify his generalizations in accordance with the reliability of the facts obtained.

A third set of techniques will enable the student to describe quantitatively the degree of relationship existing between measures of different traits for any group of individuals or between any other types of paired measures. It is a matter of common observation that there is some relationship between, for example, intelligence

and school achievement: pupils of superior intelligence tend also to be superior in achievement and pupils of low intelligence tend to be low in achievement. One cannot, however, obtain any accurate quantitative idea of the closeness of this relationship on the basis of direct observation alone, or make any quantitative comparisons of the relationship for different school subjects. Mathematical techniques are required for these purposes. These techniques are useful in the study of cause-and-effect relationships between mental traits or abilities, in the evaluation of test materials (to describe test validity and reliability), and in estimating or predicting certain unknown measures from known values of related measures.

The Major Aspects of Instruction in Statistics

Entirely apart from these major purposes of statistical methods, there are three aspects of statistics which have been variously stressed in introductory courses in the subject. The first of these has to do with the mathematical theory underlying the derivation of the techniques, the second with the computational procedures involved in practical applications, and the third with the uses of the techniques and the interpretation of results in actual practice.

In this course, the first two of these aspects will receive no more consideration than is essential to an adequate treatment of the third. Mathematical derivations will be considered only in as far as is necessary to demonstrate the reasonableness of the techniques and to draw attention to the important assumptions made in their applications. The specific mathematical skills involved in the derivations presented will in no case go beyond those which are considered as minimum essentials in elementary arithmetic and ninth year algebra. No student, therefore, need feel that he will be seriously handicapped by lack of training in advanced mathematics. No attempt, furthermore, will be made to develop in the student any degree of skill or facility in the computation of statistical measures. A great variety of computational procedures have been developed for the techniques considered in

this course, including many which involve the use of computing machines and electric tabulating equipment. These procedures are so various and complex that any early consideration of them would only confuse the beginning student and would interfere with his attainment of a real understanding of the essential nature of the techniques themselves. In this course, therefore, only the most straightforward and most readily understandable computational procedures will be considered at all. The student will be expected to apply even these procedures in only a very few problems, and then only to contribute to the better understanding of the techniques rather than to develop in him any skill in computation as such.

The maximum amount of the student's time will thus be made available for the consideration of the interpretative aspects of the course. In relation to each of the techniques considered, major emphasis in instruction will be placed upon questions such as the following:

> What are the most significant mathematical properties and major characteristics of this technique? What assumptions are involved in its application?
>
> What specific uses may be made of it? In what types of situations may it be validly applied?
>
> What are its major advantages and limitations in relation to other techniques intended for roughly the same purposes?
>
> How may the results of its application be interpreted?
>
> How must this interpretation be qualified in terms of the unique conditions under which it may be applied?
>
> What common misinterpretations are to be avoided? What common fallacies in statistical thinking are related to the use of this technique?

Importance of the Interpretative Aspect

This course, then, will be essentially a course in the *interpretation* of statistical techniques as they are applied in education and psychology. The mathematical theory of statistics and the me-

chanics of computation will be minimized as much as possible. There are a number of reasons for this distribution of emphasis in instruction. One of these is that the typical student in this course is preparing for college or public school teaching, or plans to enter the field of public school administration, and will not engage in any significant amount of research, perhaps in little more than is involved in meeting the requirements for an advanced degree. If he is to attain any real insight into professional problems, however, he must be prepared to read professional literature with understanding, and must continually keep himself intelligently informed about the current research investigations and experiments reported in professional periodicals. If only as a preparation for such reading, some training in statistics is an essential part of every student's professional equipment. Without such training, most of what he reads professionally will be rendered unintelligible by the frequent recurrence of statistical terms, such as *correlation coefficient*, *probable error*, *standard deviation*, and *significant difference*. To read these materials with comprehension, the student obviously need have no skill in computational procedure, but he must be prepared to evaluate critically the uses that have been made of statistical technique by others, and must be able to check their conclusions against his own interpretations of the results obtained. On the few occasions in which he may need to apply statistical techniques himself, the student can readily look up the preferred computational procedure in available references and handbooks, and will have no difficulty in understanding the directions given if the essential nature of the technique involved is well understood by him. The small proportion of students in this course who will later engage in extensive research on their own account will in any event go on to advanced courses in statistics, in which adequate consideration of the more economical computational procedures involved in large-scale research may be more properly given.

The lack of emphasis upon the mathematical theory of statistics is as much a matter of necessity as of choice. The majority of

students taking this course will not have had the mathematical training essential to an understanding of the derivation of most statistical techniques. Furthermore, a very satisfactory understanding of the uses and interpretations of these techniques can be acquired without tracing step by step the mathematics of their derivation. The typical student can well afford to accept the mathematical derivations on faith and to devote his time more profitably to questions of use and interpretation.

The Organization of Instructional Materials in This Course

The instructional materials employed in this course comprise two volumes: this text and an accompanying study manual. The content of these two volumes has been organized into a number of natural study units, each of which deals with a relatively homogeneous set of techniques that are intended to serve the same general purpose. In one unit, for example, all of the more widely used methods of graphical portrayal of group data are considered together; in another, all of the more important measures of central tendency (averages) are described and compared; another unit treats measures of variability; etc. Each of these units consists of one of the chapters in this text plus the corresponding questions and problems for study and discussion in the manual. In each case the chapter in this text consists of a brief explanation of the mathematical properties and essential characteristics of, and the basic assumptions underlying, each of the techniques considered. The accompanying problems and questions in the study manual are intended to assist the student *to discover for himself* the practical significance of these properties, characteristics and assumptions. The questions in the manual will suggest a very large number and variety of concrete situations which may be employed to illustrate the uses and limitations of each technique, and will draw attention to the implications of the basic assumptions underlying the derivation of the techniques in the interpretation of the results obtained from them. It is left to the student himself to develop these illustrations and to formulate and state in his own words the generaliza-

tions which they support. In order that they may not be inordinately difficult to the student, most of the questions will be presented in a highly leading form, and each step in the necessary reasoning will be so clearly indicated that the student will not be likely to go far astray in his thinking.

This text alone, then, is not intended to constitute a complete discussion of the techniques considered. On the contrary, many of the interpretative statements which ordinarily would be presented in a pat form in the textual discussion will be deliberately omitted in order that the student may be required to reach the same conclusions by his own reasoning. This procedure is based on the sound pedagogical principle that knowledge which the student acquires through his own independent thinking is much more likely to be understood and permanently retained by him than that which he has memorized in the words of another. Essentially, then, the student will be expected to write for himself an important part of what will eventually constitute a complete text, namely, that part which is concerned primarily with the *use and interpretation* of statistical techniques. Each chapter in this text will contain some interpretative materials and explanations, but only to present those concepts which the student cannot reasonably be expected to discover or develop for himself.

A special effort has been made through these materials to develop in the student a critical attitude toward the use of statistical method in education and psychology. Special stress has been placed upon the limitations of each technique, upon the frequent and unavoidable failure to satisfy in practice all the basic assumptions or requirements of each technique, upon the manner in which the conclusions based upon obtained results must be qualified because of such failures, and upon prevalent misconceptions and fallacies in statistical reasoning. In a misguided effort to simplify statistics, many of these necessary qualifications have often been ignored in instruction, and the student has been provided with a number of rule-of-thumb procedures and stereotyped interpretations which, because of the numerous exceptions

to them, in the long run get him into as many difficulties as they help him to avoid. Statistical methods are an aid to, not a substitute for, common sense. Each technique is designed for a certain purpose and for use under certain conditions only. When these conditions are not satisfied, the application of the technique may and often does lead to conclusions that are obviously contradictory to common sense. It is because of just such abuses of statistical techniques that people have developed a distrust of statistics and statisticians. In using these instructional materials, then, the student is strongly advised to strive consciously to develop in himself a highly critical attitude and to be on guard against the easy tendency to over-generalize or to depend unduly upon stereotyped interpretations.

This course will cover only those techniques that are generally considered *essential* in nearly all types of statistical work in education and psychology. Many techniques ordinarily included in a first course in statistics, such as the harmonic and geometric means, the coefficient of variability, the correlation ratio, partial and multiple correlation, and other special but rarely used statistical tools, will be given no consideration whatsoever in this course. This restriction in scope is based on the principle that it is better for the student to acquire a thorough understanding of a few fundamental methods and principles than only a superficial acquaintance with many. If these few techniques are well understood, the student should have little difficulty later in interpreting other techniques for himself if and when the occasion to use them arises.

In addition to the restrictions just noted, as has been previously mentioned, this course will devote the minimum of time to mathematical theory and to computational procedures. These, incidentally, are the aspects of instruction which primarily account for the reputation of being difficult that courses in statistics have so frequently acquired. Everything considered, then, the approach employed in instruction in this course should not present any inordinate difficulties to the student, but, on the contrary,

should constitute one of the *easiest* and most effective procedures by which he may derive from his studies something of real functional value and may acquire a sound statistical judgment.

How to Use These Materials in Study

The content of this course, then, has been presented in a form which is specially designed to encourage the student to think things out for himself and thus arrive at a reasoned understanding of statistical techniques. Consequently, as has already been explained, none of the chapters in this text is intended to be complete in itself. The full significance of some of the statements made in these chapters may not be wholly appreciated by the student until he has also considered the problems and the questions for discussion in the study manual. To use these materials most effectively, the student is advised to employ a procedure somewhat as follows with reference to each unit:

1. Read carefully the complete chapter in this text once or twice before considering any of the problems or questions in the study manual.

2. Begin *writing out* your own answers to the questions in the manual in the order in which they are presented, referring to the chapter in this text wherever necessary. If a question at first seems beyond your comprehension, leave it temporarily and go on to the others. Some of the later questions may give you a hint to how to answer the earlier question.

3. *Do not ask for help* from your instructor or fellow students on any question until you have first done your best to answer *all* of the questions corresponding to the chapter on which you are working. Your final objective, of course, is to arrive at a thorough and reasoned understanding of the techniques considered. Letting the other fellow do your thinking will only interfere with your realization of this objective, even though it may seem the easiest immediate solution.

4. *After* you have done your best to answer the questions yourself, take every opportunity to discuss them with other students

and to compare answers. You will, of course, wish finally to make certain that your answers are correct. If your instructor follows the recommended procedure, he will in due time check your work for you or will consider all of the questions in his lectures or class discussions and will indicate the correct responses to you.

5. When all of your answers have been checked, read the chapter in this text again very carefully and attempt in this final reading to integrate your reasoning and conclusions about the techniques considered.

CHAPTER II

THE FREQUENCY DISTRIBUTION

ANYONE who has worked with test scores collected from a large group of individuals knows that it is extremely difficult to derive any adequate idea of the performance of the group as a whole from the individual measures in the unordered form in which they were originally collected. Consider, for example, the following scores (Table 1) obtained from a group of 100 high-school pupils, each score representing the number of words spelled correctly in a 200-word spelling test.

TABLE 1

SCORES OF 100 HIGH-SCHOOL PUPILS ON A 200-WORD SPELLING TEST

132	126	87	94	107	191	174	105	133	129
171	93	112	123	106	85	105	80	93	63
128	179	105	127	88	112	170	87	154	120
56	82	131	126	141	89	92	109	138	121
164	156	111	89	146	146	163	121	75	115
137	146	56	104	102	100	71	110	134	150
159	102	65	79	126	153	112	159	132	65
139	120	147	68	102	101	96	148	108	152
153	138	93	118	92	98	108	112	67	68
145	86	112	83	103	76	157	96	134	106

To hold so many scores in mind at once is obviously impossible; to derive any generalized concepts of *group* performance from a brief inspection of these scores is extremely difficult. Certain characteristics of the group can, of course, be noted at once. It is not difficult to see that no pupil made a perfect score, that most pupils spelled more than 100 words correctly, that every pupil spelled some words correctly, that a "good many" of the pupils scored between 110 and 150, etc., but such statements hardly constitute a meaningful, accurate, or useful description of the group as a whole, nor do they provide an adequate basis for the evaluation of the relative performance of any individual within

the group. To add very much to the precision and meaningfulness of this description would require a most painstaking "hunt and count" process. Through such a process it is possible, for example, to find the lowest and the highest scores in Table 1, or to determine exactly how many pupils scored above 100 or any other given value, or to determine the exact number of scores between 110 and 150 or between any other pair of values, etc. The student has only to try to do these things for himself, however, to discover how time-consuming is the process, how inaccurate it is likely to be, and how inadequate it is, after all, for the purpose of providing him with a composite mental picture of the group performance.

What is needed, then, is some way of classifying or arranging the scores so as to make more convenient the task of interpreting them as a group. One obvious possibility would be to rearrange the scores in order of their size, from the highest to the lowest. With such a rearrangement it would be very much easier to note the highest and lowest scores, or to count the number of scores between any two given values, or to evaluate roughly any given score by noting how far down in the list it occurs, etc. Rearrangement of the scores in this manner, however, would also require a considerable amount of time, and would still not enable one to note quickly and easily the performance of the pupils as a group.

A better procedure would be to list, in order of their size, all *possible* score values within the range of all the scores obtained, and then to indicate after each score value the number of times it occurred, as has been done in Table 2.

It is immediately evident that this form of arrangement markedly facilitates interpretation. The more frequently occurring scores now stand out clearly, the points of concentration are quite readily noted, the total number of scores may be quickly secured by simply adding the numbers in the frequency column, the number of scores between any given values can likewise be readily obtained through simple addition, etc. Most important is the fact that this form of table shows in a graphic way how the scores

TABLE 2

SIMPLE FREQUENCY DISTRIBUTION OF THE SPELLING SCORES IN TABLE 1

(Intervals of one unit)

S	*f*	*S*	*f*	*S*	*f*	*S*	*f*	*S*	*f*
191	1	164	1	137	1	110	1	83	1
190		163	1	136		109	1	82	1
189		162		135		108	2	81	
188		161		134	2	107	1	80	1
187		160		133	1	106	2	79	1
186		159	2	132	2	105	3	78	
185		158		131	1	104	1	77	
184		157	1	130		103	1	76	1
183		156	1	129	1	102	3	75	1
182		155		128	1	101	1	74	
181		154	1	127	1	100	1	73	
180		153	2	126	3	99		72	
179	1	152	1	125		98	1	71	1
178		151		124		97		70	
177		150	1	123	1	96	2	69	
176		149		122		95		68	2
175		148	1	121	2	94	1	67	1
174	1	147	1	120	2	93	3	66	
173		146	3	119		92	2	65	2
172		145	1	118	1	91		64	
171	1	144		117		90		63	1
170	1	143		116		89	2	62	
169		142		115	1	88	1	61	
168		141	1	114		87	2	60	
167		140		113		86	1	59	
166		139	1	112	5	85	1	58	
165		138	2	111	1	84		57	
								56	2

are *distributed* along a linear scale of values. This latter advantage would be more evident were the scores arranged in a single vertical column (which is the usual practice) instead of in five separate columns as the limitations of space here necessitated.

Table 2 has the serious disadvantage of bulkiness. With the scores distributed over so wide a range, too much space is necessary to list all possible values. This fact suggests that the interpretation of the data would be further facilitated if Table 2 were *condensed* by indicating the number of scores falling within *equal intervals* along the linear scale, instead of indicating the number of times each integral value occurred. This has been done in Table 3. In this table, illustrating what is known as a *grouped* frequency distribution, each interval is identified in the "S" column (S rep-

resents scores or measures) by the highest and lowest integral scores in the interval, and each "frequency" value indicates the total number of scores contained in the corresponding interval. In this case, each interval includes three units along the scale — any other size of interval could, of course, have been employed.

TABLE 3

GROUPED FREQUENCY DISTRIBUTION OF THE SPELLING SCORES IN TABLE 1

(Intervals of three units)

S	*f*	*S*	*f*
191–193	1	122–124	1
188–190		119–121	4
185–187		116–118	1
182–184		113–115	1
179–181	1	110–112	7
176–178		107–109	4
173–175	1	104–106	6
170–172	2	101–103	5
167–169		98–100	2
164–166	1	95– 97	2
161–163	1	92– 94	6
158–160	2	89– 91	2
155–157	2	86– 88	4
152–154	4	83– 85	2
149–151	1	80– 82	2
146–148	5	77– 79	1
143–145	1	74– 76	2
140–142	1	71– 73	1
137–139	4	68– 70	2
134–136	2	65– 67	3
131–133	4	62– 64	1
128–130	2	59– 61	
125–127	4	56– 58	2

Obviously, the degree of compactness in a table of this kind will depend upon the size of the interval into which we decide to classify the scores. We can secure successive degrees of compactness, for example, by using an interval of 5 units, as in Table 4; or of 10 units, as in Table 5; or of 20 units, as in Table 6; or of 50 units, as in Table 7.

It should be noted that Tables 3–7 differ in one fundamental respect from Table 2. In Table 2 each original score is retained intact, that is, the exact value of *each* score is indicated. In the later tables, however, we lose in varying degrees the *identity* of the original scores. For example, we may read in Table 4 that

TABLE 4
GROUPED FREQUENCY DISTRIBUTION OF THE SPELLING SCORES IN TABLE 1
(Intervals of five units)

S	*f*
188–192	1
183–187	
178–182	1
173–177	1
168–172	2
163–167	2
158–162	2
153–157	5
148–152	3
143–147	5
138–142	4
133–137	4
128–132	5
123–127	5
118–122	5
113–117	1
108–112	10
103–107	8
98–102	6
93– 97	6
88– 92	5
83– 87	5
78– 82	3
73– 77	2
68– 72	3
63– 67	4
58– 62	
53– 57	2

TABLE 5
GROUPED FREQUENCY DISTRIBUTION OF THE SPELLING SCORES IN TABLE 1
(Intervals of ten units)

S	*f*
190–199	1
180–189	
170–179	4
160–169	2
150–159	9
140–149	7
130–139	10
120–129	11
110–119	9
100–109	16
90– 99	9
80– 89	10
70– 79	4
60– 69	6
50– 59	2

TABLE 6
GROUPED FREQUENCY DISTRIBUTION OF THE SPELLING SCORES IN TABLE 1
(Intervals of twenty units)

S	*f*
180–199	1
160–179	6
140–159	16
120–139	21
100–119	25
80– 99	19
60– 79	10
40– 59	2

TABLE 7
GROUPED FREQUENCY DISTRIBUTION OF THE SPELLING SCORES IN TABLE 1
(Intervals of fifty units)

S	*f*
150–199	16
100–149	53
50– 99	31

there were 6 scores in the interval 93–97, but we have no way of telling how these 6 scores were distributed *within* the interval itself. We are therefore unable to determine from Table 4 the frequency of occurrence of any single score value. However, we can now more conveniently derive an adequate idea of how the scores were *distributed*, in general, over the entire range. The coarser the interval, the more serious this loss of identity of individual scores becomes. In Table 7 it causes most of the scores to fall in a single interval, and thus hides most of the characteristics of the original distribution.

The size of the interval to be used is thus a matter of arbitrary choice — dependent upon the nature of the data, upon the uses to which the grouped frequency distribution is to be put, or upon the kind of interpretations that one desires to draw from it. If high precision in description is desired, if fluctuations in frequency over small parts of the range are to be studied, and if the number of scores tabulated is large enough to permit such detailed study, then the interval used should be small, as is illustrated in Tables 2, 3, and 4. If, however, only a very rough picture of the distribution of scores is needed, a very broad interval, as in Table 6 or even in Table 7, may prove quite satisfactory.

It is therefore dangerous to set up any general rule concerning the number of intervals into which a series of measures should be classified. Experience has shown, however, that for most types of data there is usually no real need for more than 20 intervals, and that the use of less than 12 intervals usually obliterates too many important characteristics of the distribution.

The purpose of the preceding discussion has been to point out as simply as possible the *major* purposes, advantages, and limitations of the frequency distribution as a means of presenting group data. It now becomes necessary to consider more specifically the detailed questions that arise in the construction of frequency distributions of data of various types.

Frequency Distributions of Integral Test Scores

Different types of data require different methods of handling — factors important in one situation are not important in others. It is therefore impossible to provide any single set of rules that the student can apply in any and all situations and to all types of data. The data for which the majority of students in this course will have to construct frequency distributions, however, will most often consist of integral scores on educational and psychological tests. The construction of frequency distributions for such data is a relatively simple matter, and will therefore be considered first. The procedure required for other types of data can then be more easily explained as variations of this simpler procedure.

STEPS IN THE CONSTRUCTION OF A GROUPED FREQUENCY DISTRIBUTION OF INTEGRAL TEST SCORES

1. Arrange a *data sheet* with the three headings *Score*, *Tabulation*, and *Frequency*. The abbreviated notations *S*, *Tab.*, and *f*, respectively, may be used if desired. (See illustration in Table 8.)
2. Determine the *range* of the scores: Find the highest score and the lowest score in the series. Find the difference between these scores. This difference is called the *range* of the scores.
3. Divide the range by 15. (Carry the result to only one decimal place.)
4. Select from the following preferred list the number nearest the quotient obtained in Step 3. The number thus selected will represent the size of the interval to be used.
   Preferred intervals: 1, 2, 3, 5, 7, 10, 15, — or any higher multiple of 5.
5. Write the integral limits of each interval, in descending order, in the first (*S*) column of the table. Begin at the top with the interval which contains the highest score and continue until the interval containing the lowest score is reached. The "integral limits" of an interval are the highest and lowest scores in the interval. Determine these limits as follows:
   a) When the number of units in the interval (as selected in Step 4) is an odd number, find the *multiple* of this number which is nearest to the highest score in the series. Select the integral limits of the upper interval so that this multiple is the middle score in the interval. The limits of the other intervals will, of course, be automatically determined when those of the top interval are fixed.
   b) When the number of units in the interval is an even number, let the lower integral limit of each interval be a multiple of this number.
6. Tabulation: Begin with the first score in the original un-

ordered list of scores. Determine in which interval this score is included. Place a tally mark, in the *Tabulation* column, opposite the appropriate interval. Proceed in the same way for the remaining scores in the original list. The subsequent counting is facilitated if every fifth mark in a row is made slanting across the preceding four marks.

7. Count the number of tally marks opposite each interval and write the result in the frequency column. Add the numbers in the frequency column as a partial check on the accuracy of tabulation. The result should agree with the total number of scores in the original list.

ILLUSTRATIVE PROBLEM

These steps may be made clearer by considering their application to the data in Table 1. These scores have been properly arranged in a frequency distribution in Table 8 following. The steps in the construction (numbered to correspond to those used in the preceding general description) were as follows:

1. A data sheet was first prepared. The form of this data sheet is shown in Table 8.
2. The highest score in Table 1 is 191. The lowest is 56. The range is therefore 135.
3. 135 divided by 15 is 9.0.
4. The number in the list of preferred intervals nearest to 9 is 10. The scores were therefore grouped into intervals of 10 units each.
5. In accord with Step 5 *b* in the preceding rules, the lower integral limit of the interval containing the highest score (191) is 190. The upper limit of this interval is then 199. The values 190–199 were therefore written at the top of the *Score* column, and the rest of the interval limits were determined by building down from this interval.
6. The first score in the original list is 132. The first tally mark was therefore placed in the *Tabulation* column opposite

the interval 130–139. The second score is 171. The second tally mark was therefore placed opposite 170–179. The procedure was the same for the remaining scores.

7. The tally marks in each row were counted and these numbers placed in the corresponding positions in the frequency column. The sum of these frequencies was found to be 100, the same as the number of scores in Table 1.

TABLE 8

FREQUENCY DISTRIBUTION OF SCORES IN TABLE 1: SCORES OF 100 HIGH-SCHOOL PUPILS ON A 200-WORD SPELLING TEST

S	*Tab.*	*f*
190–199	/	1
180–189		
170–179	////	4
160–169	//	2
150–159	//// ////	9
140–149	//// //// /	7
130–139	//// ////	10
120–129	//// //// /	11
110–119	//// ////	9
100–109	//// //// //// /	16
90– 99	//// ////	9
80– 89	//// ////	10
70– 79	////	4
60– 69	//// /	6
50– 59	//	2

Comments on Procedure Suggested for Test Scores

As has already been noted, different situations may call for different procedures, even for the same type of data. The steps suggested on pages 17 and 18 only describe the procedure that may *usually* be followed. There are many situations, however, in which exceptions must be made to these rules. It is therefore essential that the reason for each step be clearly understood, in order that the student may recognize the situations in which variations are desirable.

Step 3: Dividing the range by 15 obviously results in a number which is contained in the whole range 15 times. The procedure suggested in Steps 3 and 4, then, will result in about 15 intervals for the whole distribution. Experience has shown that approxi-

mately this number of intervals is adequate for most purposes. If for any reason it is desired to group the scores into finer or coarser intervals, the number of intervals desired should be substituted for 15 in this step.

Step 4: The suggestion in Step 4 is made for reasons of convenience only, and has no bearing on the accuracy of any results obtained from the frequency distribution. We could, of course, dispense with this step and use as the size of the interval the "rounded" integral value of the quotient obtained in Step 3. For example, in the illustrative problem, we could have used an interval of 9. There are certain objections, however, to this procedure. One is that people in general are multiple-of-five or multiple-of-ten "minded." It is easier for them to think in terms of multiples of 5 or 10 than in terms of numbers such as 6, 9, 13, 16, 19, etc., which are representative of the numbers that we would frequently get as the size of the interval if we used the rounded quotient of Step 3 directly. In general, the use of an interval containing an odd number of units results in a more convenient midpoint for each interval. Because of the loss of identity of the original scores, it will be necessary in later computations to use the midpoint of each interval to represent the value of all the scores contained in the interval. If the interval contains an even number of units, the midpoint will be a decimal value and therefore inconvenient to use. In any interval containing an odd number of units, however, the midpoint will be an integral number.

One more advantage of the limitation in choice suggested in Step 4 is that it results in uniformity in the solutions of problem work handed in by the class. This is administratively quite important from the point of view of the instructor or the reader who has to correct these problems. This step should therefore be rigidly observed by the student in all problems in this course that are not affected by any of the qualifications made elsewhere in this chapter.

Step 5: After the size of the interval has been selected, where to "start" each interval must still be decided. For example, if an

interval of three units is to be used, and if the highest score in the series is 113, we could write the limits of the top interval as 111-113, or 112–114, or 113–115. A definite basis for settling this type of question has been provided in Step 5. The provision in *a* under Step 5 results in a midpoint that is easy to "read," especially where the interval is one of 5 units or an odd multiple of 5 units. As suggested in *b* under Step 5, it is self-evident that a grouping of 10–19, 20–29, 30–39, etc., is more natural and convenient than, say, 13–22, 23–32, 33–42, etc. For even intervals, other than intervals of 10, Step 5 is important only to secure uniformity in the solutions of problems assigned in this course.

A specific illustration of Step 5 *a* might be helpful. Consider a series of scores in which the highest score is 151 and the lowest is 54. The range is then 97, which divided by 15 yields 6.5. We therefore select the interval of 7 from the preferred list. The multiple of 7 nearest 151 is 154, which will be the midpoint of the top interval. The limits of the interval may then be determined by counting out three in either direction from 154, and are 151 and 157 respectively. Beginning with the interval 151–157, we then build down in the score column to obtain the limits of the remaining intervals. It will be noted that *each* interval will have as a midpoint a multiple of 7.

An exception to Steps 4 and 5: "Natural" Grouping: Sometimes the measures in a series will lend themselves more naturally to another grouping than that determined by the "rule-of-thumb" suggestions in Steps 4 and 5. For various reasons, the measures may tend to concentrate at or about points on the scale which are equal distances apart. For example, salaries of high school teachers are usually multiples of $50 or $100. Salaries such as $913.00, $879.00, $1192.00 will be found much less frequently than salaries of $900, $950, $1050, etc. In such instances, the interval "imposed" on these data should be equal to or a multiple of this uniform distance between successive points of concentration, and the midpoint of each interval should coincide with one of these points (or should be such that within each interval these points

are placed as symmetrically as possible with reference to the midpoint). If the intervals are not so chosen, the measures within each interval will show an unbalanced distribution, and a systematic error will be introduced into any computation in which the midpoint of each interval is used to represent the average value of the measures in the interval.

Variations in Procedure for Other Than Test Score Data

The problem of constructing frequency distributions of test scores is simplified by the fact that such scores are almost invariably expressed only in integral values; that is, fractional test scores are of very rare occurrence. There are many situations, however, in which continuous variables are measured to the nearest given fraction of a whole unit. Heights of individuals, for example, may be measured to the nearest eighth or tenth of an inch. It may also happen, in such cases, that the range of measures is so narrow that in order to get sufficient discrimination between the measures in the frequency distribution an interval of a fraction of a whole unit must be used. For example, heights of individuals might be classified into intervals of one-half or one-quarter inch.

In cases where the measures to be tabulated have been determined to the nearest multiple of a given fraction of a unit (for example, to the nearest multiple of a sixteenth of an inch, tenth of a pound, or fifth of a second) the following rules may usually be applied.

1. Divide the range by the number of intervals desired (usually 15). Choose as the size of the interval that *convenient* multiple of the given fraction which is nearest this quotient.
2. Let the interval "limits" (corresponding to the integral limits in distributions of test scores) be multiples of the given fraction. If the interval is an odd multiple of the given fraction, let the *midpoint* of each interval be a multiple of *the size of the interval.* If the interval is an even multiple of the given fraction, let the lower "limit" be a multiple of the

size of the interval. (As will be noted later, these "limits" are not the *real* limits of the intervals.)

3. Proceed as in Steps 6 and 7 on pages 17 and 18.

The following illustration may help to make these rules clear. Suppose a measure of height of each of a number of individuals has been determined to the nearest tenth of an inch. Suppose the tallest individual is 72.8 inches and the shortest 64.3 inches in height. The range of the distribution would then be 8.5 inches. One fifteenth of this range would be .566 inches. Since it would be futile to express the size of the interval in units finer than those used to express the measures themselves, we round this result to the nearest tenth of an inch, or to .6 inch. This value could be used as the size of interval, but in this case a half-inch interval would be more convenient to use and would result in a number of intervals sufficiently close to that desired (15).

In accord with the second rule, the midpoint of each interval would be a multiple of .5. The *midpoints* would therefore run as follows: 64.5, 65.0, 65.5, 66.0, and so on up to 73.0, which would be the midpoint of the interval containing the highest measure.

In accord with the third rule, we would express the "limits" of the intervals in the *S* column in tenths of an inch, as follows: 64.3–64.7, 64.8–65.2, 65.3–65.7, and so on up to 72.8–73.2, the limits of the top interval. (As will be explained later, these are not the *real* limits of the intervals.)

It is very important to note that these rules for "other than test score" data again only constitute a rule-of-thumb procedure that is *usually* satisfactory, and to remember that there are many situations where exceptions may and should be made to these rules.[1] The exceptions may be of the same nature as those described

[1] The moral, "Beware of rule-of-thumb procedures," should be continually preached throughout a first course in statistics. Such rules of convenience are very valuable devices for simplifying and facilitating routine statistical work when applied by persons who understand their limitations, but they are also extremely dangerous in the hands of beginners in that they tend to foster an uncritical attitude. Throughout this course the student should strive consciously to develop the habit of examining critically each new technique, of remaining keenly aware of the assumptions that are made in its development and that are necessary in its application, and of

in the case of integral test scores. If a very large number of measures are to be distributed, and if changes in frequency within a small part of the whole range are to be studied, then more than 15 intervals may be desirable. Similarly, there may be situations where a small number of intervals will yield all the information that is desired. Again, if the measures tend to group about equally spaced points into "natural" intervals, the student should not hesitate to depart from the rule-of-thumb to let the imposed interval conform to this natural interval.

Real Limits and the Meaning of Integral Measures

For the sake of simplicity in presentation, certain important considerations have been omitted from or only very briefly mentioned in the preceding discussions. These have to do with the *real* limits of intervals, with the distinction between real and integral limits, with interval midpoints (sometimes called class-values), and with the meaning of an integral score or measure.

The numerical data collected in statistical work in education and psychology may be classified as either *continuous* or *discrete*. *Discrete* data are always expressed in whole numbers or integers, and ordinarily represent counts of indivisible entities or units. The linear scales employed with discrete data are always characterized by *gaps* at which no real measures may ever be found. School enrollments, sizes of families, and census enumerations are examples of discrete data. *Continuous* measures are those which may conceivably be found at *any* point along a continuous linear scale. Weights of school children, for example, may be measured in as fine units as we please, and (between certain limits) there is no point along the scale of weights at which we may not conceivably find the weight of some pupil, no matter how finely

noting the characteristics of situations in which exceptions must be made to any arbitrary rules of convenience. It is the failure of statisticians to develop this attitude, and the consequent careless application of techniques to situations for which they are not intended, that lead so often to conclusions which are obviously absurd and contrary to common sense, and that hence have weakened the confidence of people in general in "statistics."

we subdivide the scale. Any trait or characteristic in which individuals may differ by amounts which would approach zero if sufficiently refined measuring instruments were employed may be considered as a continuous variable. Intelligence, school achievement, arithmetic ability, height, and strength are examples of continuous variables.

While continuous variables may theoretically be measured in as fine units as we please, the measuring instruments which we employ in actual practice are usually relatively crude, and the measures obtained are only *approximations* to absolutely accurate determinations. We seldom measure weights of persons, for example, in smaller units than pounds, or ages in smaller units than months or years.

Ordinarily, measures of continuous variables are taken to the *nearest* multiple of some convenient unit. Weights, for example, are usually read to the *nearest* pound. If, when one weighs himself, he finds that the pointer on the scale is closer to 146 than to 145, he reads his weight as 146 pounds. When a person gives his weight as 181 pounds, we interpret this to mean that his real weight is nearer 181 than either 180 or 182 — that is, that it is somewhere between 180.5 and 181.5. Similarly, height is usually read to the nearest inch, or sometimes to the nearest half or quarter of an inch, and performance in the hundred-yard dash is timed to the nearest tenth or fifth of a second.

In a frequency distribution of weights, then, an interval identified by the integral limits 163–167 must be considered as really extending from 162.5 up to 167.5 pounds, since 163 represents any real weight of from 162.5 to 163.5 and 167 any weight from 166.5 to 167.5.

Since most measurements expressed as integers may be considered as having been taken to the *nearest* integral values, the *real* limits of an interval in a frequency distribution should usually be considered as extending .5 of a unit on either side of the integral limits. The so-called integral limits are then not *limits* at all, but only the highest and lowest whole numbers *within* the interval.

This observation, as the student will later discover, is of considerable significance in the computation of certain statistical measures derived from frequency distributions.

Scores on all kinds of educational and psychological tests should, in the opinion of the writer, be interpreted in the manner just suggested. Some writers on statistical procedures, including writers of elementary statistical textbooks, have maintained that for certain types of tests an integral test score should be considered as representing an interval which extends from the given integral value up to the next integer above. They would contend, for example, that a score of 7 on an arithmetic problems test should be interpreted as representing a unit interval of 7.00–7.99, on the grounds that a pupil may have begun work on but not have had time to complete an eighth problem. As will be pointed out later, however, scores on educational or psychological tests never have any *absolute* significance, but only indicate the *relative* status of an individual in a group. The addition (or subtraction) of any constant amount to (or from) *all* measures alike clearly cannot influence the relative status of any measure. This being the case, no advantage can possibly be gained by making, in the case of test scores, any exception to the general rule given in the preceding paragraph, while to make such an exception will only unnecessarily complicate the procedures and confuse the student. Furthermore, to consider an integral test score as the *lower limit* of a unit interval is inconsistent with the known fact that errors of measurement due to test unreliability are equally likely to occur in *either* direction. For these reasons, it is suggested that integral scores on all educational and psychological tests and scales be considered as *midpoints* of unit intervals, and that the *real* limits of any interval in a grouped frequency distribution of such scores be considered as extending .5 of a unit on either side of the integral limits.

It is important to note that there are certain types of data (other than test scores) which require a different treatment from that suggested in the preceding paragraphs. In the collection of chronological age data, for example, it is the usual practice to

express an individual's age in years on his *last* birthday. Ordinarily, we think of a "13-year-old boy" as one who is anywhere between 13 and 14 years of age. A boy whose age is 13 years 7 months would usually be tabulated as a 13-year-old. Similarly, "five years of teaching experience" would, as such data are often collected, mean more than five but less than six years of experience. For data collected in this manner, we *must*, in order to avoid important systematic errors, consider an integral measure as the *lower limit* of a unit interval. The *real* limits of any interval in a grouped frequency distribution of such data would have to be considered as extending from its lower integral limit up to the lower integral limit of the next interval above. The real limits of the interval 16–18 would in this case be 16.00–18.999. It should be noted, however, that age data *may* be and often are otherwise collected. Many questionnaires, for instance, include the item "Give your age in years to your *nearest* birthday." In this case, of course, no exception should be made to the usual interpretation of integral measures and real limits. How an interval in a grouped frequency distribution should be interpreted, then, depends upon the manner in which the data were collected, or in which the measurements were made.

The *midpoint* of any interval is always midway between the *real* limits, however these real limits may be placed with reference to the integral limits. The midpoint of the interval 16–17 would, in the case of a distribution of integral test scores, be halfway between 15.5 and 17.5, or 16.5. The interval 16–17 in a distribution of ages "to last birthday" would have a midpoint of 17.00, halfway between 16.00 and 17.999. The midpoint is significant because it is so frequently used in statistical computations to represent the average value of the measures within the interval (the identity of the original measures having been lost).

It might appear, because of the discontinuous character of *discrete* data, that the preceding suggestions concerning the determination of real limits and midpoints may not be applied when the data are discrete. Some textbook writers, in fact, have given

special consideration to the construction and interpretation of frequency distributions of discrete data, and have described slightly modified procedures for their treatment. In the writer's opinion, this has only served to confuse the student with qualifications which are of no practical consequence so far as the majority of students are concerned. In this course, therefore, no distinctions in the statistical treatment of continuous and discrete data will be made, either with reference to the frequency distribution or to techniques later considered.

CHAPTER III

PERCENTILES

The Nature of the Measuring Scales on Educational and Psychological Tests

THE linear scales along which the scores on educational and psychological tests are expressed differ in several fundamental respects from those employed in physical measurement. In physical measurement each scale is based upon a constant *unit*, and measurements are made from a reference point which either represents an absolute zero or has a known relation to the absolute zero. The units employed in physical measurement are also usually capable of description in more fundamental terms, which permit us to transform measures from one system of measurement into another — for example, to transform inches into centimeters, ounces into grams, or degrees Centigrade into degrees Fahrenheit. Scores on educational or psychological tests have none of these characteristics. A test score usually represents the number of test items to which the person tested has made the correct response. For example, if a pupil makes a score of 80 on a 150-word spelling test, this score indicates that he has spelled 80 of the words correctly. The meaningfulness of this score depends, of course, upon the range and distribution of difficulty of the words constituting the test. If the test contains 100 very easy words, this score does not necessarily mean that the individual making it is a very good speller. On the other hand, if the test consists exclusively of very difficult words, a score of 80 may represent a remarkable performance.

The meaning of a *difference* between two scores on the same test likewise depends upon the range and distribution of difficulty of the items. Suppose, for example, that one 150-word spelling test consists of words which are evenly distributed over a very wide range of difficulty and that a second 150-word test consists of

words all of which are very nearly of the same difficulty as the average word in the first list. A pupil scoring 120 on the first test is, then, probably a very much better speller than one scoring 20, since the hardest words spelled by the first pupil would be very much more difficult than the hardest words spelled by the second. On the second test, however, two pupils making the scores of 20 and 120 respectively may not differ in ability by nearly so much, since the words spelled by the first pupil would be only slightly easier than the most difficult of those spelled by the second. For similar reasons, a given *difference* between two scores on the same test might have a different significance at different points along the scale. Suppose that on a certain test pupil A spelled 30 words, B spelled 60, and C spelled 90 words correctly. Suppose, further, that the test contains 70 very easy words and 70 very difficult words, with only 10 words of intermediate difficulty. In this case the difference in ability between C and B would probably be very much greater than that between B and A, since A and B might both have been able to spell only very easy words while C was able to spell some of the very difficult. On the scale of scores for this test, then, the "unit" employed would be much larger at some points than at others. Similarly, a score of zero on a test of this kind would have no absolute significance. If a pupil fails to spell any word in a spelling test — that is, if he makes a score of zero — obviously it does not follow that he has *no* spelling ability, since other easier tests might contain some words that he can spell.

In general, then, the magnitude of the "unit" employed on the scales for any educational or psychological test depends upon the number of test items and upon the distribution of their difficulty for the test as a whole. Since the number of items making up the test is arbitrarily determined by the test author, and since the difficulty of the individual items and the form of the distribution of difficulty for all items cannot be accurately anticipated or controlled by him but usually is more or less accidentally determined, the magnitude of the "unit" employed is indeterminate

and usually fluctuates in value even within the same scale. The meaning of a given score on any test is *unique* to that test; that is, it is not exactly the same as on any other test. The meaning of the arbitrary zero point on any scale of test scores is also unique to the test and never corresponds to an absolute zero. Furthermore, scores on such tests can never be described in more fundamental terms by means of which direct comparisons of scores may be made from test to test or readings on one scale transposed into those on another.

For these reasons, a single score obtained on most educational or psychological tests has little if any *absolute* significance — that is, it is not capable of meaningful interpretation when considered alone. Neither can it be meaningfully compared directly with a score obtained on another test. Scores on such tests usually have relative meaning only; that is, they are ordinarily useful only to determine an individual's *relative status* in a given group. The fact that a given pupil has made a score of 70 on a test in United States history, for example, in itself tells us nothing about the quality or magnitude of his achievement. In order to interpret this performance, we must not only be intimately acquainted with the test itself but must also know what scores have been made on the same test by other pupils in a group to which the given individual belongs, and must know something about the nature of that group, that is, whether it is made up of college or high school or elementary school pupils, what kind or amount of instruction they have had, what is the level and range of their intelligence, etc.

Ranks, Percentiles, Deciles, and Quartiles

Because of the characteristics of test scores that have just been considered, it is essential in the analysis of test data that we have some means of deriving from the original or raw scores other measures which are directly indicative of the relative status of each of these scores in a distribution of such scores. Such measures of relative status will enable us to interpret more adequately a single test performance and to make comparisons of performances

on different tests. One of the devices commonly used for this purpose is that of determining the rank of each score in the series of scores in which it is found. The *rank* of the score indicates its position in a series when all scores have been arranged in order of magnitude. A rank of 30 for a given score would indicate that the score is the 30th from the top (or from the bottom) when all scores have been arranged in order of size. The meaningfulness of any given rank obviously depends upon the number of scores in the series. To rank 30th in a group of 50, of course, does not mean the same thing as to rank 30th in a group of 100. For this reason, ranks are ordinarily expressed in relative terms as *percentile* ranks. The percentile rank of a given score in a distribution is the per cent of measures in the whole distribution which are lower than the given score. If, for example, an individual makes a score higher than that which is made by 89 per cent of the individuals in a given group, we would say that he is at the 89th percentile. In general, then, the pth percentile in a distribution of scores or measures may be defined as that point on the scale below which p per cent of the cases fall. Thus the 90th percentile is the point below which 90 per cent and above which 10 per cent of the measures lie. The 75th, 50th, and 25th percentiles are known as the quartile points in the distribution, or simply as the quartiles. The 75th percentile is the Upper Quartile, and is usually denoted by Q_3. The 25th is the Lower Quartile or Q_1, while the 50th is the Middle Quartile or median. The even 10th percentiles are often referred to as the *deciles*. Hence, the 20th percentile is the second decile, the 30th the third, etc. According to the definition given above, the 100th percentile would be a point above the highest score earned, and the zero percentile below the lowest, and hence could not correspond to any actual scores. In practice, however, the highest and lowest scores are frequently arbitrarily considered as corresponding to the 100th and zero percentiles respectively.

The student should distinguish carefully between the terms *percentile* and *percentile rank*. The *percentile rank* of a given

score is the number representing the *per cent* of the cases in the total group lying below the given score value, while the percentile is the *score* or measure below which a given per cent of the cases lie. The 28th percentile in a distribution of weights may be 112 pounds, but the percentile rank of an individual of this weight in this distribution is 28.

The Computation of Percentile Ranks in Grouped Frequency Distributions

In any given frequency distribution of scores, we may wish (a) to determine the percentile rank of a given score, or (b) to determine the score with a given percentile rank, that is, to determine a given percentile. We shall consider first the procedure involved in determining the percentile rank of a given score.

If all of the original scores were arranged in order of magnitude (and if there were no ties in rank), we could determine the percentile ranks by dividing the percentile scale (of from 0 to 100) into as many equal divisions as there are individuals in the group, and by assigning as the percentile rank of each individual the *midpoint* of the division in which he belongs. For example, if there were 40 individuals in the group, we would divide the scale of from 0 to 100 into 40 equal divisions. The first division would extend from 0 to 2.5; the second, from 2.5 to 5.0; etc. The individual ranked third from the bottom would then belong in the division 5.0–7.5, and his percentile rank would be the midpoint of this division, or 6.25. Similarly, the individual scoring at the bottom of the list would belong in the division 0.0–2.5, and would have a percentile rank of 1.25.

If we are to work directly from a grouped frequency distribution of the scores, the fact that we have lost the identity of the original scores requires that we follow a somewhat more complicated procedure. Suppose, for example, that we wish to find the percentile rank of the score 95 in the distribution presented in Table 9. To do this we must first determine *how many* scores in the distribution lie below the score 95. The number of scores lying

below 95 is equal to the number of scores in all of the *intervals* below that which contains this score, *plus* the number of scores *within* the interval 90–99 which are below 95. The number of scores below the interval 90–99 can be readily determined by adding the class frequencies below this interval. In this case, the sum of these frequencies would be $3 + 1 + 2 + 5 + 8 + 8 + 8 = 35$. To determine how many of the scores *within* the interval 90–99 lie below 95, we must make an assumption concerning the way in which these scores are distributed throughout the interval. The most convenient and reasonable assumption that we can make is that these scores are *evenly* distributed within the interval. The point 95 is 5.5 units above the lower real limit (89.5) of this interval. Since the interval consists of 10 units, 5.5 units represents $\frac{5.5}{10}$ of the distance from the bottom to the top of the interval. Since we have assumed that the scores are evenly distributed within the interval, it follows that $\frac{5.5}{10}$ of the 11 scores in the interval or 6.05 of these scores will lie below 95. Hence, the total number of scores below 95 will be $35 + 6.05 = 41.05$. This number is 51 per cent of the total number of cases in the distribution $\left(\frac{41.05}{80} \times 100\right)$. Hence, the score of 95 has a percentile rank of 51.

If the percentile ranks of many score values in the distribution are to be computed by this method, it is best to begin by preparing a cumulative frequency (*cf*) column to the right of the frequency column in the distribution, as has been done in Table 9. The cumulative frequency column is prepared by "adding in" successive class frequencies from bottom to top. The entry opposite the lowest interval is the frequency in that interval; the entry opposite the second interval is the sum of the frequencies in the first and second intervals; the entry opposite the third interval is the sum of the frequencies in the first, second, and third intervals, etc. The entry opposite the top interval must, of course, be

equal to N, the total number of cases in the distribution. The following rules may then be applied in general to compute the percentile corresponding to any given score.

1. Subtract from the given score value the lower *real* limit of the interval in which it is contained.
2. Divide this difference by the size of the interval.
3. Multiply this quotient by the frequency in the given interval.
4. Add this product to the cumulative frequency below the given interval.
5. Divide the result by the total number of cases (N) and multiply by 100. This last result should rarely be carried to more than one decimal place, and ordinarily should be rounded to the nearest whole value.

TABLE 9

ILLUSTRATION OF COMPUTATION OF A PERCENTILE RANK AND OF A PERCENTILE FROM A GROUPED FREQUENCY DISTRIBUTION

M	*f*	*cf*
160–169	3	80
150–159	5	77
140–149	2	72
130–139	8	70
120–129	8	62
110–119	5	54
100–109	3	49
90– 99	11	46
80– 89	8	35
70– 79	8	27
60– 69	8	19
50– 59	5	11
40– 49	2	6
30– 39	1	4
20– 29	3	3

Computation of the Percentile Rank of the Score 95:

(1) 95 is $95. - 89.5 = 5.5$ units from the lower real limit of the interval 90–99.

(2) 5.5 is $\frac{5.5}{10}$ of the size of the interval.

(3) $\frac{5.5}{10} \times 11 = 6.05$ scores lie between 95 and 89.5.

(4) 35 scores lie below 89.5; hence, $35 + 6.05 = 41.05$ scores lie below 95.

(5) $\frac{41.05}{80} \times 100 = 51\%$ of the measures lie below 95.

Computation of the 50th Percentile:

(1) 50% of N = 40 scores lie below the 50th percentile.

(2) 35 of these scores lie below 89.5; $40 - 35 = 5$ of these scores lie between 89.5 and the 50th percentile.

(3) $\frac{5}{11}$ of the interval 90–99 is below the 50th percentile.

(4) $\frac{5}{11} \times 10 = 4.54$ score units = distance from 89.5 to 50th percentile.

(5) 50th percentile $= 89.5 + 4.54 = 94.04$.

Computation of a Given Percentile

The procedures to be followed in determining a given percentile is suggested by the preceding discussion. If we had a list of the

original scores arranged in order of size and wished to determine, for example, the 50th percentile, we would first determine how many scores constituted 50 per cent of the total number in the series, and would then count up from the bottom of the list until we had reached this number. If there were an even number of scores, the score halfway between the score last counted and the next above would then be the 50th percentile. If the number of scores were an odd number, the score corresponding to the 50th percentile would be the middle score in the series. Again, however, if we are working directly from a grouped frequency distribution, we must follow a more complicated procedure. Suppose, for example, that we wish to determine the 50th percentile in the distribution given in Table 9. To do this, we must first determine how many scores constitute 50 per cent of the total number of scores in the distribution. Fifty per cent of 80 is 40.

We wish to determine, then, below what point along the scale 40 scores will lie. By examining the cumulative frequency column, we note that this point must lie in the interval 90–99, since 35 of the scores lie below this interval and 46 below the one above. This means that we must find the point *within* the interval 90–99 below which $40 - 35 = 5$ of the frequencies in that interval lie. Five frequencies represent $\frac{5}{11}$ of the total number of scores within the interval. Since the interval contains 10 units, the point desired is $\frac{5}{11} \times 10 = 4.54$ score units above the lower real limit (89.5) of the interval. Thus the 50th percentile is $89.5 + 4.54$ or 94.04.

Expressed in more general terms, the procedure for computing any given percentile (that is, the point below which a given per cent of the measures lie), is as follows:

1. Find the given per cent of N.
2. Subtract from this number the number in the cumulative frequency column which is next below it.
3. The desired percentile will lie in the interval corresponding to the cumulative frequency which just *exceeds* the result of

Step 1. Divide the difference obtained in Step 2 by the frequency in this interval.

4. Multiply the quotient by the size of the interval.
5. Add this product to the lower limit of the interval. The result is the desired percentile.

Note: The procedures that have just been described for computing percentiles and percentile ranks are by no means the most convenient that can be followed, particularly if a large number of percentiles or percentile ranks are to be computed. The preceding methods have been presented primarily in order to acquaint the student with the essential nature of the percentile. A more convenient graphic method of transposition will be presented in the following chapter.

The Uses and Interpretation of Percentiles

The preceding discussions have been concerned primarily only with the definition of percentiles and with their computation. The more important questions of "In what situations and for what specific purposes may percentiles be employed?" and "How may percentiles be interpreted?" have yet to be considered. Described in general terms, the major uses of percentiles in education and psychology are:

1. To facilitate the interpretation of a single measure in a distribution of such measures;
2. To make possible comparisons between and combinations of measures originally expressed in different units — particularly to permit comparisons and combinations of scores on different tests (for individuals in the same group or in groups of comparable ability); and
3. To provide a condensed description of a frequency distribution — particularly to describe its variability and form.

A number of illustrations of the first two of these uses are suggested in the manual in the study exercises corresponding to this chapter. These questions will also draw attention to some of the more important *limitations* of percentiles in practical work. It

will be left to the student to develop these illustrations and to discover these limitations for himself. It is believed that this procedure will result in the acquisition of a more thorough understanding than if the answers to these questions were provided in the textual discussion. An attempt has been made, however, to make these questions sufficiently leading so that most students should have little difficulty in supplying the answers required. The questions in the manual, then, and the answers to them which are supplied by the student, should be considered as an integral and essential part of this whole discussion of percentiles.

CHAPTER IV

GRAPHICAL REPRESENTATION OF FREQUENCY DISTRIBUTIONS

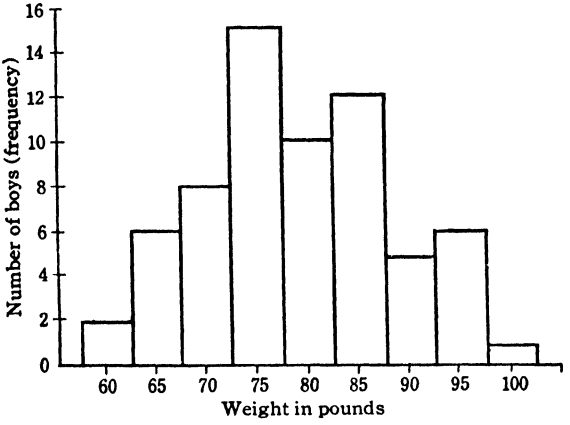
IN ORDER to describe or interpret a given frequency distribution, we may often wish to have answers to questions such as the following: Which measures occur most frequently? How are the measures distributed? Are they evenly distributed over the whole range, or do they tend to concentrate or pile up at certain points more than at others? How much do they tend to pile up at these points? What is the general form of the distribution — for example, is it symmetrical in shape?

These and similar questions can, of course, be answered through a detailed examination and comparison of the individual class frequencies. Most of these characteristics of the frequency distribution, however, can be readily determined at a glance if the distribution is portrayed in graphic form. Graphical representations can be much more easily read than statistical tables, and are particularly desirable if the data are to be presented in a report intended for readers untrained in the use of statistical methods. Such representations, furthermore, are essential even to the trained statistician in any study concerned primarily with the shape of the distribution.

The Histogram

The simplest form of graphical representation of a frequency distribution is the histogram. This type of representation is illustrated in Figure 1. The histogram in Figure 1 is based on the frequency distribution given beside it. From this figure we may note several general characteristics of the histogram. The vertical and horizontal lines at the left and at the bottom of the figure are known as the *axes*. The scale along the vertical axis is that along which the frequencies in the individual intervals, or

the class frequencies, are represented, and is referred to as the frequency scale. The horizontal scale is that along which the scores or measures are represented. The horizontal scale is divided into a number of equal units, each of which usually corresponds to one of the intervals in the distribution. The numbers given below the



FREQUENCY DISTRIBUTION

Weights	f
98–102	1
93–97	6
88–92	5
83–87	12
78–82	10
73–77	15
68–72	8
63–67	6
58–62	2

FIG. 1.
Frequency distribution and histogram of weights of a group of 65 boys (closed rectangle type of histogram).

horizontal scale sometimes represent the midpoints of the intervals and sometimes the limits of the intervals. In Figure 1 the numbers below the horizontal scale correspond to the *class measures* or interval midpoints. The base of each of the rectangles or columns of the histogram corresponds to one of the intervals in the distribution. The height of each column is proportional to the frequency in the corresponding interval in the distribution. Sometimes the lines between the adjacent rectangles are omitted.

The manner in which the histogram is constructed is too obvious to warrant any very detailed explanation. The scale along the vertical axis is laid off so as to provide for the largest class frequency in the distribution and so as to result in the desired proportions (height to width) in the completed histogram. The vertical scale *always begins with zero* at the intersection of the two axes. The horizontal scale is divided into a number of equal

intervals. The number of these intervals usually is two or three more than the number of intervals in the frequency distribution to be portrayed, so that a space may be left between the histogram and the vertical axis and between the histogram and the right-hand margin of the total space used. The use of arithmetically ruled paper will make it much easier to lay off these scales and to draw the rectangles.

The Frequency Polygon

Another type of graphical representation quite commonly employed is the frequency polygon. The frequency polygon in Figure 2 is based on the same distribution as the histogram in Figure 1. Figure 3 presents both these figures on the same chart.

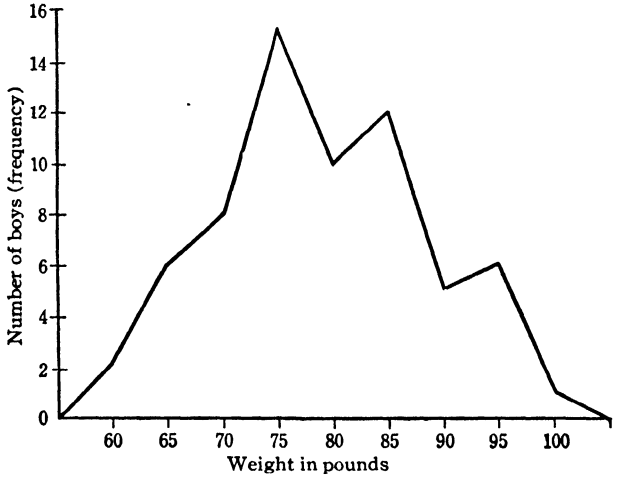


FIG. 2.
Frequency polygon of distribution of weights of a group of 65 boys.

The frequency polygon may be considered as having been derived from the histogram by drawing straight lines joining the midpoints of the upper bases of adjacent rectangles (or columns). The polygon is closed at each end by drawing a line to the base line from the midpoint of the upper base of each of the end columns to the midpoint (on the base line) of the next outlying interval (of zero frequency). It is, of course, not necessary to construct

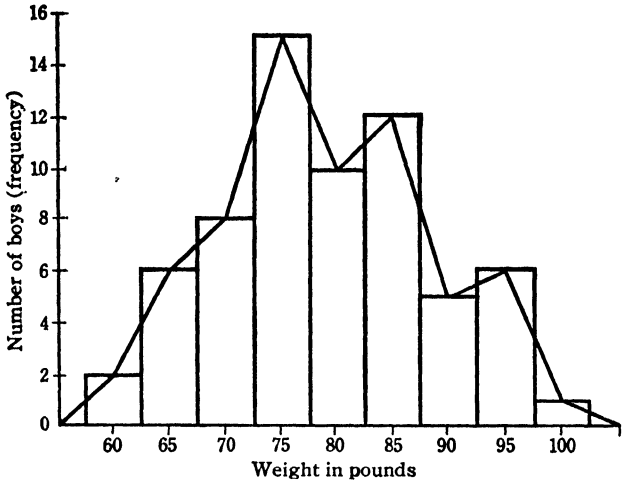


FIG. 3.
Frequency polygon and histogram (superimposed) of distribution of weights of a group of 65 boys.

the histogram first in order to construct the polygon. The polygon may be constructed directly by marking points directly *above the midpoint* of each interval at a distance from the base line proportional to the frequency in the interval. These points are then joined by straight lines, and the polygon is closed as before.

The Cumulative Frequency Curve or the Ogive

Another method of representing distributions graphically — much less frequently used than the histogram or polygon but superior to them for certain purposes — is the cumulative frequency curve or *ogive*, sometimes known as the *percentile curve*. It is constructed in very much the same fashion as the polygon except that the *cumulative* frequency is plotted for each interval rather than the frequency within the interval, and that the points joined by the straight lines are directly above the *upper limit* of each interval instead of above its midpoint. Figure 4 presents a cumulative frequency curve based on the same data as Figures 1 and 2. In order to construct this curve, a cumulative frequency column was first prepared for the distribution in the manner explained on page 34. The distribution and the cumulative frequency column

are given in the table beside the ogive. In the construction of this ogive, the axes were prepared in the same fashion as for a polygon, except that the vertical scale was laid off so as to include the highest *cumulative* frequency. With reference to these axes, then, a point was located which was directly above the upper limit of the first interval (58–62) and 2 units from the base line along the vertical scale. A second point was located which was directly above the upper limit of the second interval and 8 units from the base line. Other points were similarly located for each of the remaining intervals, and these points were joined by straight lines. The curve was then closed at the bottom in the same fashion as in the construction of a polygon. If now the vertical scale is divided into 10 or 100 equal parts, as has been done at the right of Figure 4, decile or percentile values corresponding to any given

FREQUENCY DISTRIBUTION

(Cumulative)

Weights	f	cf
98–102	1	65
93–97	6	64
88–92	5	58
83–87	12	53
78–82	10	41
73–77	15	31
68–72	8	16
63–67	6	8
58–62	2	2

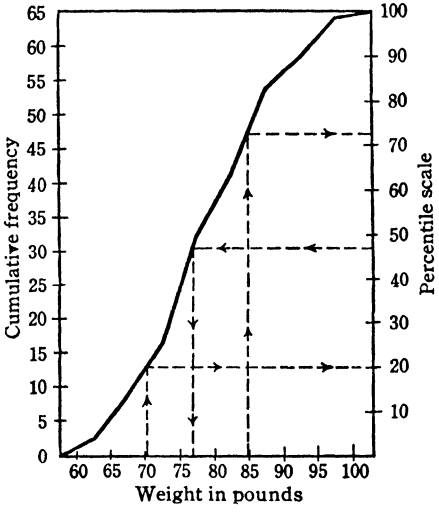


FIG. 4.

Cumulative frequency curve or ogive of weights of a group of 65 boys.

weight can be read directly from it. For example, if we erect a perpendicular from the base line at the point 85 until it meets the ogive, and then draw a horizontal line from this point until it meets the scale at the right, we find that 73 per cent of the measures in the distribution are below 85 pounds. Similarly, 20 per

cent of the cases lie below the point 71 on the weight scale; in other words, the weight measure 71 is at the 20th percentile. If arithmetically ruled paper had been used, we would not have needed actually to draw these vertical and horizontal lines, but could have followed along the ruled lines on the chart to reach the points desired, or could have used a ruler as a guide to determine them more conveniently. To determine the percentile rank of a weight of 83 pounds, for example, we could lay a ruler on the chart in a vertical position such that its edge fell at the point 83 on the base line, and would then mark the point on the ogive at which the same edge cut it. We would then hold the ruler in a horizontal position such that its edge coincided with the point just determined, and would read the desired percentile rank from the right-hand scale at the point at which it was cut by the ruler's edge.

The procedure just described can be reversed in order to find any given percentile. For example, if we wished to find the 47th percentile, we would lay our ruler horizontally across the chart so that its upper edge corresponded to the point 47 on the percentile scale, then mark the point at which that edge of the ruler cut the ogive, and then adjust the ruler in a vertical position so that its edge passed through the point just determined. The desired weight would then be that at which the edge of the ruler cut the base line — in this case, about 77 pounds.

This graphic method of transforming scores into percentile ranks or percentile ranks into scores is much more convenient to apply (if a large number of scores or percentile ranks are to be transformed) than the computational procedure explained on pages 33 to 35. This method may not be quite so accurate, because of possible errors made in plotting the ogive or in reading values from it, but considering the inherent unreliability of the percentile, it is sufficiently accurate for all practical purposes.

Sometimes, when the absolute values of the cumulative frequencies are of no interest in themselves, it is more convenient to plot the percentile rank of the upper limit of each interval directly,

instead of plotting the cumulative frequencies. To do this, we first express each cumulative frequency as a *per cent* of the total frequency; that is, we prepare a column of *relative* cumulative frequencies. (Such a column has been prepared for each of the distributions in Table 2 on page 22 of the manual. The quickest way to compute these numbers, particularly if a computing machine of the multiplying type is available, is first to compute $\frac{100}{N}$, and then to multiply each cumulative frequency by this number. For the distribution of ninth grade scores in Table 2, $\frac{100}{N} = \frac{100}{3845} = .0260$. Each cumulative frequency was multiplied by this number to obtain the numbers in the column headed "Cumulative frequency in per cents.") Each of these relative cumulative frequencies, of course, represents the percentile rank of the upper real limit of the corresponding interval. We can then lay off the vertical percentile scale directly, using the ruled lines on our coordinate paper, instead of later subdividing a scale into 10 or 100 equal parts.

There is no very real distinction between a cumulative frequency curve and a percentile curve. A distinction sometimes made, however, is that the curve is called a cumulative frequency curve if the vertical scale shows only cumulative frequencies, and is called a percentile curve if the vertical scale shows only the percentiles. Figure 4 is then both a cumulative frequency curve and a percentile curve, since both types of scales are provided. The term *ogive* refers to the shape of the curve, and may be applied either to the cumulative frequency or the percentile curve.

Ogives are sometimes drawn with the percentile or cumulative frequency scale on the horizontal axis and the score scale on the vertical axis.

Supplementary Suggestions for the Construction of Histograms, Polygons, and Ogives

1. Note that the sides of the rectangles in the histogram and the turning points in the ogive always come above the *real limits* of

the intervals. Since most of the integral measures will be considered as having been taken to the nearest whole value, these real limits will ordinarily lie .5 of a unit beyond the integral limits of each interval. This fact must be taken into consideration in indicating numerical values along the base line and in plotting the figure.

2. Any of these figures should always carry a complete, clear, and concise *title*. This title should always completely identify the data represented, independently of any accompanying textual description. In other words, the title should be such that if the chart is removed from context — for example, for the purpose of preparing a lantern slide — it will contain all the information needed for its interpretation.

3. The vertical and horizontal scales should always be definitely *labeled* so that it is perfectly clear what each scale represents and what units are employed on each.

4. If more than one figure is drawn on the same chart, each should be drawn with a different kind of line (solid, broken, dotted, etc.), and the meaning of each line indicated by a neat *legend* in an upper corner of the chart or in some other convenient space on the chart. In general, if there is any possibility that the chart may be later reproduced in printing, do not use colored inks for distinguishing between superimposed figures, because of the expense involved in color reproduction in printing.

"Smoothing" Frequency Polygons and Ogives

It will be noted that because of the erratic manner in which frequencies change from one interval to the next in many distributions, the straight lines joining the points determined in plotting the polygon or ogive will form a very irregular line. Irregularities in the form of the figure will be much more prominent in the polygon than in the ogive. Many of these irregularities may be considered as only accidental or of little or no significance, since they may be peculiar to the particular grouping (or choice of interval) employed in the construction of the frequency dis-

tribution, or may be characteristic only of the one sample of individuals considered and not generally characteristic of other similar groups. In order to obtain a more highly generalized picture, therefore, the practice of "smoothing" the original figure is sometimes followed. This may be done by drawing free-hand a smooth curved line which comes as close as possible to passing through all of the points used in plotting the original figure or, in other words, which most nearly coincides with the irregular straight line outline. Such a line has been drawn free-hand for the polygon and ogive in Figures 5 and 6 respectively.

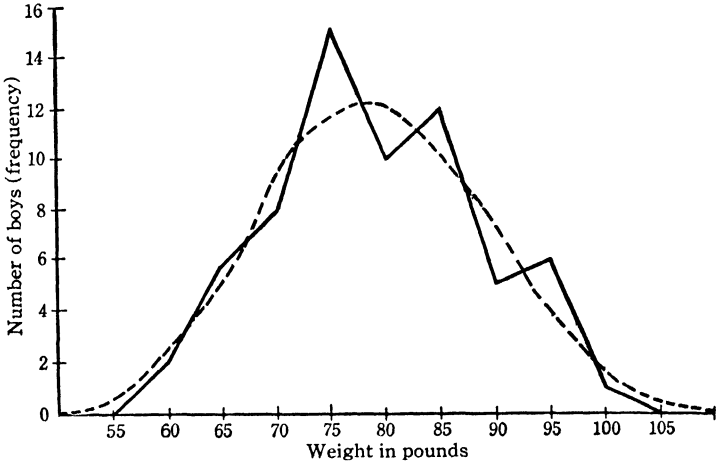


FIG. 5.
Smoothed frequency curve of weights of a group of 65 boys.

Smoothing should be resorted to only when the group of individuals involved is not being studied for its own sake but is only being considered as a *sample* which is presumably representative of some still larger group or population. The purpose of smoothing, then, would be to remove from the polygon or ogive for the sample those irregularities which would not be characteristic of the distribution for the entire population. The principal danger in this smoothing procedure is that it sometimes removes irregularities which are not accidental, but which are real and sometimes significant characteristics of the distribution for the whole popula-

tion. There is, of course, no way of telling by inspection whether or not a given irregularity is accidental.

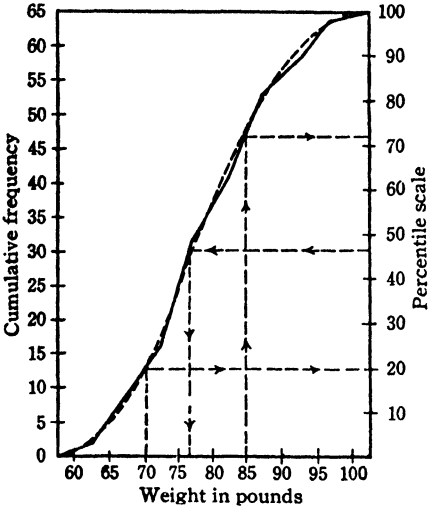


FIG. 6.
Smoothed ogive of distribution of weights of a group of 65 boys.

There are other and more objective ways of smoothing figures than the free-hand method just described. In general they are not sufficiently better than the free-hand method to warrant their consideration here. Any smoothed figure, no matter how derived, represents at best only a guess as to how the more highly generalized figure would look, and no method of smoothing is highly reliable for this purpose. The only highly dependable method of eliminating these accidental irregularities is to collect data from larger numbers of cases, that is, to plot the results for larger samples.

The Form of a Frequency Distribution

There are a number of terms used to describe the form of a frequency distribution with which the student should become familiar in order that he may more readily comprehend the subsequent discussions.

A distribution is said to be *bilaterally symmetrical* if the polygon or frequency curve can be folded along a vertical line so that the two halves of the figure coincide. C, D, E, F, G, and H in Figure 7 (on next page) are illustrations of symmetrical curves.

A distribution is said to be *skewed* if it is lacking in symmetry, that is, if the measures tend to pile up at one end or the other of the range of measures. A distribution is said to be negatively skewed or skewed to the left if the measures pile up at the upper

end of the scale, and positively skewed or skewed to the right if the measures pile up at the lower end of the scale. Curve B in Figure 7 is very markedly skewed to the right, while curve A is moderately skewed to the left.

A curve is said to be *bell-shaped* if, as its name implies, it is symmetrical, has one broad smooth hump in the middle, and "tails off" gradually at either end. Curves C, D and E in Figure 7 are bell-shaped, but exhibit various degrees of flatness or peakedness.

The normal curve is a peculiar bell-shaped curve which can be exactly defined only in terms of the equation used to plot it. This type of curve will be discussed in Chapter VII. Curve F

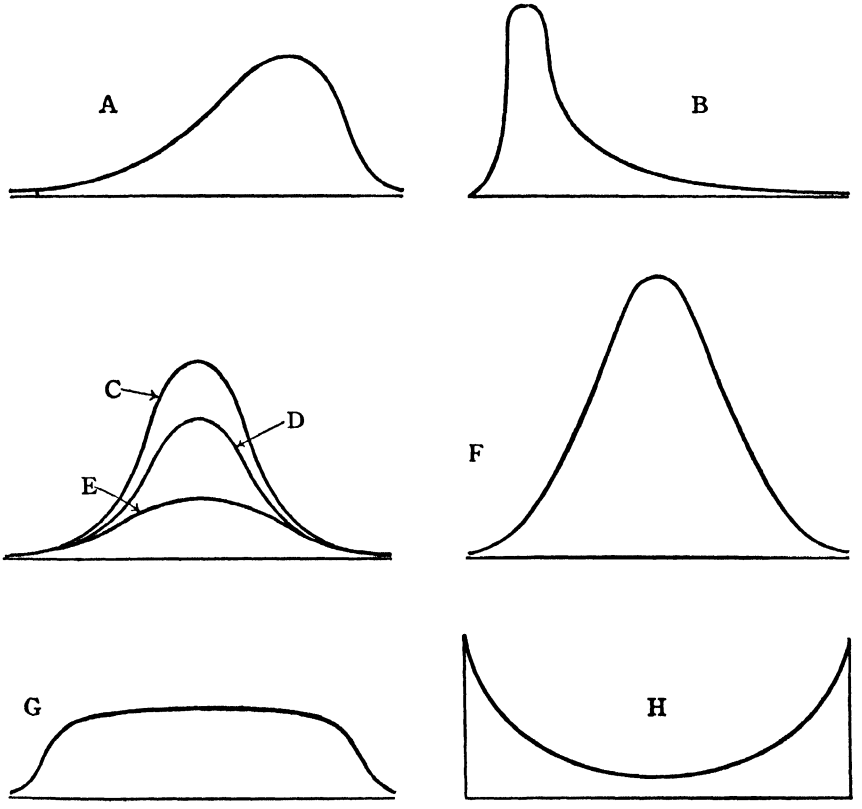


FIG. 7.
Typical forms of frequency distributions.

in Figure 7 is a normal curve, as is also curve D. The *apparent* difference in peakedness between D and F is due only to the difference in the choice of units used in plotting the same curve.

A frequency distribution is said to be *rectangular* to the degree that all class frequencies have the same value. Curve G in Figure 7 approaches rectangularity. Curve H is representative of "U-shaped" curves. Curves A–G inclusive are said to be "unimodal," since they have only one pronounced peak. (See discussion of the *Mode* in Chapter V.) A curve with two pronounced peaks, even though both are not of the same height, would be described as "bi-modal."

The Uses and Interpretation of Histograms, Polygons and Ogives

As in the preceding chapter, the student will be left to discover for himself, with the aid of the questions in the study manual, the various uses and limitations of the histogram, polygon and ogive. Again, these study exercises and the answers supplied by the student must be considered as an integral part of this chapter. It is essential that each of these questions be very carefully considered.

CHAPTER V

MEASURES OF CENTRAL TENDENCY

THE term "average" — already familiar to the student before beginning this course — is one whose popular meanings are extremely loose and ambiguous. We use the same term indiscriminately in speaking of, for example, the "average American," the "average personality," the "average yield of corn per acre," the "average high school," the "average length of life," etc. Synonyms for the term in its popular usages are such expressions as "typical," "usual," "representative," "normal," and "expected." If asked to define the term more accurately, the "average" man might respond that it is the single measure or individual that best represents a group of measures or of individuals, but if asked how to *select* the most representative measure in a group of measures he is likely to become less specific. He may say that in order to find the average of a series of figures you simply add them all up and divide by the number of them, but such a concept becomes meaningless when applied to data that cannot be numerically represented, as in the case of "the average personality" or "the average school teacher," and even for data which may be numerically represented this process does not in all cases yield the most "typical" or "representative" result.

Whatever may be the specific meanings of the term "average," it is reasonably clear to anyone, from a knowledge of the general meaning of the term, that the use of an "average" adds greatly to the convenience with which we can reason about groups or make comparisons between groups. No person can bear in mind simultaneously the characteristics of *all* members of a large group of individuals, but he has little difficulty in handling such groups in his thinking when he can let a single quantitative measure represent the whole group, that is, when he can use an "average"

as a concise and simple picture of the large group from which it is derived. Nothing further need be said about the general utility of averages, although the distinction between different kinds of "averages" will require the careful attention of the student.

To the statistician, "average" is a general term applying to all kinds of measures of central tendency derived from group data. There are at least five such measures in common use, but only three of them — the arithmetic mean, the median, and the mode[1] — are used with sufficient frequency in practical applications in education and psychology to warrant their inclusion in a first course in statistics for students in those fields.

THE ARITHMETIC MEAN

The *arithmetic mean* of a series of measures is equal to the sum of the measures divided by their number. It is the "average" most often referred to in popular usage. Using the algebraic notation in which

M represents *arithmetic mean*,
Σ means "the sum of,"[2]
X represents an individual score or measure,
N represents the number of measures,

it may be defined by the formula,

$$M = \frac{\Sigma X}{N}$$

The arithmetic mean is usually referred to simply as the *mean*. The letter M is most frequently used to represent it, but the notation $A.M.$ is also often used. While the mean may, of course, be computed directly from the original measures, its computation is made more convenient when the measures have first been arranged into a frequency distribution. Consider the following frequency distribution of test scores:

[1] The other two are the *harmonic mean* and the *geometric mean*. Descriptions of these averages may be found in any good reference book in statistics.

[2] Σ is the upper case Greek letter "Sigma."

TABLE 10

COMPUTATION OF THE MEAN BY THE "LONG METHOD"

(Applied to a Distribution with a Unit Interval)

Score (X)	Frequency (f)	($f \times X$)
19	1	19
18	1	18
17	3	51
16	8	128
15	17	255
14	13	182
13	10	130
12	4	48
11	1	11
10	2	20
	$N = 60$	$862 = \Sigma f \times X$

$$M = \frac{\Sigma f \times X}{N} = \frac{862}{60} = 14.37$$

The mean could have been computed by the method indicated in the original definition by simply summing all the individual scores and dividing by N. The score 17 would then have entered into the addition column 3 times, the score 16 would have entered 8 times, etc. The process is simplified by adding the products $3 \times 17 = 51$, $8 \times 16 = 128$, etc., and exactly the same result is secured. To facilitate the computation, a third column (headed $f \times X$) is added to the frequency distribution, in which is written the product of each score and the frequency with which it occurred, and these products are added to secure the sum of all the measures. In the case of the illustration, therefore, it was necessary to add only 10 numbers to obtain the sum of all the measures, instead of adding 60 separate measures. The notation used in the problem is self-explanatory.

The mean of a grouped frequency distribution can be computed in a similar fashion by letting the midpoint of each interval represent all the scores in the interval, but in this case the accuracy of the result is affected by the loss of identity of the original measures. Consider the following grouped frequency distribution.

The uppermost interval contains one score. That score may have had an original value anywhere within the limits 93–97 in-

TABLE 11

COMPUTATION OF THE MEAN OF A GROUPED FREQUENCY DISTRIBUTION BY THE "LONG" METHOD

Interval	Midpoint of Interval (X)	(f)	($f \times X$)
93–97	95	1	95
88–92	90	3	270
83–87	85	2	170
78–82	80	7	560
73–77	75	12	900
68–72	70	10	700
63–67	65	9	585
58–62	60	5	300
53–57	55	3	165
48–52	50	2	100
		$N = 54$	$3845 = \Sigma f \times X$

$$\text{Mean} = \frac{\Sigma f \times X}{N} = \frac{3845}{54} = 71.20$$

clusive, but nothing more concerning its original identity can be determined from the table itself. We therefore assume that the best guess of its original value is the midpoint of the interval (95), and we then use this value in the subsequent computations. In the same way, for the interval whose midpoint is 90, we assume that the mean value of the three original scores is 90, or that their sum is 270, and so on for the rest of the intervals.

Now in the case of this specific illustration, it happens that the *actual* value of the single score tabulated in the uppermost interval was 94. Hence, the number (95) entered in the third column was 1 too large, and an error (due to grouping) was therefore introduced into the computation.

In the interval whose midpoint is 90, the *actual* values of the three scores were 92, 89, and 91. The *actual* value of their sum was therefore 272, and the number entered in the third column (270) was 2 too small. In a similar fashion, an error due to grouping will be present in the number written in the third column for most of the intervals, the only exceptions being those rare intervals where by chance the mean value of the scores in the interval is exactly equal to the midpoint. When all intervals are considered together, however, the errors in one direction are just about

balanced by those in another, so that the final value obtained for the mean is usually a very close approximation to the actual value, that is, the value of the mean that would have resulted by summing all of the *original* scores and dividing by their number.

It is left as an exercise for the student to show more specifically why the errors in the mean that are due to grouping tend to cancel out to zero when all intervals are considered.

The "Short" Method of Computing the Mean

Consider the following numbers:

217,011
217,009
217,006
217,012
217,005

What is the quickest way of finding the mean of these numbers? We note at once that each number is equal to 217,000 plus a small number. We can compute the mean of the original numbers by simply finding the mean of these small numbers and adding this value to 217,000, as in the following illustration.

Original Numbers	*Difference between Number and 217,000*	
217,011	11	
217,009	9	
217,006	6	
217,012	12	
217,005	5	
	43	(Sum of differences)

$\frac{43}{5} = 8.6$ (Mean of small numbers or differences)

$217{,}000 + 8.6 = 217{,}008.6$ (Mean of large numbers)

Compare the operations in the preceding illustration with those in

the following application of the long, or direct, method to the same numbers.

$$\begin{array}{r} 217{,}011 \\ 217{,}009 \\ 217{,}006 \\ 217{,}012 \\ 217{,}005 \\ \hline 1{,}085{,}043 \end{array} \quad \text{(Sum of numbers)}$$

$$\frac{1{,}085{,}043}{5} = 217{,}008.6 \text{ (Mean)}$$

The advantages of the first or "short" method are in this case quite obvious. By eliminating the necessity of dealing with *any* large numbers in arithmetic computation (except for the last step) we not only reach the final result more easily and quickly, but also with less likelihood of making arithmetical errors.

The process involved in the preceding illustration represented a specific application of the following generalized rules for computing the mean of a series of numbers by the so-called "short" method. The language used in these rules differs from that used in the preceding illustration, but the student should have no difficulty in recognizing that the process is essentially the same.

SHORT METHOD OF COMPUTING THE MEAN

1. Select any convenient value as an "arbitrary reference point." (It is usually best to select a value likely to be close to the actual mean.)
2. Express each measure as a "deviation" from this arbitrary reference point. (Each "deviation" is equal to the difference between the measure and the arbitrary reference point. If a measure is below the arbitrary reference point, its deviation will have a *negative* sign.)
3. Find the mean of these deviations by the usual method. (Add algebraically, and divide the sum by the number of measures.) Call this mean of the deviations the "correction" to the arbitrary reference point. (If the sum of the negative

deviations exceeds the sum of the positive deviations, this correction will be negative.)

4. Add (algebraically) this correction to the arbitrary reference point. The result is the mean of the original measures.

The student may demonstrate for himself that this process will always yield the same result as that obtained by the usual method of adding the original measures and dividing by their number, regardless of the value chosen as the arbitrary reference point.

The Short Method Applied to the Frequency Distribution

The occasion will very rarely if ever arise in which the student will compute a mean in exactly the manner illustrated in the preceding discussion. The purpose of the preceding discussion was simply to explain the fundamental nature of the short method, in order that the student might better understand its more practical application to data arranged in frequency distributions.

Let us first consider the application of the short method to a frequency distribution with a unit interval.

TABLE 12

SHORT METHOD OF COMPUTING MEAN APPLIED TO A FREQUENCY DISTRIBUTION WITH A UNIT INTERVAL

Score	f	d	fd	
19	1	5	5	
18	1	4	4	
17	3	3	9	$\Sigma fd = 51 - 29 = 22$
16	8	2	16	$\frac{\Sigma fd}{N} = \frac{22}{60} = .37$
15	17	1	17	
A.R. = 14	13	0	/+ 51	
13	10	− 1	− 10	
12	4	− 2	− 8	
11	1	− 3	− 3	
10	2	− 4	− 8	
$N =$	60		/− 29	

$$\text{Mean} = \text{A.R.} + \frac{\Sigma fd}{N} = 14 + 0.37 = 14.37$$

To facilitate computation, the deviation of each score value from the A.R. (abbreviation for arbitrary reference point) is written in a third column called the "deviation" or d column.

The product of each frequency and the corresponding *d* value is then written in the *fd* (frequency × deviation) column. The purpose of this step is clear. The score 17 occurs 3 times. The deviation of a single score of 17 from 14 is 3; the sum of the deviations of 3 such scores is 3 × 3 or 9. The sum of the deviations of all scores is then obtained by adding the numbers in the *fd* column. This sum is most conveniently secured by adding the positive and negative deviations separately, and then obtaining the algebraic sum of these partial sums.

The application of the short method to the grouped frequency distribution is essentially the same as in the foregoing illustration, the only difference being the way in which the deviations are expressed. Consider the following illustration.

TABLE 13

SHORT METHOD OF COMPUTING MEAN APPLIED TO A GROUPED FREQUENCY DISTRIBUTION

Midpoint of Interval	*f*	*d*	*fd*	
95	1	4	4	$\Sigma fd = 24 - 65 = -41$
90	3	3	9	
85	2	2	4	$\frac{\Sigma fd}{N} = \frac{-41}{54} = -.759$
80	7	1	7	
A.R. 75	12	0	+24	or −.76 (rounded)
70	10	−1	−10	(correction in *interval* units)
65	9	−2	−18	$5 \times -.76 = -3.80$
60	5	−3	−15	(correction in score units)
55	3	−4	−12	
50	2	−5	−10	Mean = A.R. + correction
	$N = 54$		−65	= 75 − 3.80 = 71.20

In this case each deviation is expressed in units of *intervals*, rather than in score units. The midpoint 80, for example, deviates 1 interval from the A.R., the midpoint 95 deviates 4 intervals, etc. The sum of the deviations divided by N, therefore, tells *how many intervals* the A.R. deviates from the mean. In other words, $\frac{\Sigma fd}{N}$ gives the correction to the A.R. in *interval*, rather than in score units. Since in this case the interval is 5 times as large as a score unit, one must multiply the correction by 5, transforming

it into score units, before applying it to the A.R. in the final step.

The process of computing the mean of a grouped frequency distribution may now be recapitulated in the following general rules:

STEPS IN THE COMPUTATION OF THE MEAN OF A GROUPED FREQUENCY DISTRIBUTION BY THE SHORT METHOD

1. Select as the arbitrary reference point (A.R.) the midpoint of the interval which you think is most likely to contain the actual mean.[1] (The midpoint of any other interval will do, but the *fd* products will in general be smaller and therefore more convenient to handle if the A.R. is selected as suggested here.)
2. Indicate, in a column headed "*d*," the number of intervals between each interval-midpoint and the A.R. (Simply count away from the A.R. one unit at a time in either direction. All deviations below the A.R. must be preceded by a negative sign.)
3. Multiply the frequency in each interval by the corresponding *d* value, and write the products in a column headed "*fd*." (All *fd* products below the A.R. will have a negative sign.)
4. Find the sum of the positive products in the *fd* column, then the sum of the negative products. Then add these sums *algebraically*.
5. Divide this result by N, the total number of cases. (This quotient may be denoted by c', and represents the "correction" in *interval* units to the A.R.)
6. Multiply the quotient by the size of the interval. (This product, which may be denoted by c, represents the "correction" to the A.R. in *score* units.)

[1] Because of this manner of selecting it, the arbitrary reference point is sometimes called the "Guessed Mean" or the "Assumed Mean" and is often denoted by the abbreviation "G.M." This notation has sometimes misled beginning students because it seems to imply that the short method will not be accurate if a "good" guess is not made of the value of the actual mean.

7. Add this product algebraically to the A.R. (Subtract if negative, add if positive.)

The reasons for calling this method the "short" method may not be too apparent. Since the number of intervals in a grouped frequency distribution seldom exceeds 20, and since the A.R. is usually taken near the middle of the distribution, it follows that the numbers in the deviation column are usually only one-digit numbers. All multiplications required in filling the *fd* column may, therefore, be done mentally. For this reason, and because all long column addition is eliminated, the computation of the mean by this method is extremely simple arithmetically, and also provides fewer opportunities for error than does the long method.

It is, nevertheless, true that, as far as the computation of the mean is concerned, this method is "short" in name only. When the mean of a series of measures is the only measure desired (when no measures of variability are to be computed later) and when an adding or calculating machine is available, time is saved by simply adding the original measures and dividing by their number, without taking time to construct the frequency distribution and to fill in the *d* and *fd* columns called for by the short method. Usually, however, it is desirable to construct the frequency distribution for other reasons than for calculating the mean. Usually, also, some measure of variability, such as the standard deviation, is required in addition to the mean. As will be explained later, the short method is highly essential in the computation of the standard deviation. From the point of view of the time consumed in the entire process of constructing a frequency distribution and computing the mean and the standard deviation, the "short" method undoubtedly is a significant time-saver.

THE MEDIAN

The median may be most simply defined as the *middle* measure in a series in which all measures have been arranged in the order of their size. Since the median is usually computed from a fre-

quency distribution, the best definition in general is that the median is that point on the scale above and below which half of the scores or frequencies lie. The median is thus the same as the 50th percentile. The method of computing the 50th percentile or median has been explained (see page 36) and need not be repeated here. The usual abbreviation for the median is *Mdn*.

THE MODE

The *mode* of a frequency *curve* may be defined as that value along the horizontal scale at which the height of the curve is greatest. It is sometimes also defined as the most frequently recurring score in the distribution. For example, in the distribution in Table 2, page 13 of the text, the modal score would be 112, since it occurs five times and no other occurs more than three times. The modal score is obviously a very unstable measure. In Table 2, for instance, if the individual scoring 145 had scored 146 instead, and if two of those scoring 112 had each scored 113, the mode would have been changed from 112 to 146. A more meaningful measure, usually called the "crude" mode, is the midpoint of the interval containing the highest frequency in a relatively coarsely grouped frequency distribution. For example, the crude mode of the distribution in Table 5, page 15, is 104.5. Even the crude mode is highly unstable for distributions of small numbers of cases.

When there is more than one outstanding frequency in a distribution (and these frequencies are not in adjacent intervals) we describe the distribution as multi-modal.

In all subsequent discussions and questions, the "mode" referred to may be taken to mean the "crude" mode as here defined.

THE NUMBER OF SIGNIFICANT DIGITS IN THE MEAN

It is reasonably apparent that the accuracy of the result of an arithmetic computation depends upon the accuracy of the original data. If each of the measures in a series represents only a rough estimate of or coarse approximation to an accurate measure of the same thing, then the mean of these approximate measures

must itself be considered as only an approximation. In this case, it would not be consistent with the nature of the original data to compute the mean to a large number of decimal places; on the contrary, to do so would give the computed mean the appearance of an accuracy which it does not possess. It is, therefore, important that the student know to how many decimal places a mean may be computed.

Suppose that we have a measure of the weight of each of seven similar objects, but that these weights have been determined with various degrees of accuracy. These measures (in pounds) are as follows: 12.34, 10.15, 9.2, 14., 7.363, 8., 10. The first measure has been taken to the nearest hundredth of a pound, and the real weight of the thing measured may therefore be anywhere between 12.335 and 12.345 pounds. The third measure has been taken to the nearest tenth of a pound, and the fourth only to the nearest pound. We know, then, that if the weight of the fourth object had been more accurately determined, the digits following the decimal point might have had any value.

We can then write these numbers in column order as follows:

$$\begin{array}{r} 12.34 \\ 10.15 \\ 9.2 \\ 14. \\ 7.363 \\ 8. \\ 10. \\ \hline 71.053 \end{array}$$

Most persons would write the *sum* of these numbers as 71.053, and would then compute the mean to be $\frac{71.053}{7} = 10.1504\ldots..$, the number of decimal places to which the result was carried depending only upon the whim of the computer.

The fallacy in this procedure becomes apparent if we substitute for each measure the highest *actual* weight that each object *might* have had. The first object, for example, *might* really have weighed almost 12.345 pounds, the fourth almost 14.5 pounds,

etc. The sum of these *maximum* weights is 72.6135, as shown below.

Maximum Values

12.345
10.155
9.25
14.5
7.3635
8.5
10.5
72.6135

$$M = \frac{72.6135}{7} = 10.3733$$

Minimum Values

12.335
10.145
9.15
13.5
7.3625
7.5
9.5
69.4925

$$M = \frac{69.4925}{7} = 9.9275\ldots$$

The *minimum* value of the actual sum, as shown above, is 69.4925. The actual mean, therefore, may lie anywhere between 10.3733 . . . and 9.9275 . . . The only digits in the mean first computed, then, that we know to be correct are the two digits to the left of the decimal point. This mean should, therefore, have been rounded to 10, in order to avoid giving a misleading impression of high accuracy.

In any number, the digits *known to be correct* are called the *significant* digits. The mean (10.1504), originally obtained in the preceding illustration, contains only two significant digits, as does the sum 71.053.

It is possible to set up a general rule for determining the number of significant digits in any sum. This rule is as follows: *The last significant digit in a sum cannot lie any farther to the right of the decimal point than the last significant digit in the least accurate of the measures added.* The least accurate of the weight measures just considered are the fourth, sixth, and seventh (14, 8, and 10),

which contain *no* significant digit to the right of the decimal point. Hence, the sum of these measures can contain no significant digit to the right of the decimal point.

This rule may be made clearer by the following illustration. Given the following measures, to find their mean: 11.17343, 10.2, 14.49. We can write these as follows:

$$\begin{array}{r} 11.17343 \\ 10.2???? \\ 14.49??? \\ \hline 35.7???? \end{array}$$

The question marks indicate that the digit in that place is unknown and may have any value from 0 to 9. The sum of the digits in the hundredths column is $7 + ? + 9 = ?$, since 16 plus an indeterminate number is still unknown. Hence, we can be sure of only the first three digits (35.7) in the sum.[1]

The mean of these measures is then $35.7???? \div 3 = 11.9????$

$$\begin{array}{r} 3\overline{)35.7????} \\ 11.9???? \end{array}$$

Similarly, if the sum of a series of 117 numbers is 246.532 (assuming that all of these digits are significant), their mean may be written 2.10711, as shown below.

$$\begin{array}{r} 2.10711 \\ 117\overline{)246.532????} \\ \underline{234} \\ 125 \\ \underline{117} \\ 832 \\ \underline{819} \\ 13? \\ \underline{117} \\ 1?? \\ \underline{117} \\ ??? \end{array}$$

[1] The last of the digits (7) is not absolutely correct — its actual value might be 8 or 9, depending upon the value of the fourth digit in the second number — but it is sufficiently accurate to be worth retaining, which is not true of the remaining digits in the sum.

We may be sure 117 is contained just once in 13?, regardless of the value of ?, and that it is also contained once (or very nearly once) in 1??, but we cannot carry the division farther.

The preceding examples illustrate the general rule that *the number of significant digits in the mean of any series is the same as the number of significant digits in its sum.*

We now have the two rules that will enable us to determine to how many decimal places any mean may be carried. Fortunately, the application of these rules is simplified in the case of integral test scores. Since such scores are never expressed in decimals, and since all of their digits are significant, the sum of any set of test scores always consists only of significant digits. Hence we can immediately establish the following simpler rule: *The number of significant digits in the mean of any distribution of test scores is equal to the number of digits in the sum of the scores (or measures).*

Thus, the mean of the distribution in Table 10 contains only *three* significant digits, since the sum of the scores (862) contains only three digits. This mean should, therefore, have been rounded to 14.4. Similarly, the mean of the distribution in Table 11 contains four significant digits (the sum is 3845) and hence cannot be meaningfully carried any farther than to 71.20.

When the mean of a distribution has been computed by the short method, we do not determine the sum of the original scores directly. However, we can readily determine *how many digits* there are in the sum by dropping the decimal places in the mean (as first computed) and multiplying by N. For example, in Table 12 the value of the mean as first computed is 14.37. Dropping the decimal places, we get 14, which when multiplied by 60 is 840. This is a close approximation to the sum of the original scores and tells us that that sum would contain three digits. Hence, this mean should be rounded to three digits, or to 14.4.

Similarly, in Table 13 the sum of the original scores would contain four digits ($71 \times 54 = 3834$), and hence the mean should contain only four digits (as does the mean given, 71.20).

The rule for determining the maximum number of digits

in which the mean of any distribution of integral measures may be expressed may then be stated as follows: *Drop the decimal places in the mean first computed, multiply by N, and round the mean to the number of digits in this product.*

It should be noted that the preceding rule applies, not only in the case of test scores, but also to any continuous data originally expressed in whole numbers.

These rules should be rigidly observed in all statistical work that the student may do with measures of continuous variables. They do not apply to *discrete* data. The mean of a distribution of sizes of families, for example, may, as far as the accuracy of the individual measures is concerned, be carried to any number of decimal places. In such cases, other considerations will determine the manner in which the result is expressed.

The injunction in the first sentence of the preceding paragraph does not mean that the student must always *retain* all significant digits in the mean, but only that he should retain none that are not significant. There will be many instances in which the character of the original data will permit a higher degree of accuracy in the mean than is actually needed for its interpretation. Other sources of error, in addition to the approximate character of the individual measures, also determine the accuracy or reliability of the mean. Certain of these other sources of error will be considered later. In general, then, the student should carry a mean only as far as is demanded by the uses to which it is to be put, even though these rules and other considerations will permit him to carry it farther.

It should be noted that the median and mode are not arithmetic in character — being counting or observational measures only — and hence are not subject to the rules here given. The student may safely follow the practice, however, of never carrying a median to any more decimal places than the mean of the same distribution.

The Importance of "Errors" in Statistical Work

This discussion of significant digits may at first appear to the student as pedantic and much ado about little. It may be observed that so much attention has been given to this issue, not only because of its intrinsic significance, but because it is one of the first concrete instances met in this course of the many sources of error which must be considered in the interpretation of statistical data. Another source of error — one which will be considered in the study exercises for this chapter — is the loss of identity of the individual measures which results when measures are grouped in a frequency distribution. Other more important types of error which will be considered later are errors in random sampling and errors due to lack of validity and reliability in the measuring instruments used. These other sources of error are far more serious than the one just discussed. Hence the rules here considered only indicate the *maximum* accuracy which a mean may have; the number of significant digits which it actually contains is nearly always less, and often much less, than these rules would indicate.

One of the worst mistakes that can be made in statistical work is that of uncritically accepting all statistical facts at their face value, or of presenting approximate or unreliable data without drawing attention to the errors which the data probably contain. Statistics as a body of knowledge and a system of techniques is in spirit exact and accurate. Precision and accuracy of statement are highly desirable for their own sake. There is enough of loose and careless thinking in education and psychology without statistics itself making any contributions of this sort. Among the most important elements in statistical judgment, then, are keen awareness of probable errors, and a disposition to qualify accordingly any conclusions based upon statistical analyses. Statistical "smoke screens" should never be permitted to hide or obscure the unreliable and ambiguous character of the original data with which we so often have to deal.

The Uses and Interpretation of the Measures of Central Tendency

As in previous instances, the student will be left to write for himself the most important section of this discussion of measures of central tendency. The foregoing descriptions and explanations should be adequate to enable him to identify the essential characteristics and the mathematical properties of each of the "averages" considered. With the aid of the study exercises, he should be able to appreciate readily how much it may matter which type of average is used in any given situation and for any given purpose — to discover that in many instances the choice of the wrong measure of central tendency may be as serious in its consequences as a deliberate falsification. Again, the fact that he has, in part at least, reasoned these things out for himself should result in their more permanent retention and more complete assimilation.

CHAPTER VI

MEASURES OF VARIABILITY

IT SHOULD be readily apparent that a measure of central tendency alone can describe only one of the important characteristics of a distribution, and that it is equally essential to know how *compactly* the measures are distributed about this point of central tendency, or, conversely, how far they are scattered away from it. In describing the distribution of intelligence for a given class of pupils, for example, it would not be sufficient to know only the average I.Q. of the class. For instructional purposes it is equally if not more important to know how large are the individual *differences* in intelligence within the class, or how heterogeneous the group is in intelligence. In other words, we should like to know whether the class is made up exclusively of students of average and near-average intelligence or contains a large proportion of very bright and very dull pupils.

This condition in a frequency distribution is variously referred to as *dispersion*, *spread*, *scatter*, *deviation*, and *variability*. There are several ways of describing this characteristic quantitatively. One of the simplest but least adequate of these methods is to state the values of the highest and lowest measures, or the range of the distribution. To describe the variability in intelligence in a given school class, for example, we might say that the highest I.Q. is 140 and the lowest is 73, or that the range of intelligence is 67 I.Q. points. This type of description is not very meaningful, since it is dependent only upon the two extreme individuals in the group, and since almost anything may be true of the form of the distribution between these extremes.

Another way of describing the variability of a distribution is to state the values of the 10th and 90th, or of the 25th and 75th, percentiles. For example, the knowledge that 10 per cent of the

pupils in a given group are below 80 pounds in weight while the upper 10 per cent are above 140 pounds in weight gives us a fairly accurate quantitative notion of the variability in weight of the individuals in the group. The *Semi-Interquartile Range* (Q), which is half of the distance between the upper and lower quartiles, or half of the difference between the 75th and 25th percentile scores, is one of the most frequently used measures of variability. These and other measures based upon percentiles, however, do not take into consideration the value of each individual score within the distribution, and are therefore unreliable and lacking in descriptive value. Two distributions may show the same semi-interquartile range, and yet the outlying scores in one distribution may be far more extreme than in the other.

The variability of the scores in a distribution clearly depends upon the amounts by which the individual scores *deviate* from the measure of central tendency. To describe the variability of a frequency distribution, therefore, we could determine the amount by which the score for each individual differs from the mean score, considering all of these differences (deviations) as positive, and could then compute either the median or the mean of these deviations. The first of these measures would be known as the *Median Deviation* from the mean, and is sometimes called the *Probable Deviation* or, in distributions of sampling errors, the *Probable Error* (P.E.). The mean of the deviations from the mean is sometimes called the *Mean Deviation* (or M.D.) but more frequently the *Average Deviation* (or A.D.).[1] Each of these measures is relatively easy to interpret. The median deviation is the absolute amount of deviation from the mean that is exceeded by half of the measures in the distribution. Thus, to say that the median deviation in height for a given group of individuals is two inches is to say that half of the individuals in the group differ in height from the average individual in the group by two inches or

[1] *Mean Deviation* is a better name, since it recognizes the distinction between *average* as a general term and *mean* as a specific term. The name *Mean Deviation* will therefore be used in this course, in spite of the prevalence of the name *Average Deviation* in educational and psychological literature.

more. The mean deviation is only a little more difficult to interpret. The mean deviation (M.D.) is ordinarily larger than the median deviation, for reasons that the student should be able to deduce for himself. (See Question 8, page 47 of the manual.)

The *Standard Deviation* is by far the most widely used measure of variability. It is similar to the average deviation except for the fact that each deviation is squared before averaging and the result then reduced to a magnitude comparable to the original deviations by extracting its square root. To compute directly the standard deviation of weights for a group of individuals, we would first find the amount by which the weight of each individual differed from the mean weight. We would then square this deviation for each individual, add these squared values together and divide by their number, and then extract the square root of the result. No attempt will be made here to explain the advantage of thus squaring each deviation and later extracting the square root of the average. The student will have to take it on faith that this procedure results in a more reliable measure of variability than the simpler M.D., and one that is better adapted for use in more complicated statistical theory, as in sampling error and correlation theory. It is this fact that the standard deviation is essential in the calculation of other statistical constants that results in its being used so much more widely than the M.D. or Q. If we were concerned only with the description of variability and had no occasion to use more complicated statistical techniques, we would probably use the M.D. in preference to the S.D. in most cases. Since, however, in most statistical analyses we must in any event compute the standard deviation in order to calculate other statistical measures, we use the S.D. instead of the M.D. for the simpler descriptive purposes.

Computation of the M.D.

As has already been suggested, the M.D. of a series of measures can be computed by finding the difference between each individual measure and the mean of the series, and then finding the mean of

these differences, all differences being considered as positive. The basic formula for the M.D. may be written as follows:

$$\text{M.D.} = \frac{\Sigma x}{N} \qquad (2)$$

in which x represents the deviation of any measure from the mean (not from the A.R.). Σ means "the sum of." Only the absolute magnitudes of the deviations are taken, that is, all deviations are considered as positive.

To compute the M.D. by the direct method just suggested would ordinarily be very time-consuming, particularly if the mean had been carried to several decimal places and if the number of cases were large. A much more practicable procedure is to compute the M.D. from the grouped frequency distribution by a "short" method similar to that used for computing the mean.

The steps in this short method are described below. The statements in brackets following each step show how it is applied in the illustrative problem in Table 14 on page 74.

STEPS IN COMPUTATION OF THE M.D. BY THE SHORT METHOD

1. Prepare a grouped frequency distribution, and compute the mean by the short method. If it is then found that the mean is not contained in the same interval with the arbitrary reference point, it will be necessary to construct a new pair of *d* and *fd* columns with the A.R. taken as the midpoint of the interval which contains the mean. The following steps assume that the *d* and *fd* columns used will satisfy this *essential* condition.

   [In the illustrative problem, the arbitrary reference point used in the original computation of the mean was the midpoint of the interval 73–77. Later it was found that the mean (71.20) did not lie in this interval. The original *d* and *fd* columns were therefore discarded, and another pair constructed with 70 (the midpoint of the interva which contains 71.20) as the A.R.]

2. Find the sum of the *frequencies* for the intervals whose *midpoints* are *above* the mean. Call this f_a.

   [The interval 73–77 is the first whose midpoint is above 71.20. The sum of the frequencies in this and higher intervals is 25. Hence, $f_a = 25$.]

3. Find the sum of the frequencies for the intervals whose midpoints are below the mean. Call this f_b.

   [The interval 68–72 is the highest whose midpoint is below 71.20. The sum of the f's in this and all lower intervals is 29. Hence, $f_b = 29$.]

4. Find the difference between f_a and f_b. That is, find $(f_a - f_b)$.

   [$25 - 29 = -4$.]

5. Find the difference between the mean and the arbitrary reference point, that is, find A.R. − M. Call this result c'.

   [A.R. − M. = 70 − 71.20 = −1.20. Hence, $c' = -1.20$. If the mean had been originally computed from an A.R. of 70, the value of c' would already have been found in computing the mean.]

6. Find the *product* of the results of Steps 4 and 5. That is, find $c'\ (f_a - f_b)$.

   [$(-4) \times (-1.20) = 4.80$. Hence, $c'\ (f_a - f_b) = 4.80$.]

7. Add the numbers in the fd column *without regard to sign.* Denote the result by $\Sigma\,|fd|$.

   [The *absolute* sum of the fd's is 85. Hence, $\Sigma\,|fd| = 85$.]

8. Multiply this result by the size (i) of the interval. That is, find $i\ (\Sigma\,|fd|)$.

   [The interval used is one of five units. Hence, $i\ (\Sigma\,|fd|) = 5 \times 85 = 425$.]

9. Add to this result the result of Step 6. That is, find $i\,(\Sigma\,|fd|) + c'\ (f_a - f_b)$.

   [$425 + 4.80 = 429.80$.]

10. Divide this result by N to get the M.D. That is, find

$$\text{M.D.} = \frac{i\ (\Sigma\,|fd|) + c'\ (f_a - f_b)}{N} \tag{3}$$

[M.D. $= \frac{429.80}{54} = 7.959\ldots$ or M.D. = 7.96.]

11. Round the result to the desired degree of accuracy, not to exceed the number of digits in the mean (which itself should contain no more digits than indicated by the rules, pages 65–66).

[The result is rounded to 7.96. The M.D. of a distribution of integral test scores should very rarely be carried to more than two decimal places.]

TABLE 14

ILLUSTRATION OF COMPUTATION OF M.D. BY THE SHORT METHOD

	f	Discarded because reference interval does not contain the mean. d	fd	d	fd	
93–97	1	4	4	5	5	$f_a = 25 \quad f_b = 29$
88–92	3	3	9	4	12	$f_a - f_b = -4$
83–87	2	2	4	3	6	$c' = \text{A.R.} - \text{M.} = 70-$
78–82	7	1	7/+ 24	2	14	$71.20 = -1.20$
73–77	12	0	0	1	12	$c'(f_a - f_b) = -4 \times$
68–72	10	− 1	− 10	0	0	$-1.20 = 4.80$
63–67	9	− 2	− 18	− 1	− 9	$i(\Sigma \lvert fd \rvert) = 5 \times 85 = 425$
58–62	5	− 3	− 15	− 2	− 10	$\text{M.D.} = \frac{425 + 4.80}{54} =$
53–57	3	− 4	− 12	− 3	− 9	
48–52	2	− 5	− 10/− 65	− 4	− 8	$\frac{429.80}{54} = 7.96$
N =	54					$85 = \Sigma \lvert fd \rvert$

Original computation of mean.

$$\text{Mean} = 75 + 5 \cdot \frac{(-41)}{54}$$
$$= 75 - 5 \times .7593$$
$$= 71.2035$$
$$= 71.20 \text{ (rounded)}$$

If the student is curious about the *reasons* for the various steps in this procedure, he may find them explained in any good reference book on statistics.[1]

Since the computational procedure used has no bearing on the *interpretation* of the M.D., no explanation of these steps will be attempted here. The student is asked to take this computational procedure on faith, and to do all of his thinking about the M.D. in terms of the fundamental formula (2), or in terms of the defini-

[1] See Holzinger, Karl J. *Statistical Methods for Students in Education*, pp. 102–107. Ginn and Company, 1928.

tion: *The M.D. is the mean of the deviations taken from the mean of the distribution.*

Computation of the S.D.

The standard deviation of a distribution may be defined as the square root of the mean of the squared deviations from the mean of the distribution. It may be found by finding the difference ($x = X - M$) between each individual measure and the mean of the distribution, squaring these differences individually, adding the squared deviations and dividing the sum by N, and then extracting the square root of the result. The fundamental formula for the S.D. is

$$\text{S. D.} = \sqrt{\frac{\Sigma x^2}{N}} \tag{4}$$

Again, because the direct method of computation just described is too time-consuming to be practicable, the short method described below should generally be used. As before, the statements in brackets following each step show how it is applied in an actual problem — that presented in Table 15.

STEPS IN COMPUTATION OF THE S.D. BY THE SHORT METHOD

1. Prepare a grouped frequency distribution of the measures and complete the d and fd columns as in the computation of the mean. Unlike the computation of the M.D., the midpoint of *any* interval may be used as the A.R.
2. Multiply each number in the fd column by the corresponding number in the d column, and write the result in a third column headed fd^2.

   [In the illustrative problem, the product of the d and fd values for the top interval is $4 \times 4 = 16$. The remaining numbers in the fd^2 column are similarly obtained.]
3. Find the *algebraic* sum of the numbers in the fd column and divide by N. That is, find $\frac{\Sigma fd}{N}$.

   $[\frac{\Sigma fd}{N} = \frac{24 - 65}{54} = \frac{-41}{54} = -.759$ (rounded).]

4. Square this result. That is, find $\left(\frac{\Sigma fd}{N}\right)^2$.

   [$(-.759)^2 = .576$ (rounded).]

5. Add the numbers in the fd^2 column and divide by N. That is, find $\frac{\Sigma fd^2}{N}$. Note that all numbers in this column are positive.

   [The sum of this column in Table 15 is 247. Hence $\frac{\Sigma fd^2}{N} = \frac{247}{54} = 4.574$.]

6. Subtract from this quotient the result of Step 4. That is, find $\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2$.

   [$4.574 - .576 = 3.998$.]

7. Extract the square root of this difference. That is, find

$$\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2}$$

   This is the standard deviation in *interval* units.

   [The square root of 3.998 is 1.99.]

8. Multiply this square root by the size of the interval to get the S.D. That is, find

$$\text{S.D.} = i\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} \qquad (5)$$

   [The S.D. of the distribution in Table 15 is $1.99 \times 5 = 9.95$.]

9. Round the result to the desired accuracy. (In general, the S.D. of a distribution of integral test scores should not be carried to more than two decimal places.)

As in the case of the M.D., the student is advised to accept on faith the statement that this computational procedure will yield very nearly the same result as that obtained when the direct method, described at the beginning of this section, is applied to the original measures. The standard deviation computed from a grouped frequency distribution will be slightly inaccurate because of the loss of indentity of the original measures (that is, because

TABLE 15

ILLUSTRATION OF COMPUTATION OF S.D. BY THE SHORT METHOD

M	f	d	fd	fd^2	
95	1	4	4	16	$\frac{\Sigma fd}{N} = \frac{-41}{54} = -.759$
90	3	3	9	27	
85	2	2	4	8	$\left(\frac{\Sigma fd}{N}\right)^2 = (-.759)^2 = .576$ (rounded)
80	7	1	7	7	
75	12	0	0		$\frac{\Sigma fd^2}{N} = \frac{247}{54} = 4.574$
70	10	−1	−10	10	
65	9	−2	−18	36	
60	5	−3	−15	45	
55	3	−4	−12	48	$\sqrt{\frac{\Sigma fd^2}{N} - \left(\frac{\Sigma fd}{N}\right)^2} = \sqrt{4.574 - .576} = \sqrt{3.998} = 1.99$
50	2	−5	−10	50	
					S.D. $= 5 \times 1.99 = 9.95$
$N = 54$		$\Sigma fd = -41$		$247 = \Sigma fd^2$	

of grouping errors), but this inaccuracy is nearly always too slight to be of any practical significance if the frequency distribution has been properly constructed. The short method of computation does not in itself result in any error.

The student should make no attempt, then, to *interpret* the S.D. in terms of Formula (5), but should do all of his thinking about this measure in terms of its definition or of Formula (4).

Important Characteristics of the Various Measures of Variability

Measures of variability are in general much more difficult to interpret than measures of central tendency. The observations in the following paragraphs, however, may help to make their meaning more clear.

We may note first that the several measures of variability may in one sense be considered as special types of "averages." Instead of representing "average" *position* on a scale, they represent "average" amounts of *deviation* from such a position. Q represents the mean amount by which the upper and lower quartiles deviate from the median, the *median deviation* represents the median amount by which the *individual* measures deviate from the mean of the distribution, the M.D. represents the mean amount by which the various individuals in a group differ from the mean individual,

while the *square* of the S.D. represents the mean value of the squared deviations from the mean.

It may be helpful, also, to observe that in any bell-shaped distribution, the S.D. will always be larger than the M.D., and the M.D. larger than the median deviation and Q. If the distribution approximates the form of the normal curve, the M.D. will be about five-sixths as large as the S.D., and the Mdn. Dev. (median deviation) and Q will each be about two-thirds as large as the S.D.

Each of the measures of variability may be thought of as a unit of *distance* along the scale in terms of which the position of any measure may be described with reference to the mean. If the distribution closely approximates the form of the normal curve, roughly two-thirds of the measures will lie within one S.D. of the mean, about 95 per cent of the measures will lie within two S.D.'s of the mean, and only a negligible proportion (usually less than 1 per cent) will deviate from the mean by more than three S.D.'s. Similarly, again for distributions closely approximating the form of a normal distribution, about 57 per cent of the measures will lie within one M.D. of the mean. In *any* symmetrical distribution, of course, 50 per cent of the measures will lie within one median deviation or within one Q of the mean. The trouble with the preceding generalizations is that they apply only to distributions that are very nearly normal or symmetrical, and that the majority of distributions with which we actually deal are neither approximately normal nor approximately symmetrical. Distributions may be found in which the S.D., when measured off on both sides of the mean, subtends more than 90 per cent of the cases, and others in which it subtends only slightly more than 50 per cent. In some distributions the M.D. is smaller than Q, and in some very much larger. Any generalizations such as those given in this and the preceding paragraph must therefore be used with extreme caution.

The complexity of the mathematical character of the S.D. makes it the most difficult to interpret of the various measures

of dispersion. This difficulty is further increased by the fact that the S.D. is so often expressed in units which are not in themselves meaningful. Little meaning, for example, can be derived from the statement that the distribution of scores on the *Iowa Every-Pupil Test in Algebra* for the pupils in the ninth grade of the Jonestown High School shows a standard deviation of 6.5. We cannot conclude from this statement that these pupils are either highly variable or very much alike in achievement, primarily because we do not know what amount of difference in achievement 6.5 units on this test represents, but also because of the complexity of the S.D. If, however, we know that the S.D. of scores on the same test is 8.2 for the ninth graders in the Smithville High School, we can say that the latter group is *more* variable than the first in whatever the test is measuring. In spite of the complexity of the S.D., it is apparent that the group with the larger S.D. must be that in which the individual differences (or individual deviations from the mean) are more extreme.

The interpretation of the other measures of variability is similarly affected by the ambiguity of the measuring scale used. In general, therefore, these measures are most useful in education and psychology for *comparisons* of variability in two or more groups. Their usefulness in the description of a single group is largely limited to those instances in which they may be referred to a meaningful *standard*, but such descriptions, of course, also involve comparison. For example, if we knew that in the *typical* Iowa high school the S.D. of scores on the aforementioned algebra test was 4.8, and that the largest S.D. reported in any Iowa high school was 8.5, then we could say that the pupils in the Smithville High School constitute "an unusually heterogeneous group."

The Uses and Interpretation of Measures of Variability

Certain of the uses most frequently made of measures of variability will be suggested in the study exercises. In addition to these uses the S.D. finds important applications in sampling error theory, in correlation theory, in transforming test scores into

comparable derived measures, in "scaling" the difficulty of test items, and in the description of test reliability. Most of these latter uses depend upon the relationship of the standard deviation to the normal curve, and will be discussed in later chapters, subsequent to a consideration of the properties of the normal curve.

As was true of the measures of central tendency, each measure of variability has unique characteristics which make it superior to the other measures for certain purposes and inferior for others. The study exercises will assist the student to recognize the significance of these characteristics.

CHAPTER VII

THE NORMAL CURVE OF DISTRIBUTION

The Characteristics of the Normal Curve

THE normal curve of distribution, more commonly known simply as the *normal curve*, is a mathematical concept of great significance in statistical theory. Why it is so significant will be explained later in this chapter and in those to follow, but before considering its applications it may be well to consider first just what the normal curve is — what are its mathematical properties and general characteristics.

The normal curve of distribution may be most rigidly defined as the frequency curve whose height at any point is inversely proportional to the antilogarithm of half of the square of the distance (measured in units of the standard deviation) of that point from the mean, or as a curve in which the ordinate (y) at any given number of sigma-units from the mean is given by the expression:

$$y = y_o e^{-\frac{z^2}{2}} \tag{6}$$

in which y_o is the ordinate at the mean, e is the base (2.7183) of the Naperian system of logarithms, and z is the distance of the given ordinate from the mean, measured in units of the standard deviation of the distribution.

This definition, of course, will not be very meaningful to any student in this course who has not had advanced training in mathematics, nor is he advised to attempt to derive much meaning from it. It is presented here primarily in order to emphasize early in this discussion that the normal curve is essentially a mathematical *ideal* — an ideal, not in the sense of a standard of perfection or excellence, but in the sense of a product of the imagination. Many similar ideal curves have no counterpart in reality; the normal curve, however, happens to describe quite accurately

the form of distribution of certain types of actual data, and thus is of practical significance in statistical work.

A description which will be more meaningful to most students than the preceding definition is that the normal curve is a symmetrical bell-shaped frequency curve which exhibits a certain unique set of relationships between the ordinate at the mean of the distribution and the ordinates at various sigma-distances from the mean. This unique set of relationships is presented in part in Figure 8. Since the curve is bilaterally symmetrical, only half of the curve is shown.

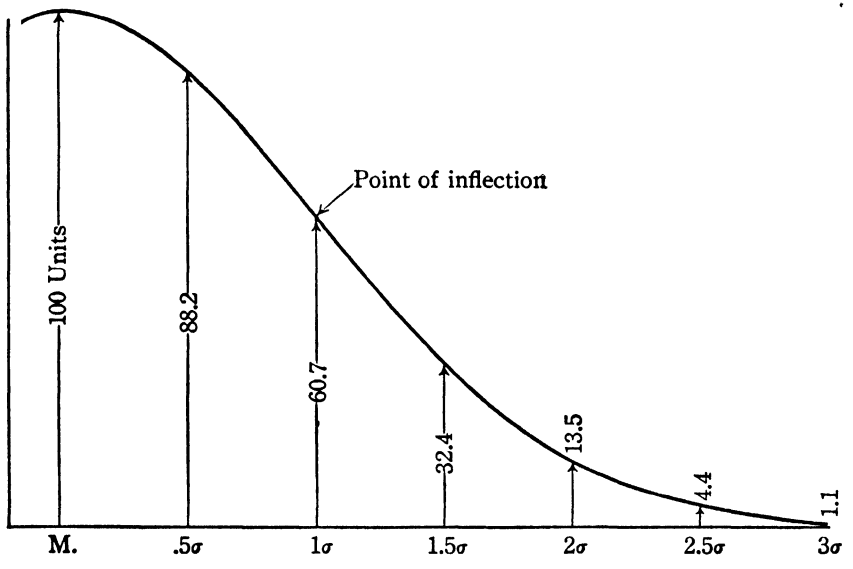


FIG. 8.
Ordinates under the normal curve.

As is indicated in Figure 8, in *any* normal curve the ordinate 1 S.D. from the mean is 60.7 per cent of the ordinate at the mean. The ordinate at 2σ from the mean is 13.5 per cent of the mean ordinate, and that at 3σ from the mean is 1.1 per cent of the mean ordinate. A similar statement can be made about the ordinate at any given sigma-distance from the mean. Table 16 presents these relationships more accurately for ordinates at one-hundredths

of a standard deviation intervals.[1] Table 16, then, may be considered as an alternate definition of the normal curve.

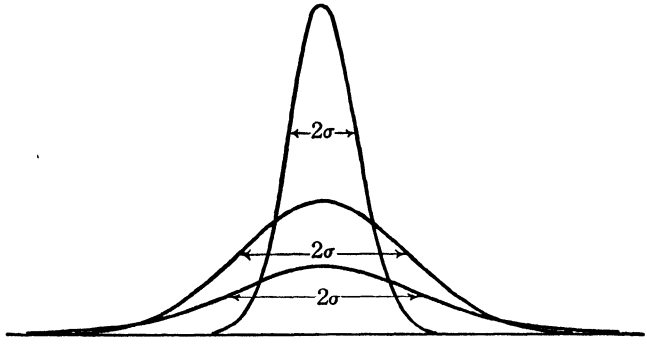


FIG. 9.
Normal curves of varying ratios of height to "width."

Any size of linear unit, of course, can be used to represent 1 sigma in plotting the curve, and the curve may be drawn with any desired height of the mean ordinate. The appearance of the curve will differ markedly, depending upon the choice of scale units in plotting it. Each of the curves in Figure 9,[2] for example, is a true normal curve. In each of them the ratio between the mean ordinate and the ordinate at any given sigma-distance from the mean is the same as in Figure 9, or as given in Table 16. Each of these curves is equally "flat" or "peaked," as these terms are used in statistics, although their *apparent* flatness or peakedness may differ considerably. The effect of these variations in plotting is to make it very difficult to recognize by inspection whether or

[1] Table 16 may be read as follows: Suppose we wish to find the ratio between the mean ordinate and the ordinate at 2.17 sigma-units from the mean. We look for 2.1 in the column under $\frac{x}{\sigma}$ at the left of the table, and then follow along the *row* thus identified until we get to the column headed .07. The ratio desired is that which is in both the 2.1 row and the .07 column, and is .0950. The height of the curve at 2.17 S.D.'s from the mean is then .0950 or 9.5 per cent of its height at the mean. Similarly, the ordinate at .62 sigma-units from the mean is .8251 or 82.5 per cent of the mean ordinate.

[2] The ideal normal curve has no definite *width*, since it is asymptotic to the base line. In picturing such curves, however, usually we arbitrarily cut off the curve at about 3 S.D.'s from the mean at either end, since only a negligible proportion of the area under the curve is beyond these limits.

TABLE 16

ORDINATES UNDER THE NORMAL CURVE AT VARIOUS SIGMA-DISTANCES FROM THE MEAN (ORDINATES EXPRESSED AS PROPORTIONS OF THE MEAN ORDINATE) [1]

$\frac{x}{\sigma}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	1.0000	1.0000	.9998	.9996	.9992	.9988	.9982	.9976	.9968	.9960
0.1	.9950	.9940	.9928	.9916	.9903	.9888	.9873	.9857	.9839	.9821
0.2	.9802	.9782	.9761	.9739	.9716	.9692	.9668	.9642	.9616	.9588
0.3	.9560	.9531	.9501	.9470	.9438	.9406	.9373	.9338	.9303	.9268
0.4	.9231	.9194	.9156	.9117	.9077	.9037	.8996	.8954	8912	.8869
0.5	.8825	.8781	8735	.8690	.8643	.8596	.8549	.8501	.8452	.8403
0.6	.8353	.8302	.8251	.8200	.8148	.8096	8043	.7990	.7936	.7882
0.7	.7827	.7772	.7717	.7661	.7605	.7548	.7492	.7435	7377	.7319
0.8	.7262	.7203	.7145	.7086	.7027	.6968	.6909	6849	6790	.6730
0.9	.6670	.6610	.6550	.6489	.6429	.6368	.6308	.6247	.6187	.6126
1.0	.6065	.6005	.5944	.5883	.5823	.5762	5702	.5641	.5581	.5521
1.1	.5461	.5401	.5341	.5281	5222	.5162	.5103	.5044	.4985	.4926
1.2	.4868	.4809	.4751	.4693	.4636	.4578	.4521	.4464	.4408	.4352
1.3	.4296	.4240	.4185	.4129	.4075	.4020	.3966	.3912	3859	.3806
1.4	.3753	.3701	.3649	.3597	.3546	.3495	.3445	.3394	.3345	.3295
1.5	.3247	.3198	.3150	.3102	.3055	.3008	.2962	.2916	.2870	.2825
1.6	.2780	.2736	.2692	.2649	.2606	.2563	.2521	.2480	.2439	.2398
1.7	.2358	.2318	.2278	.2239	.2201	.2163	.2125	.2088	.2051	.2015
1.8	.1979	.1944	.1909	.1874	.1840	.1806	.1773	1740	.1708	.1676
1.9	.1645	.1614	.1583	.1553	.1523	.1494	.1465	.1436	.1408	.1381
2.0	.1353	.1327	.1300	.1274	.1248	.1223	.1198	.1174	.1150	.1126
2.1	.1103	.1080	.1057	.1035	.1013	.0991	.0970	.0950	.0929	.0909
2.2	.0889	.0870	.0851	.0832	.0814	.0796	.0778	.0760	.0743	.0727
2.3	.0710	.0694	.0678	.0662	0647	.0632	.0617	.0603	.0589	.0575
2.4	.0561	.0548	.0535	.0522	.0510	.0497	.0485	.0473	.0462	.0451
2.5	.0439	.0429	.0418	.0407	.0397	.0387	.0378	.0368	.0358	.0349
2.6	.0341	.0332	.0323	.0315	.0307	.0299	.0291	.0283	.0276	.0268
2.7	.0261	.0254	.0247	.0241	.0234	.0228	.0222	.0216	.0210	.0204
2.8	.0198	.0193	.0188	.0182	.0177	.0172	.0167	.0163	.0158	.0154
2.9	.0149	.0145	.0141	.0137	.0133	.0129	.0125	.0122	.0118	.0115
3.0	.0111									

[1] The data in this table were taken from *Tables for Statisticians and Biometricians*. Edited by Karl Pearson. Cambridge University Press.

not a given curve is normal, or to distinguish between one that is normal and one that is not. The polygon in Figure 10, for example, looks very much like a normal curve, and, if seen alone, would pass unchallenged as such by most persons, but it is actually too "flat" to be normal, as is shown by the superimposed normal curve.

The normal curve has the additional properties that it is *asymptotic* to the base line (when extended to greater distances beyond the mean, it continues to approach but never reaches the base line) and that its *points of inflection* (the points where the curvature changes in direction) are each 1 S.D. from the mean ordinate

Another characteristic — most important of all — is considered in the next section ("Area Relationships under the Normal Curve").

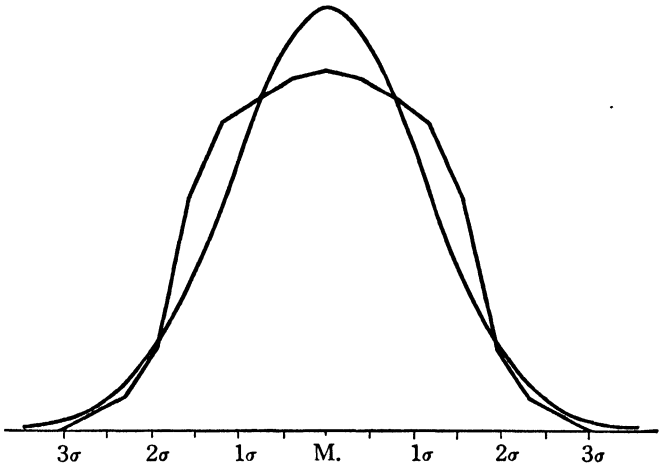


FIG. 10.
Normal curve "fitted" to a frequency polygon.

It should be noted that the term *normal* as here used is simply a *name* for this particular curve, and does not have any of the usual connotations as, for example, in speaking of "a normal child." *Normal,* as a technical term in statistics, does not mean "the ordinary or usual condition" or "free from abnormalities."[1]

In the subsequent discussion, a "normal distribution" means any frequency distribution whose form corresponds to that of the normal curve.

Area Relationships under the Normal Curve

Since the ordinate at a given sigma-distance from the mean of the normal curve always has the same relationship to the mean ordinate, it follows that the *area* under the curve included between the mean ordinate and an ordinate a given sigma-distance from

[1] "There is nothing arbitrary or mysterious about variability which makes the so-called normal type of distribution a necessity, or any more rational than any other sort, or even more to be expected on a priori grounds. Nature does not abhor irregular distributions." — Thorndike, E. L., *Mental and Social Measurements,* pp. 88–89.

the mean is always the same *proportion* of the total area under the curve. If under a normal curve we erect perpendiculars from the base line at the mean and at a point 1 S.D. from the mean, the part of the area under the curve included between these perpendiculars will always be 34.13 per cent (rounded) of the total area. The shaded area in Figure 11 corresponds to that just described.

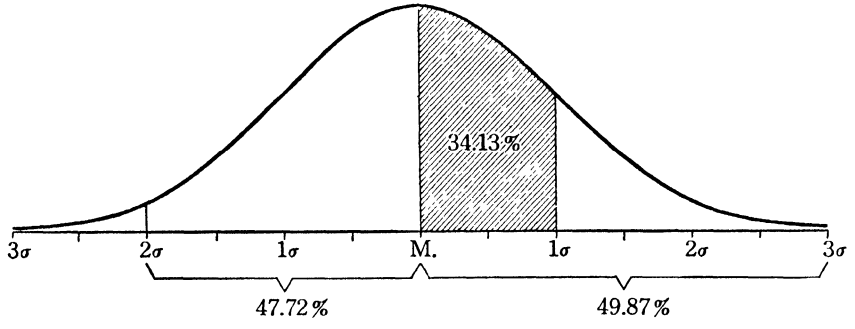


FIG. 11.
Area relationships under the normal curve.

Similarly, 47.72 per cent of the total area will be included between the mean ordinate and an ordinate 2 S.D.'s from the mean, and 49.87 per cent of the total area will be subtended between the mean and a point 3 S.D.'s from the mean. These statements and similar statements for ordinates at other distances from the mean apply alike to any normal curve (regardless of the choice of units in which it is plotted). These relationships for ordinates at one-hundredths of sigma intervals are given in Table 17. (In this table, x represents the distance from the mean, and hence $\frac{x}{\sigma}$ represents that distance in standard deviation units.) These $\frac{x}{\sigma}$ ratios are expressed in tenths along the vertical margins and in hundredths along the horizontal margins of the table. The numbers within the body of the table represent the *per cents* of the total area which are included between the mean ordinate and the ordinates at these various sigma-distances from the mean. Since in any frequency curve the number of units in its area is proportional to the number of cases in the distribution — that is, since the area

represents the frequency — the numbers in Table 17 also represent per cents of the total frequency, as the title of the table indicates.

Table 17 may be read as follows: Suppose we wish to find what per cent of the total area under the normal curve is between the mean and a point 1.36 S.D.'s from the mean. To find this percentage, we run down the column under $\frac{x}{\sigma}$ until we get to 1.3. We then follow along the row thus determined until we get to the column headed .06. The number which is both in the 1.3 row and in the .06 column is then seen to be 41.31, which is the per cent desired.

TABLE 17

PER CENT OF TOTAL AREA UNDER THE NORMAL CURVE BETWEEN MEAN ORDINATE AND ORDINATE AT ANY GIVEN SIGMA-DISTANCE FROM THE MEAN [1]

$\frac{x}{\sigma}$	.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 90	33 15	33 40	33 65	33 89
1 0	34 13	34 38	34 61	34 85	35 08	35 31	35 54	35 77	35 99	36 21
1 1	36 43	36 65	36 86	37 08	37 29	37 49	37 70	37 90	38 10	38 30
1 2	38 49	38 69	38 88	39 07	39 25	39 44	39 62	39 80	39 97	40 15
1 3	40 32	40 49	40 66	40 82	40 99	41 15	41 31	41 47	41 62	41.77
1 4	41.92	42 07	42 22	42 36	42 51	42.65	42.79	42 92	43 06	43.19
1 5	43 32	43 45	43 57	43 70	43 83	43 94	44 06	44 18	44 29	44 41
1 6	44 52	44 63	44 74	44 84	44 95	45 05	45 15	45 25	45 35	45 45
1 7	45 54	45 64	45 73	45 82	45 91	45 99	46 08	46 16	46 25	46 33
1.8	46 41	46 49	46 56	46 64	46 71	46 78	46 86	46 93	46 99	47 06
1 9	47 13	47 19	47 26	47 32	47.38	47 44	47 50	47 56	47.61	47 67
2 0	47 72	47 78	47 83	47 88	47 93	47.98	48 03	48 08	48 12	48 17
2 1	48 21	48 26	48 30	48 34	48.38	48 42	48.46	48 50	48.54	48 57
2 2	48 61	48 64	48 68	48 71	48 75	48 78	48 81	48 84	48.87	48 90
2.3	48 93	48 96	48 98	49 01	49 04	49 06	49 09	49.11	49 13	49 16
2 4	49 18	49 20	49 22	49 25	49 27	49 29	49 31	49.32	49 34	49.36
2 5	49 38	49 40	49 41	49 43	49 45	49 46	49.48	49 49	49 51	49.52
2.6	49 53	49 55	49.56	49 57	49.59	49 60	49 61	49 62	49 63	49.64
2 7	49 65	49 66	49 67	49 68	49.69	49 70	49 71	49 72	49.73	49 74
2 8	49 74	49.75	49 76	49.77	49.77	49.78	49.79	49 79	49.80	49 81
2.9	49.81	49.82	49 82	49 83	49 84	49.84	49 85	49 85	49 86	49.86
3.0	49.87									
3.5	49 98									
4.0	49.997									
5.0	49.99997									

[1] The data in this table were taken from *Tables for Statisticians and Biometricians.* Edited by Karl Pearson. Cambridge University Press.

This table may be employed to derive a number of important types of information about any normal distribution of measures. These types of information, and ways in which Table 17 is used to derive them, are explained in the numbered paragraphs below.

1. To find the number, proportion, or per cent of the cases in a normal distribution which lie above (or below) any point along the scale.

   Illustration: Given a normal distribution with M = 90, S.D. = 15, and N = 150. To find the per cent of the cases in the distribution which lie above 110. This measure is 110–90 = 20 units above the mean, or (since the S.D. is 15) $\frac{20}{15}$ = 1.33 (rounded) sigma units above the mean. According to Table 17, 40.82 per cent of the measures in the distribution will lie between the mean and this point, that is, between 90 and 110. Since the distribution is symmetrical, 50 per cent of the cases will lie above the mean. Hence, 50 − 40.82 = 9.18 per cent of the cases will lie above 110. This result may also be expressed either as a proportion (.0918) of the whole distribution, or as a *number* of cases (9.18 per cent of 150 = 13.77 cases).

   As a further example, suppose we wish to find the per cent of cases that are below 63. Since this measure is below the mean, we would first find the per cent between this score and the mean, and then *subtract* this percentage from 50 per cent, the total per cent below the mean. The result will be 3.59 per cent, which could be expressed also as a proportion or as the number of cases in the manner already described.

2. To find the number, proportion, or per cent of the cases which lie between any two given points along the scale.

   Illustration: To find the per cent of the total number of cases which lie between 85 and 100 in the distribution just considered. This percentage may be considered as the sum of two percentages: the percentage between 85 and the mean

(90), and the percentage between the mean and 100. 85 is .33σ below the mean, and hence, according to Table 17, 12.93 per cent of the cases would lie between it and the mean. Similarly, 24.86 per cent would lie between the mean and 100, making a total of 37.79 per cent between 85 and 100.

If both of the given points lie on the same side of the mean, the percentage of cases included between them must be considered as the *difference* between the percentages included between each and the mean. For example, in this distribution, 28.81 per cent of the cases would lie between 78 and the mean, and 7.93 per cent would lie between 87 and the mean. Hence, $28.81 - 7.93 = 20.88$ per cent would lie between 78 and 87.

3. To find the point on the scale above (or below) which a given number, proportion, or per cent of the cases in a distribution lie.

Illustration: To find the point above which 30 per cent of the cases lie, in the distribution used in the preceding illustrations. If 30 per cent of the cases lie above a desired point, then 20 per cent must lie between that point and the mean. We must first find, then, how many sigma units we must go away from the mean in order to subtend 20 per cent of the cases. To do this, we search *within the body of* Table 17 to find the number nearest 20. This number is 19.85, which corresponds to a deviation of .52 sigma units, since it lies in the .5 row and in the .02 column. The desired point, then, is .52σ above the mean. Since $\sigma = 15$, this point is 7.8 units above the mean, and is equal to $90 + 7.8 = 97.8$. Thirty per cent of the measures in the distribution, then, lie above 97.8. Similarly, to find the point above which 75 per cent of the measures lie, we would note that, since 50 per cent of the measures are above the mean, 25 per cent of the measures must be between the desired point and the mean. We would then look within the body of the table to find the number nearest 25 per cent, which is 24. 86. This corresponds

to a sigma deviation of .67. Hence, the desired point is .67 × 15 = 10.05 units below the mean, or is at the point 90 − 10.05 = 79.95. Seventy-five per cent of the cases, then, lie above the point 79.95 in the given distribution.

4. To find the distance on either side of the mean which subtends a given number, proportion, or per cent of the cases.

   Illustration: To find the distance on either side of the mean which subtends the middle one-third of the cases in the distribution already considered. We must first find two points at equal distances from the mean in either direction between which 33⅓ per cent of the cases lie. This means that 16.666 or 16.67 per cent (rounded) of the cases will lie between the mean and either one of these points. We then look in the body of the table for the number nearest 16.67. This number is 16.64, which corresponds to a sigma deviation of .43. The desired distance, then, is .43 of a standard deviation, or 6.45 units. Hence, the middle one-third of the cases in the distribution are within 6.45 units of the mean. Accordingly, two-thirds of the cases in the distribution will deviate from the mean by *more* than 6.45.

5. To find the *probability* that a single case selected at random from a distribution will lie above (or below) a given point on the scale.

   Illustration: To determine the number of chances in 100 that a single score selected at random from the distribution already considered will have a value above 120. We note, following the procedure described under 1 above, that 2.28 per cent of the cases in the distribution will lie above 120. Since each score in the distribution has an equal chance of being the one drawn, and since 2.28 out of every 100 scores in the distribution lie above 120, the chances are 2.28 in 100 that the single score selected will exceed 120 in value.

   The *probability*, expressed in terms of "chances in 100," that a given score selected at random will satisfy any given condition is the same as the *per cent* of the cases in the whole

distribution that satisfy this condition. The reasonableness of this statement may be more apparent if considered with reference to other types of situations. Suppose, for example, that we wish to state the probability that a single card selected at random from a deck of playing cards will be a diamond. We know that each card in the deck has an equal chance of being drawn and that one-fourth of the cards in the entire deck are diamonds. Hence, we say that the chances are 1 in 4 or 25 in 100 that a single card drawn from the deck will be a diamond. Similarly, the chances are 1 in 13 or 7.7 in 100 of drawing a card of any given denomination, as, for example, a king. Similarly, if a bag contains a large number of marbles, 27 per cent of which are black, 60 per cent white, and 13 per cent red, the chances are 27 in 100 that a single marble drawn at random from the bag will be black, 60 in 100 that it will be white, and 13 in 100 that it will be red.

In the majority of the applications which the student will make of Table 17, the results will be expressed in terms of probability, and hence it is particularly important that he understand thoroughly this and the following uses of Table 17. (Numbers 5 to 8 in this series.)

6. To find the point with reference to which the probability is of a given value that a single case selected at random will lie above (or below) that point.

   This, as the student will recognize from the preceding discussion, is equivalent to 3 above, since the probability desired is the same as per cent of the cases that lie above (or below) a given point.

7. To find the probability that a single case selected at random will lie between two given points.

   This is equivalent to 2 above.

8. To find the amount of deviation from the mean for which the probability is of a given value that a single case selected at random will deviate from the mean by more or less than that amount.

This is equivalent to 4 above. For example, if we wish to find a deviation from the mean such that the chances are even, or 50 in 100, that any given score selected at random will deviate from the mean by less than this amount, we would find that sigma deviation from the mean which subtends 25 per cent of the cases. To do this, we would look in the table for the number nearest 25, which, as we have seen under 3 above, corresponds to $\frac{x}{\sigma} = .67$. If we wished to determine this value more accurately, we could interpolate between the values given in the table. For example, the numbers in the body of the table nearest 25 per cent are 24.86 and 25.17. The difference between these numbers is $25.17 - 24.86 = .31$. $25 - 24.86 = .14$. Hence, 25 lies $\frac{14}{31}$ of the distance between 24.86 and 25.17. The sigma-distance corresponding to 24.86 is .67, and that corresponding to 25.17 is .68. Hence, the sigma deviation desired corresponding to 25 per cent is $\frac{14}{31}$ of the distance between .67 and .68, or $.67 + \frac{14}{31} \times .01 = .67 + .0045 = .6745$. Thus, the middle one-half of the cases in the distribution are within $.6745\sigma$ of the mean. In other words, the chances are even that any score selected at random from a normal distribution will deviate from the mean by less than $.6745\sigma$ (or by more than $.6745\sigma$). For this reason, $.6745\sigma$ is known as the *probable deviation* (from the mean) of any measure selected at random from a normal distribution.

In general, it is not essential in most applications in education and psychology to interpolate between the values given in Table 17 as was done in the illustration under 8 above. In other words, the student may use as a sufficiently close approximation the value given in the table which is *nearest* that desired. For the relatively few situations in which higher accuracy is demanded in

practice, there are other tables [1] available in which the results are carried to a larger number of decimal places and which may be used instead of resorting to interpolation in coarser tables.

The Significance of the Normal Curve in Education and Psychology

If the student were to make a broad and representative collection of frequency distributions from the actual data which may be found in the research literature of education, psychology, anthropometry, and other related fields, and if he were to plot a smooth frequency curve for each of these distributions (making them comparable by plotting all to the same sigma scale and all with the same total area), he would find that his collection contained a wide variety of forms of distributions. Some curves would be badly skewed to the right, some moderately skewed to the left, some bimodal, some "U-shaped," some "J-shaped," and some almost rectangular. A large proportion of them could be roughly described as bell-shaped and as approximately symmetrical in form, with a single mode near the center of the range and with gradually decreasing frequencies in each direction, but among these bell-shaped curves some would have a high narrow peak with long flat "tails," others would have broad flat "humps," and would tail off more sharply at the extremes, and still others would show intermediate degrees of flatness or peakedness. (See Figure 7, page 49, and Figure 10, page 85.) How great may be the variation in forms of distributions, even of a single trait, is strikingly illustrated by the age distributions presented in Figure 12 (page 94), which are taken from the *Report of the Fifteenth Census of the United States* (1930).

Because of this extreme variation in form, the student would find it impossible to phrase a single generalized description that would apply accurately to more than a small proportion of the distributions collected. He might be able to classify all distributions into a number of fairly distinct types, and to provide a

[1] *Tables for Statisticians and Biometricians.* Edited by Karl Pearson. Cambridge University Press.

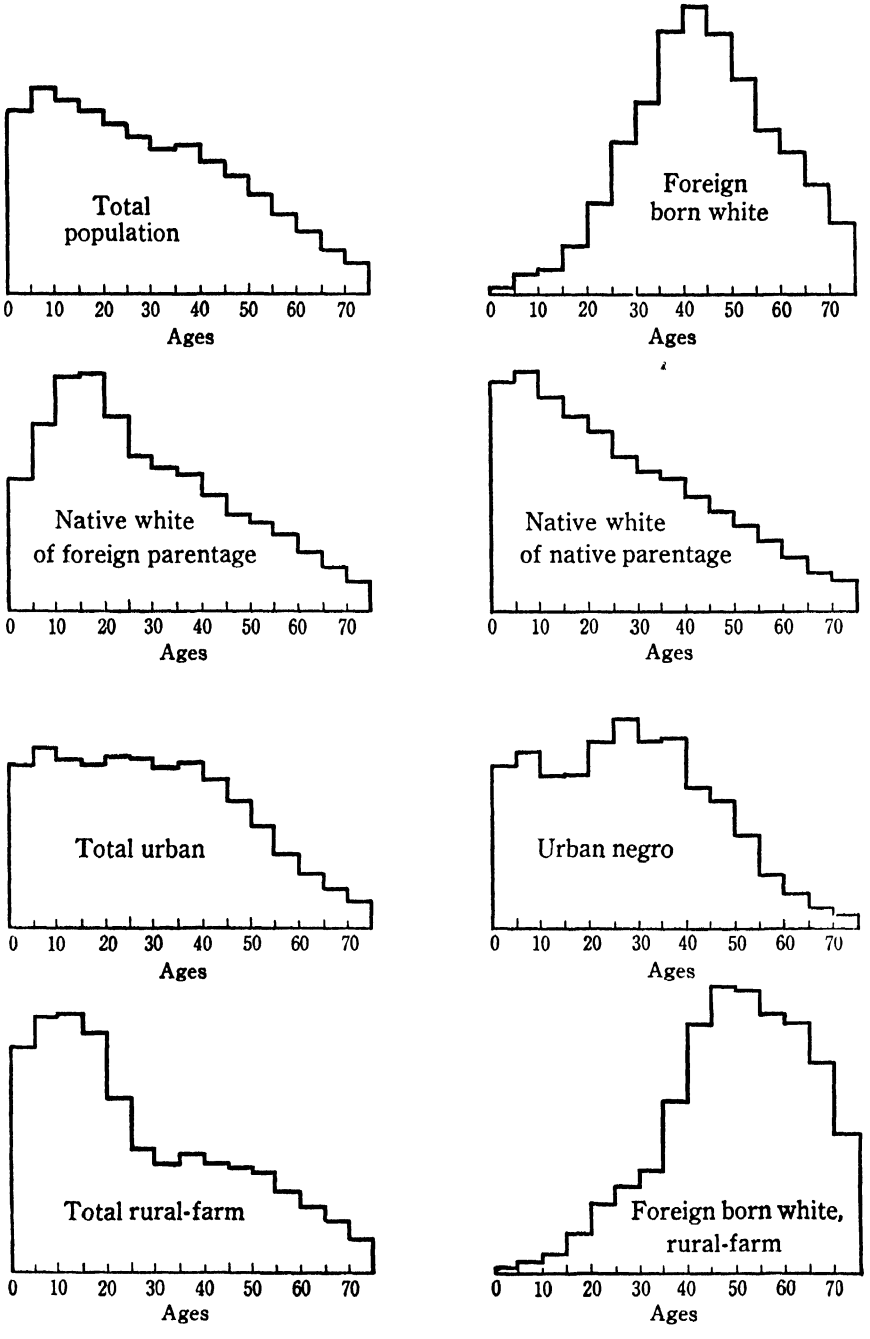


FIG. 12.

Age distributions of male population of the United States (from fifteenth census of United States, 1930).

generalized description of the form of the distributions in each classification. For example, he might find that many of the curves are of the general type represented by curve A in Figure 7, that many others are of the type represented by curve B, while others are roughly of the C type, etc. Yet he could find no *single* generalized curve which would provide a close fit to each of the distributions in this collection.

There is, then, no universal "law" concerning the form of frequency distributions in general. Unfortunately, however, there appears to have been built up in the literature of education and psychology the false conception that there *is* a single generalized frequency curve which does accurately describe the fundamental form of nearly all distributions of educational and psychological data. This misconception has been encouraged by the discussions in many textbooks in elementary statistics in these fields. Specifically, students have been led to believe erroneously that the *normal* curve constitutes such a generalized description, and that there is an underlying "law of normality" which applies to *all* or nearly all types of educational and psychological data.[1] Because of the very wide prevalence of this erroneous notion of a universal law of normality, and because many students beginning

[1] The following are direct quotations from a number of statistics texts in education and psychology

"Most mental and biological measures are distributed according to the normal curve if a sufficiently large number of such measures are distributed."

"Measures of natural phenomena, as well as measures of mental and social traits, tend to be distributed symmetrically about their central tendency in proportions which are determined by the laws of chance."

"If a reasonably large number of measures of some trait or characteristic are tabulated, they will in most cases approximate a normal distribution."

"This symmetrical or bell-shaped distribution is so nearly universal in statistics that it has come to be called the normal curve.... Many scientists have come to accept with some reservations the view that distributions of traits and abilities from representative groups tend to be symmetrical or normal.... Therefore, any serious departure from the normal curve... is in general interpreted that the traits or abilities measured do not represent a random sampling of such traits or abilities.... Consequently, if we wish to be sure that our computations of central tendency or variability are accurate, we must measure these traits or abilities in a sufficient number... to obtain a normal distribution."

In most instances, these statements have later been qualified in the same discussions, but not sufficiently to impress upon the student how numerous and important are the exceptions to these broad and loose generalizations.

this course may already have this misconception firmly established in their minds, it is essential that we begin our consideration of the true place of the normal curve in education and psychology by demonstrating the falsity of this conception.

We may note, first of all, that it is impossible to talk meaningfully about *the* form of distribution of measures of any human trait, simply because the measures of any given trait may show different forms of distribution for different "populations" or classifications of individuals. The statement, "The form of the distribution of height is normal," for example, is not meaningful because we have not specified what particular classification or type of individuals is involved. To illustrate, it is meaningful to refer to the form of distribution of height for all seven-year-old boys in Iowa, or for white adult men in the United States, or for women between the ages of 20 and 30 in the United Kingdom, but since the form of the distribution would undoubtedly differ for each of these and other populations, and since no one of them can be considered as *the* population, we may not consider any single frequency curve as representing *the* form of the distribution of height measurements in general. Since, then, we cannot talk meaningfully about *the* form of the distribution of a *single* trait, it obviously is even less fruitful to attempt to describe in general the distribution of any and all traits.

Among all of the populations which might be considered, however, there are many for which measures of various human traits are distributed in a form closely approximating that of the ideal normal curve, just as there also are many for which measures of the same or of other traits may show a skewed distribution or a distribution of some other characteristic form. There are many physical traits, for instance, which will show approximately normal distributions *if* the population in question is highly homogeneous with reference to certain related traits. For example, the distribution of height will closely approximate the form of the normal curve if measures are plotted for a large sample of individuals who are all of the same race, age, and sex. The distribution of height, however

for a sample of mixed ages, races, and sexes might show any form of distribution, depending upon the proportion of persons of various ages, or various races, or of the two sexes in the whole sample. Again, weight is fairly [1] normally distributed for individuals of the same race, age, sex, and height.

Since measures of many physical traits do show approximately normal distributions for many homogeneous populations, it seems probable that the same would also be true for many mental traits. It is dangerous, however, to argue thus by analogy from one type of trait to another. No *assumptions* concerning the form of distribution of any trait should be made on this basis alone for any population. The important consideration in this connection is that we are *not* justified in talking loosely about any underlying "law of normality" as if such a law applied to the distribution of measures of *any* trait regardless of the character of the population considered. This is particularly important since so many of the populations in which we are interested in education and psychology are only very vaguely or ambiguously defined, and are seldom highly homogeneous with reference to other traits related to the one under consideration.

Perhaps one of the principal reasons that we have exaggerated or misrepresented the importance of the normal curve in education and psychology is the fact that the *scores* obtained on educational and psychological tests, for almost any unselected group of pupils, so frequently present what may be roughly described as a bell-shaped form of distribution. This, however, is not of any very fundamental significance, since most of these tests are deliberately constructed so as to yield approximately symmetrical distributions of scores. In nearly all objective achievement test construction, for instance, it is the aim of the test author so to adjust the difficulty of the items and the distribution of their difficulty that the average score made by the group to be tested will approximate half of the possible score and that the range

[1] Actually, these distributions are slightly skewed positively. It is a matter of common observation that excessive "over-weight" is much more common and extreme than excessive "under-weight."

of scores will extend from near zero to near the possible score. If he desired to do so, the test author could as easily prepare a test that would yield a distribution markedly skewed to the right or one markedly skewed to the left, or of almost any other form. Because the "units" employed in educational and psychological test scales are arbitrarily established, because they fluctuate in value even within the same scale, and because the amount of such fluctuation cannot be accurately controlled by the test author, we cannot conclude, simply because the obtained scores are symmetrically distributed, that if the same traits or abilities could be measured along a "true" scale with a constant unit, these "perfect" measures would also be symmetrically distributed. Furthermore, the scores obtained on educational and psychological tests are always characterized by accidental errors of measurement due to the limited sampling of items constituting the test itself, that is, due to the unreliability of the test. These accidental errors, as is true of certain other types of chance data, do tend to be normally distributed, and therefore tend to produce a normal distribution of the scores which contain these errors. It may be noted in this connection that the fact that a test shows a fairly normal distribution is in itself not necessarily an indication that the test is of high quality; in fact, the more completely worthless or unreliable a test may be — that is, the more the scores obtained on it are due only to chance — the more likely it is to present a normal distribution of scores.

It is not implied, because of the foregoing considerations, that the normal curve does not have a very important place in statistical methods as applied to educational and psychological data. On the contrary, there is one general type of data, with which the statistician must be very seriously concerned, which under certain conditions is almost invariably normally distributed. This type of data may be described in general as consisting of the various kinds of "errors" which characterize educational and psychological measurements, including errors or chance fluctuations in random sampling, errors in measurement (due to unre-

liability of the tests and measuring instruments used), errors of observation and judgment, and errors in prediction based on regression equations. These types of errors and the uses of the normal curve in their statistical analysis will be considered in later chapters.

The so-called "law of normality," then, may be safely considered as applying only to certain types of chance data, or, more specifically, to certain types of "errors" in the quantitative analysis of educational and psychological data. In the interests of sound thinking, the student should guard carefully against any tendency to over-generalize concerning the normal curve or to make too many *assumptions* of normality, particularly with reference to distributions of individual measures of mental or physical traits, and, most especially, when the population involved is not highly homogeneous with reference to other factors related to those studied.

"Fitting" a Normal Curve to a Frequency Polygon or Histogram

It has already been noted that it is very difficult to tell by inspection whether or not a given histogram or frequency polygon approximates closely the form of the normal curve. The only sure method of judging the "normality" of a distribution is that of direct comparison, and involves superimposing on the histogram or polygon a true normal curve of the same mean, standard deviation, and total area. The procedure in "fitting" a normal curve to an observed frequency distribution is described in the following paragraphs. Since the student will have few occasions to apply this method, he is advised to give this explanation only a cursory examination.

The procedure will be explained in terms of a concrete problem — to fit a normal curve to the distribution in Table 18 on page 100.

The first step is to determine the height of the fitted normal curve at the mean. This may be found by the formula

$$y_o = \frac{N}{\sigma\sqrt{2\pi}}$$

in which y_o represents the mean ordinate, N the number of cases

in the distribution, σ the standard deviation *in interval units*, and $\pi = 3.1416$.

The S.D. of the distribution in Table 18 is 9.95, or $\frac{9.95}{5} = 1.99$ in interval units. Hence,

$$y_o = \frac{54}{1.99\sqrt{2 \times 3.1416}} = \frac{54}{1.99 \times 2.51} = 10.81$$

Since in any normal curve the ordinate at any given sigma-distance from the mean is always a definite proportion of the mean ordinate, we can now determine the ordinate at the midpoint of each interval (from Table 16). To do this, we first determine how many sigma units (this time expressing sigma in the original units) each interval midpoint deviates from the mean. The midpoint of the interval 93–97 is $95 - 71.20 = 23.80$ score units from the mean, or $\frac{23.80}{9.55} = 2.392$ sigma units from the mean. The distance between the midpoints of any pair of adjacent intervals is 5 score units or $\frac{5}{9.95} = .5025$ sigma units. Hence the deviations of the remaining midpoints can be quickly determined. These deviations are given in the third column in Table 18.

The next step is to determine, for each interval, what is the

TABLE 18

COMPUTING FREQUENCIES IN THE NORMAL DISTRIBUTION CORRESPONDING TO AN OBSERVED DISTRIBUTION

Interval Midpoints	Observed Frequencies	Deviation from Mean	Ratio of Ordinate to Mean Ordinate	Ordinate (theoretical frequency)
100	0	2.89	.0154	.2
95	1	2.39	.0575	.6
90	3	1.89	.1676	1.8
85	2	1.39	.3806	4.1
80	7	.88	.6790	7.3
75	12	.38	.9303	10.1
70	10	− .12	.9928	10.7
65	9	− .62	.8251	8.9
60	5	− 1.13	.5281	5.7
55	3	− 1.63	.2649	2.9
50	2	− 2.13	.1035	1.1
45	0	− 2.63	.0315	.3
	$N = 54$			53.7
	M. = 71.20		Mean Ordinate =	10.81
	S.D. = 9.95			

ratio between the ordinate at its midpoint and the mean ordinate. These ratios may be read directly from Table 16 (page 84) For example, the number in the body of Table 16 corresponding to $\frac{x}{\sigma} = 2.89$ is .0154. This value, and others corresponding to the deviations of the other interval midpoints, are presented in the fourth column of Table 18.

The next step is to multiply each ratio in column 4 by the height of the mean ordinate. The result in each case will be the height of the fitted curve at the midpoints of the interval in question. Since this height is expressed in terms of the frequency scale, the numbers in the last column may also be considered as the "theoretical frequencies" in a normal distribution with the same M, S.D., and N as the one given. The sum of these frequencies should always be just slightly less than the N of the original distribution. In the illustration, for example, the sum of the theoretical frequencies is 53.7 as compared to 54 for the original frequencies.

The final step is to plot both the observed and the "theoretical" frequencies on the same scale. The observed frequencies may be represented either by a histogram or a ploygon, but a smooth curve should be drawn through the points determined by the theoretical frequencies. Figure 13 presents the results of this final step for the data in Table 18.

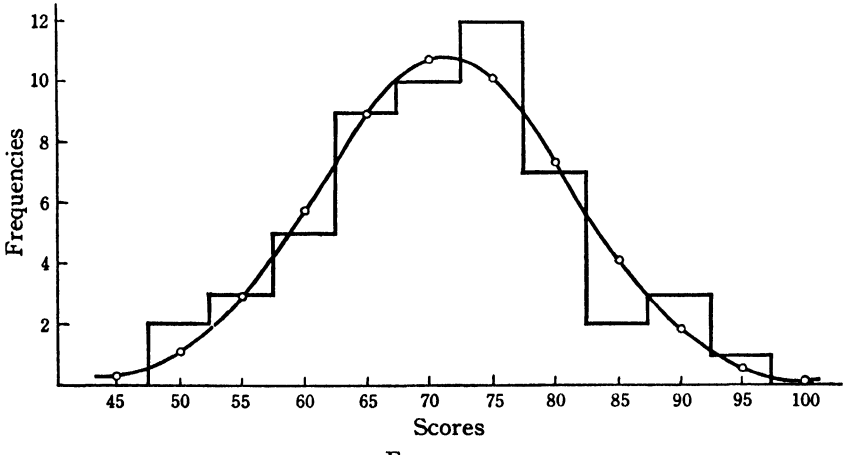


FIG. 13.
Normal curve "fitted" to a histogram.

CHAPTER VIII

SAMPLING ERROR THEORY

The General Nature of Sampling Studies

NEARLY all research studies in education and psychology are of the type known as *sampling* studies, in which measurements or observations are made of a limited number or "sample" of individuals in order that generalizations may be established about the still larger groups or "populations" of individuals that these samples are supposed to represent. Because the individuals comprising any of these populations differ from one another, and because chance or uncontrolled influences always play some part in determining which of these differing individuals are to constitute the sample used, any single fact obtained from a sample (such as a mean, median, percentile, standard deviation, etc.) is almost certain to differ by some amount from the corresponding fact for the whole population. Such "obtained" facts, therefore, may never be accepted at their face value as exactly descriptive of the population involved, but must always be considered as only approximations to, or as only *estimates* of, the corresponding "true" facts. In order that any such obtained fact may be properly interpreted, then, we need to know how "good" an estimate it is of the corresponding fact for the whole population; that is, we need to have some description of the dependability or *reliability* of the estimate and must qualify accordingly any generalization based upon it. Such descriptions of reliability are extremely important, since without them we might attribute real significance to facts that are of only accidental origin or read important meanings into mere coincidences.

Some of the more important statistical techniques used to secure these descriptions of reliability will be presented and explained later in this chapter. First of all, however, it might be well to consider, in terms of a concrete illustration, what are the major issues

and what is the general nature of the logic underlying sampling error theory.

Suppose that for some reason we wish to know the mean intelligence quotient (I.Q.), as measured by the Stanford Revision of the Binet-Simon scale, of all eighth grade pupils in the one-room rural schools in the state of Iowa. Since there are in Iowa about 9,000 one-room rural schools and a considerably larger number of rural eighth grade pupils, it obviously is beyond the facilities of any single research organization to administer an individual intelligence test to every pupil in this very large "population." In this situation, then, we would select a *sample* of rural eighth graders, consisting of a relatively small number of pupils, and would administer our intelligence test only to these pupils. We would then compute the mean I.Q. for these pupils and would consider this "obtained" mean as an *estimate* of the mean I.Q. of the entire population (the "true" mean).

The reliability of this obtained mean as descriptive of the entire population of eighth grade one-room rural school pupils would obviously depend upon the size and the representativeness of the sample employed, that is, upon how it was selected. There are a number of procedures that could be followed in the selection of the sample in a situation of this kind. (See pages 139 to 143.) One method would be to allow *chance* alone to determine which individuals from the whole population are to be selected. This could be done by securing the names of *all* eighth grade one-room rural school pupils in the state, typing each name on a slip of paper, mixing these slips very thoroughly in a container, and making a blindfold selection of the desired number of slips. An equivalent procedure would be to arrange all the names in alphabetical order and to select every fortieth or fiftieth or seventy-fifth name from the list until the desired number has been selected. A sample drawn by either of these methods would be known as a *random sample*, since the method of selection would guarantee independently to *every* individual in the whole population an equal chance of being one of those selected in the sample drawn. In actual practice, it is

rarely practicable to follow a procedure of the type just described. Other more practicable methods of sampling will be considered later in the closing section of this chapter. For the purpose of this illustration, however, we will assume that the sample involved has been selected at *random*, and that it consists of 81 pupils.

Since the individuals constituting the sample were selected by chance, we obviously could not expect the distribution of intelligence quotients for these individuals to correspond exactly to that for the whole population. By chance, our sample may contain a larger proportion of eighth graders of superior intelligence than would be found in all rural schools of the state, or it may contain a relatively large proportion of pupils of inferior intelligence. This would happen in exactly the same way and for exactly the same reason that a bridge hand dealt from a well-shuffled deck may contain more cards of one suit than of any other. In a sense, in drawing this sample the names were "shuffled" and a sample "dealt" in the same way that the deck is shuffled and hands dealt in a bridge game. The *mean* I.Q. for the pupils in this sample, then, could not be expected to agree exactly with the corresponding true mean, that is, the mean which would have been obtained had *all* pupils in the population been tested. Suppose, for instance, that the mean I.Q. for the pupils in this sample is 98.5. This fact would not enable us to infer at once that the mean I.Q. for all pupils in this population is 98.5, but only that the true mean is "somewhere near" 98.5. The next important consideration, then, is that of determining how "good" an estimate of the true mean is our obtained mean of 98.5. In other words, we need to know within what distance of the true mean we may be highly confident our obtained mean lies, or we need to know *how* confident we may be that the obtained mean is within any given distance of the true mean.

The Sampling Distribution of the Mean

Before attempting to describe thus the reliability of our obtained mean, let us first note that if we selected independently a second sample of the same size in the same fashion from this popu-

lation, we could not expect the distribution of intelligence for the individuals in this second sample to be exactly the same as for those in the first. This, again, is for exactly the same reason that we could not expect two successive bridge hands to show the same distribution of cards in the various suits. Chance would practically *guarantee* that any two successive bridge hands would differ in "value," and in the same way chance would practically guarantee that any two samples drawn from the same population would show some differences in the distributions of measures of any trait. The mean I.Q. obtained from our second sample would almost certainly differ from that obtained from the first sample. This emphasizes the fact that neither of these obtained means may be accepted as exactly descriptive of the whole population.

A third sample, similarly, would probably yield still another value of the mean. If we continued to select, independently and in the same fashion, a very large number of random samples of 81 cases each and recorded the mean I.Q. for each sample, we would find that these means would *distribute* themselves over a considerable range of values. Some samples would by chance contain unusually large proportions of pupils of superior intelligence and would yield relatively high means. Others, by reason of the accidents of sampling, would contain unusually large proportions of dull pupils and would yield relatively low means. We would find, however, that most of these means would cluster around some central value, and that only a relatively small number of the obtained means would deviate far from this value.

The distribution (like that just suggested) of the obtained means of a very large number of random samples of the same size is known as the *sampling distribution* of the mean of a sample of the given size. The *form* of the sampling distribution of the mean of any fairly large random sample will closely approximate that of the normal distribution. This has been shown to be true even though the individual measures in the population involved are not normally distributed (unless the sample is small and the departure from normality is extreme).

The Standard Error of the Mean of a Random Sample

The reliability of the mean of any *single* sample is dependent upon the variability of the sampling distribution of such means. If, in the long run, the means obtained from samples of the given size are distributed over a very wide range of values, we obviously cannot place very much dependence upon the mean obtained from any one sample of that size, because of the possibility that the particular sample considered may be one of those whose means deviate markedly from the true mean. If, on the other hand, the means obtained from a large number of similar samples are in close agreement — that is, if they show only a small variation — then any one of the means can be accepted as a close approximation to, or as a dependable indication of, the true mean. If, then, we could secure a measure of variability for a distribution of the means of a large number of random samples of 81 cases each, we could use this measure of variability to describe the reliability of the mean of any one random sample of 81 cases. The measure of variability used for this purpose is the standard deviation and, when so used, is known as the *standard error. The standard error of the mean of a given random sample is the standard deviation of the distribution of means of a very large number of random samples of the same size as the given sample*, and all, of course, drawn from the same population; that is, the standard error of the mean is the standard deviation of the sampling distribution of the mean. To say that the mean of a given random sample is unreliable is equivalent to saying that the means of other samples of the same size will fluctuate widely in value, which again is equivalent to saying that a distribution of such means would have a large standard deviation, or that the given mean has a large standard error.

Levels of Confidence

In the subsequent discussions it will frequently be desirable to indicate in *quantitative* terms what degree of confidence may be placed in certain inferences drawn from the facts obtained from a random sample. Before proceeding with the interpretation of the

standard error of the mean, therefore, it may be well to introduce and clarify the term *level of confidence.*

The degree or "level" of confidence with which a given assertion may be made may most conveniently be defined in terms of probability. Suppose that 95 of the cards in a given deck of 100 cards are marked in a certain fashion, the five remaining cards being unmarked. Suppose that after this deck has been shuffled thoroughly, we draw from it a single card at random. Since only 5 per cent of the cards are unmarked, we can, before looking at the card drawn, assert with obviously "high" confidence that we have drawn a marked card. The degree or level of confidence with which we can make this particular assertion we will call the *5 per cent level of confidence.* This name is suggested by the fact that if we continued drawing cards in this fashion,[1] each time asserting that a marked card has been drawn, we would in the long run be *wrong* 5 per cent of the time. Whenever we make any assertion — whether or not it has anything to do with cards or chance events — with the *same* degree of confidence with which we asserted in this illustrative situation that a marked card was drawn, we may say that we have made that assertion at the 5 per cent level of confidence. The card illustration, of course, is of no significance in itself, but only offers a convenient way of defining a certain degree of confidence.

Other levels of confidence may be similarly defined. For example, if only 2 per cent of the cards in the deck are unmarked, we can say at the 2 per cent level of confidence that any single card drawn at random from the deck will be marked. Again, if we have drawn a single card from a well-shuffled deck of ordinary playing cards, we may, before looking at the card, be confident at the 1.9 per cent level that the card is something other than the ace of spades (the probability of drawing the ace of spades is $1/52 = .019$). Similarly we may be confident at the $16\frac{2}{3}$ per cent level of confidence that something other than a deuce will be thrown in a single throw of a die. Note that the "per cent" specified is negatively related to the

[1] The card last drawn being replaced and the deck reshuffled before each draw.

degree of confidence involved; that is, a *small* per cent denotes a *high* degree of confidence or a low degree of *un*certainty.

This expression, "level of confidence," can be readily related to the normal distribution. For example, we know that in any normal distribution 99 per cent of the cases lie within 2.58 standard deviations of the mean, or that 1 per cent deviate from the mean by more than that amount. Accordingly, we can make the statement at the 1 per cent level of confidence that any measure drawn at random from a normal distribution will deviate from the mean by less than 2.58 standard deviations. Similarly, if a single measure has been selected at random from a normal distribution, we may be confident at the 5 per cent level that it lies within 1.96 σ of the mean, or that its absolute deviation from the mean does not exceed 1.96 σ ("absolute" meaning that we are concerned only with the size of the deviation, no distinction being made between plus and minus deviations). Similarly, we may be confident at the 2 per cent level that a measure drawn at random from a normal distribution will lie within 2.33 σ of the mean.

It may be well to note again that, while the term, "level of confidence," is most conveniently defined in terms of probability situations, it may be applied to assertions that cannot be directly related to statements of probability, as will be illustrated later in statements about the true mean of a population.

Establishing a "Confidence Interval" for the True Mean

We are now ready to illustrate, in terms of the specific example already employed, how the mean of a sample may be interpreted in relation to its standard error. For the sake of this illustration we will assume that the standard error of the mean of our sample has already been found for us — that someone else has actually taken a very large number of random samples of 81 cases each from our population of rural eighth graders [1] and has found the standard deviation of the distribution of the means of these sam-

[1] This is obviously an impracticable method of finding the standard error. A more practicable method will be suggested later.

ples to be $\sigma_M = 1.2$. Our sample mean, then, is 98.5, and its standard error is 1.2.

We have already noted that the distribution of obtained means for large random samples of any given size is approximately normal. We know, then, that our obtained mean of 98.5 belongs somewhere in a normal sampling distribution whose standard deviation is 1.2 and whose mean is the true mean of the population. Since we do not know the true mean, we cannot say just where in this hypothetical distribution our obtained mean lies. However, we can consider our obtained mean as having been drawn at random from this distribution. Accordingly, we may be "confident at the 1 per cent level" that our obtained mean is within 2.58 σ_M of the true mean; that is, we may be confident at the 1 per cent level that our obtained mean does not differ from the true mean by more than $2.58 \times 1.2 = 3.10$, or that the absolute "sampling error" in the obtained mean does not exceed 3.10. However, the sampling error may be in either direction; hence the true mean may, in the limiting cases, be either 3.10 units higher ($98.5 + 3.10 = 101.60$) or 3.10 units lower ($98.5 - 3.10 = 95.40$) than the obtained mean. We may thus be confident at the 1 per cent level that the true mean lies somewhere within the interval whose limits are 95.40 and 101.60. Similarly, we may be confident at the 2 per cent level that the true mean lies between 95.70 and 101.30, and at the 5 per cent level that it lies in the interval 96.15 to 100.85. In the same fashion, we could, if desired, set the limits corresponding to any other level of confidence, such as the 20 per cent level or the 0.1 per cent level. Any interval thus defined is known as a *confidence interval.* With reference to our sample mean of 98.5, for example, the "2 per cent confidence interval" for the true mean is 95.7–101.3.

The student may well wonder why it is deemed necessary to invent such strange expressions as "the 2 per cent level of confidence" and "the 2 per cent confidence interval" to interpret adequately the obtained mean and its standard error. It would appear much simpler merely to say, "The chances are 98 in 100 that

the true mean lies between 95.7 and 101.3." The latter type of statement is very frequently made; indeed, it is recommended in many introductory statistical textbooks. However, it is illogical, and should therefore be avoided. To say that the "chances" are 98 in 100 that the true mean lies in a certain interval is to imply that the true mean has many values, any of which may be "drawn" in a single instance. Actually, of course, the true mean is a fixed quantity; it does not fluctuate in value from time to time (or from sample to sample) and is not *distributed* either normally or in any other fashion. Statements of probability may properly be applied to randomly distributed measures or events, but not to fixed quantities. It is quite proper to say that the probability is .02 (or that the chances are 2 in 100) that the *obtained* mean of a random sample will lie more than a given distance from the true mean; we may not, however, invert the statement, that is, we may not properly say that the probability is .02 that the *true* mean deviates more than a given distance from a particular obtained mean. However, we may avoid any inconsistency by saying that we have a certain "degree of confidence" that the true mean lies within a given interval, and accordingly this is now the approved practice.

Testing an Exact Hypothesis about the True Mean

Very frequently, in situations like the one we have been considering, we may be especially interested in the possibility that the true mean has some particular exact value. In this case, for instance, we may be interested in the possibility that the mean I.Q. for the population of rural eighth grade pupils is 100 (the "norm" for the population at large). Indeed, the whole purpose of drawing the sample may have been to see if there is any evidence that rural pupils are *not* "up to the norm" of intelligence. Accordingly, in interpreting our results we might ask, "Is it reasonable, in view of what is known of our sample, to suppose that the true mean is 100?" or "Is the hypothesis tenable that the true mean is 100?" Again, recognizing that the tenability of any hypothesis is a matter of degree rather than an all-or-none proposition, we may ask, "*How*

reasonable is it to suppose that the true mean is 100?" or, inversely, "With what degree of confidence may we *reject* the hypothesis that the true mean of the rural eighth grade population is 100?"

To answer these questions, we observe that *if* the true mean *is* 100, then our obtained mean contains a sampling error of $100 - 98.5 = 1.5$. To ask, "How reasonable is it to suppose that the true mean is 100?" is therefore equivalent to asking, "How reasonable is it to suppose that the sampling error in this mean is as large as 1.5?" Since the standard error of the mean is 1.2, this hypothetical sampling error is $1.5/1.2 = 1.25$ times as large as the standard error. According to Table 17, sampling errors this large would be exceeded (in absolute magnitude) 21.12 per cent of the time. Hence, if we are to retain the hypothesis that the true mean is 100, we must accept the notion that something has happened in our one sample that (under this hypothesis) would happen in the long run only about once in five times. Since this notion is hardly to be considered as unreasonable, we conclude that the hypothesis is *tenable*. In other words, the hypothesis that the true mean is 100 is reasonably consistent with the known facts that our sample mean is 98.5 and its standard error is 1.2. While we have thus shown that the hypothesis is tenable or reasonably consistent with what was found in our sample, we have by no means *proved* that it is *true*. There are many other tenable hypotheses. The hypothesis that the true mean is 97, for instance, is equally tenable, while the hypothesis that the true mean is, say, 99.3, is even more readily accepted.

On the other hand, there are many hypotheses that we would be forced to reject in view of what we know of our one sample. Suppose, for instance, that someone suggests that the true mean of our rural eighth grade population is 95.5. If this is the true mean, then the sampling error in our obtained mean of 98.5 is 3.0, which is 2.5 times the standard error. According to Table 17, sampling errors this large would be found only 1.2 per cent of the time in the means of random samples of this size. Accordingly, to accept the hypoth-

esis that the true mean is 95.5, we must also accept the notion that something has happened in our one sample that happens only once in 100 times by chance alone. If we are unwilling, as most people would be, to believe that anything so highly improbable has actually occurred in this particular case, our only choice is to reject the hypothesis. In this case we can do so with a confidence at the 1.2 per cent level that the hypothesis is false.[1] Similarly, the hypothesis that the true mean is 96.0 may be rejected at the 3.76 per cent level of confidence.

In general, the principal steps in testing any exact hypothesis about a population, given the appropriate facts for a sample drawn from that population, are as follows:

(1) We note the discrepancy between fact and hypothesis, that is, we determine the difference between the hypothetical true

[1] Some statisticians would prefer to say that this hypothesis may be rejected at the 0.6 per cent level of confidence. They would reason that if the true mean were 95.5, then means *as high as or higher than 98.5* would be found in 0.6 per cent of all random samples of this size. (This reasoning is perhaps more consistent with the hypothesis that the true mean *is below* 95.5, which in certain respects is an indefinite or inexact hypothesis, as compared to the exact hypothesis that the true mean *is* 95.5.) They would thus take the *direction* as well as the magnitude of the hypothetical sampling error into consideration. In situations like that here illustrated, it matters little which interpretation is employed, so long as one understands clearly which definition of *level of confidence* is implied. For instance, suppose we define the level of confidence at which an exact hypothesis may be rejected in terms of the per cent of samples in which the observed discrepancy from the hypothesis would be exceeded in *absolute* magnitude (without regard to sign) if the hypothesis were true. It would then follow that this particular hypothesis may be rejected at the 1.2 per cent level. On the other hand, suppose that, with equal arbitrariness, we define the level of confidence with which we may reject the hypothesis in terms of the per cent of the time that the observed discrepancy from the hypothesis would be exceeded by other sampling errors *in the same direction* if the hypothesis were true. It would then follow that this particular hypothesis may be rejected at the 0.6 per cent level. Accordingly, when we test any exact hypothesis about the true mean, the *practical* result will be *exactly* the same under either definition, so long as we employ comparable standards. If we employ the 5 per cent level of confidence as a standard under the first definition, we should have to employ the 2.5 per cent standard under the second, the 1 per cent level under the first would be comparable to the 0.5 per cent level under the second, etc.

While the second definition may appear more fitting in situations like that here illustrated, the first definition has a decided advantage in certain tests of the "null" hypothesis about a true difference. (This advantage is explained in the footnote on page 138 following.) Because of this advantage, and because the practical result is the same in any situation, it seems desirable to use the first definition consistently in testing any exact hypothesis, and students in this course are accordingly advised to follow this procedure.

measure and the measure obtained from the sample. This difference is the *hypothetical* sampling error in the obtained measure.

(2) We determine the relative frequency with which this hypothetical sampling error would be exceeded in absolute magnitude in other similar samples *if the hypothesis were true.* (This requires that we know the sampling distribution of the obtained measure.)

(3) We may then either accept or reject the hypothesis, depending upon this relative frequency. If the relative frequency is small, we have the alternatives:

(a) of rejecting the hypothesis, maintaining that it is unreasonable to suppose that something has happened in our one sample that would happen only very infrequently if the hypothesis were true;

(b) of accepting the hypothesis, maintaining that it is reasonable to suppose that something *has* happened in our one sample that only rarely does happen by chance.

If the relative frequency is very small (say less than 2 or 1 per cent), we would ordinarily prefer the first alternative, being unwilling to accept the notion that a very rare event has actually "come off" in our one sample. However, if the relative frequency is large (say more than 5 per cent), we might admit that the hypothesis is still tenable or has not been disproved, since it is not unreasonable to suppose that something has happened in our one sample that does happen by chance at least once in twenty times.

The level of confidence at which we may reject the hypothesis, then, depends (by definition) upon the relative frequency with which the hypothetical sampling error would be exceeded in *absolute* magnitude (without regard to direction) if the hypothesis were true. If this relative frequency is 2 per cent, we may reject the hypothesis at the 2 per cent level of confidence, etc. Ordinarily, before we would *categorically* "reject" the hypothesis at all, we would require

that we be able to do so at least at the 5 per cent level, and sometimes we would "retain" the hypothesis categorically unless able to reject it at least at the 1 per cent level.

The student should have no trouble seeing how this generalized procedure applies in the specific example used. We first determined the difference between our hypothetical true mean (100) and our obtained mean (98.5). This hypothetical sampling error was $100 - 98.5 = 1.5$. We then observed that, since such sampling errors are normally distributed with a standard deviation of 1.2, it follows that sampling errors of 1.5 would be exceeded 21 per cent of the time in random samples of this size. Accordingly, we did not feel that a categorical rejection of the hypothesis was justified.

In the preceding examples the levels of confidence involved were described with greater accuracy than is needed for most practical purposes. Ordinarily, instead of describing the level so accurately, we would simply take the nearest lower [1] level in which the per cent is some convenient integer. In the examples last considered, for instance, we would ordinarily say that the hypothesis that the true mean is 95.5 may be rejected at the 2 per cent level, or that the hypothesis that the true mean is 96.0 may be rejected at the 5 per cent level. The 5 per cent, 2 per cent and 1 per cent levels are most often used in this way. If any hypothesis may be rejected at or beyond the 1 per cent level, we often say that the hypothesis is "practically certain" to be false, and we usually are not interested in discriminating between various degrees of "practical certainty." Similarly, if an hypothesis may not be rejected with at least as much confidence as is implied in the "5 per cent level," we usually would not consider rejecting it at all and hence would not be interested in discriminating between lower levels of confidence. However, the 0.1 per cent and 20 per cent levels of confidence are sometimes employed, and special provision is made for them in some probability tables. Thus, all that one needs from Table 17 for most practical purposes is the knowledge that 20

[1] The degree of confidence is lower; the numerical value of the *per cent* used to identify it is higher.

per cent of the cases in a normal distribution deviate from the mean by 1.28 σ or more, 5 per cent deviate by at least 1.96 σ, 2 per cent by 2.33 σ, 1 per cent by 2.58 σ, and 0.1 per cent by 3.3 σ.

It should now be apparent that one's decision to "reject" or "accept" an hypothesis categorically depends somewhat upon his temperament and upon the practical implications of his decision. In some instances one might be unwilling to reject an hypothesis finally and categorically even though confident at the 1 per cent level that it is false. In other instances one might reject the hypothesis though not confident even at the 5 per cent level that it is false. In general, in educational and psychological research, one does not reject an hypothesis unless he is confident at least at the 2 per cent level, or more often at the 1 per cent level, that it is a false hypothesis.[1] However, it is dangerous to recommend any single standard practice — the selection of the critical level is a matter which must be subjectively decided anew by the investigator in each independent application in terms of the peculiar nature of the situation involved.

It may be noted that the establishment of a *confidence interval* for the true mean at any chosen level of confidence consists of selecting, in turn, the lowest and highest values of the mean which constitute tenable hypotheses at that level. For instance, the hypothesis that the true mean of our eighth grade population is 95.7 may barely be rejected at the 2 per cent level, as may the hypothesis that it is 101.3. Accordingly, *any* hypothesis that it lies *outside* the interval 95.7–101.3 may be rejected at least at the 2 per cent level of confidence, that is, 95.7–101.3 is the 2 per cent confidence interval.

The Formula for Estimating the Standard Error of the Mean

We have already noted that none of the uses of the standard error of the mean that have just been considered would be practicable if

[1] It has been very frequent practice in these fields, in the past, to demand that the discrepancy between the hypothesis and the observed value be at least three times the standard error. This is equivalent to employing the 0.26 per cent level of confidence, which is considerably higher than should ordinarily be necessary.

we had to determine the value of the standard error in the direct manner suggested by the definition on page 106. Fortunately, however, we can derive a usefully accurate *estimate* of the standard error of the mean of a certain sample even though we know only the facts for that one sample. This is because it can be shown, either empirically or by mathematical derivation, that the variability of the means of a large number of random samples of the same size depends upon (1) the number of cases, N, in each sample, and (2) the S.D. of the individual measures for the whole population. This relationship is indicated by the formula:

$$\sigma_M = \frac{\sigma_{(pop.)}}{\sqrt{N}}. \tag{7}$$

This is a relationship which most students will have to accept on faith,[1] but its *reasonableness* may become apparent upon consideration of the following illustration.

Suppose that it is known that the S.D. of individual I.Q.'s for the whole population just considered is 10, and that these I.Q.'s are normally distributed, as is indicated in the upper curve in Figure 14 on page 117. Now let us suppose that we select from this population a large number of samples, each consisting of only *one* pupil selected at random from the whole population. The "mean" of each of these samples would then be the same as the I.Q. of the one pupil in the sample; hence, a distribution of the means of a very large number of such samples would show the same variability as the individual I.Q.'s for the whole population. The standard deviation of the distribution of means for samples of one pupil each would then be $\sigma_M = 10.0$, as is shown by the second curve in Figure 14. Next let us suppose that we have selected another large number of samples, each sample this time consisting of only *two* pupils drawn at random from the whole population. It should now be apparent that the *means* of these samples would show less variability than the individual I.Q.'s for the whole population, or than

[1] The derivation of this formula is relatively simple for anyone adept in algebra. See Kelley, T. L., *Statistical Method*, pp. 82–83. The Macmillan Company, 1923.

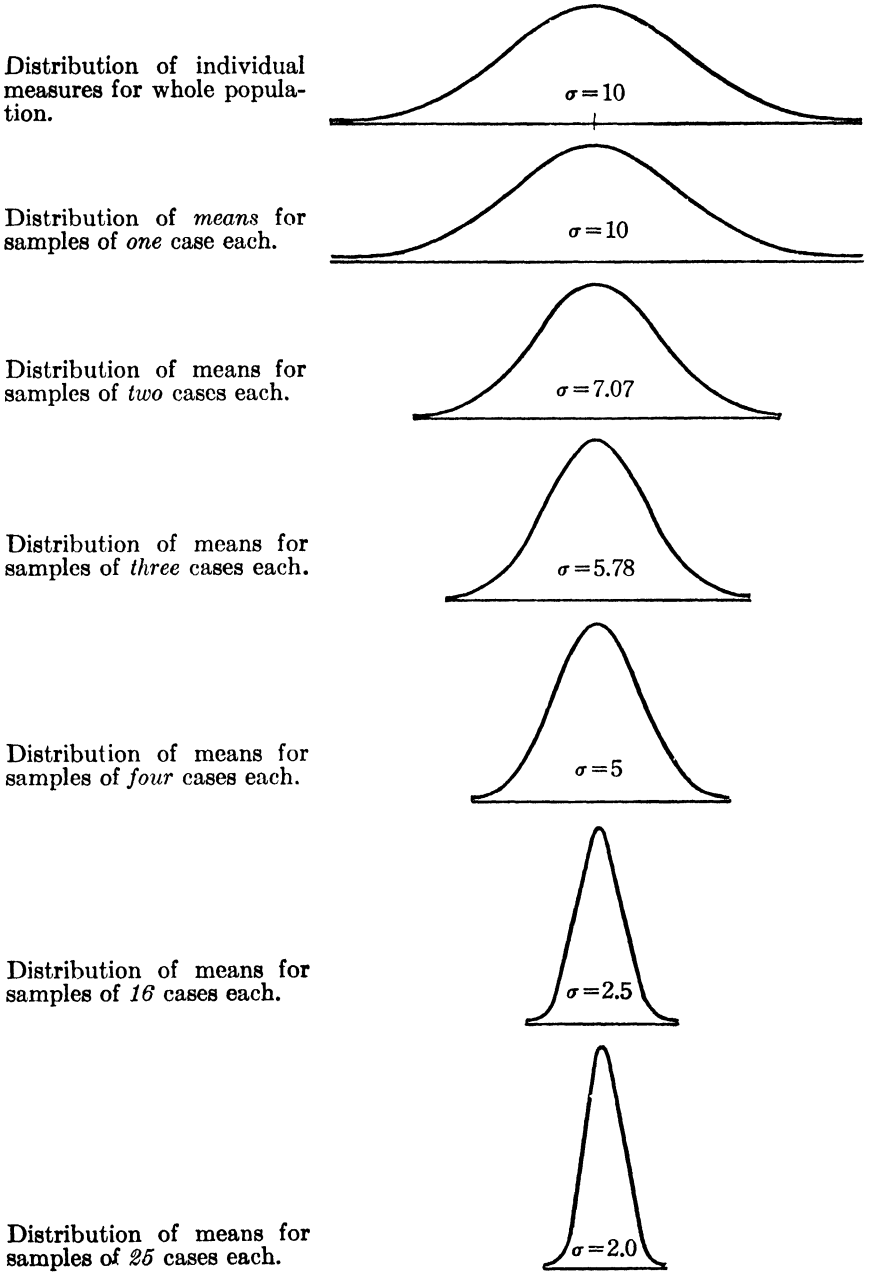


FIG. 14.

Relation of standard error of the mean to the size of a random sample.

the means of samples of one case each. This follows from the fact that while any sample of two cases may contain one individual drawn from either extreme of the distribution, it is most unlikely that *both* individuals in the same sample will deviate equally far and in the same direction from the general average. One of the two individuals drawn will almost invariably have a higher I.Q. than the other, and the mean of their two I.Q.'s will lie closer to the general average than does the I.Q. of the more extreme individual in the pair. Hence, it seems reasonable that the distribution of means of samples of two cases each should be pictured with a narrower spread than the distribution of individual I.Q.'s for the whole population, as has been done in Figure 14.

Now suppose that we select a very large number of samples of *three* cases each. The probability is now very much reduced that all of the individuals in any one sample will deviate by a large amount and in the same direction from the general average. The probability of drawing three very bright or three very dull pupils in a single sample of three cases is surely less than the probability of drawing two very bright or two very dull pupils in a single sample of two cases. Again, therefore, it seems reasonable to picture the distribution of means of samples of three cases each with a narrower spread than the distribution of means of samples of two cases each.

For similar reasons, the means of a large number of samples of four cases each would show less variability than the distribution of means of samples of three cases each. Similarly, the distribution of means of samples of any given size would show less variability than the means of samples of any smaller size.

Actual trials involving very large numbers of samples have shown that the variability of the means for samples of *any* given size is inversely proportional to the square root of the number of cases in each sample. The means of samples of four cases each, for instance, would show one half as much variability as the means of samples of one case each. The means of samples of 16 cases each would show one half as much variability as the means of samples

of four cases each, or one fourth as much as the means of samples of one case each. The means of samples of 25 cases each would be one fifth as variable as the means of samples of one case each, etc. This relationship is stated in a more general form by Formula (7).

We have now seen that, by means of Formula (7), we can state immediately the reliability of the mean of any random sample if we know the standard deviation of individual measures in the whole population and the number of cases in the sample. It may nevertheless appear, upon first consideration, that this formula can have very little practical value, since it would be just as impracticable for us to determine the standard deviation of a whole population as to select a very large number of random samples and determine the standard error of the mean empirically in the manner suggested on page 108. In actual practice we draw only one sample and must reason as best we can from only the facts for that sample.

In the practical situation, then, we must substitute for the unknown $\sigma_{(pop.)}$ in Formula (7) some estimate of it which may be derived from our sample. The obtained standard deviation of the sample is not a good estimate (particularly for small samples), since it tends to be smaller than the standard deviation of the population. However, it may be shown [1] that $\Sigma d^2/(N-1)$, in which d is a deviation from the sample mean and N is the number of cases in the sample, is an unbiased estimate of the variance (σ^2) of the population.[2] This may be expressed in terms of the standard deviation of the sample, as follows:

$$\text{est'd } \sigma^2_{(pop.)} = \frac{\Sigma d^2}{N-1} = \frac{\Sigma d^2}{N} \cdot \frac{N}{N-1} = \sigma^2_{(sample)}\left(\frac{N}{N-1}\right),$$

from which we secure

$$\text{est'd } \sigma_{(pop.)} = \sigma_{(sample)} \sqrt{\frac{N}{N-1}}.$$

[1] The proof of this is well within the understanding of any student capable of following relatively simple algebraic manipulations. See Lindquist, E. F., *Statistical Analysis in Educational Research*, pp. 48–50. Houghton Mifflin Company, 1940.

[2] The variance of any distribution is the square of its standard deviation.

If we now substitute this estimate of $\sigma_{(pop.)}$ for the actual $\sigma_{(pop.)}$ in Formula (7), we secure

$$\text{est'd } \sigma_M = \frac{\text{est'd } \sigma_{(pop.)}}{\sqrt{N}} = \frac{\sigma_{(sample)}\sqrt{\frac{N}{N-1}}}{\sqrt{N}} = \frac{\sigma_{(sample)}}{\sqrt{N-1}}.$$

The *working* formula for the standard error of the mean of a random sample is then

$$\sigma_M = \frac{\sigma_{(sample)}}{\sqrt{N-1}}. \tag{8}$$

Since when the sample is large $\sqrt{N-1}$ will not differ appreciably from $\sqrt{N}$, it has been rather general practice in the past to use the somewhat simpler expression

$$\sigma_M = \frac{\sigma_{(sample)}}{\sqrt{N}}$$

as the formula for the standard error of the mean of a large sample. However, very little is gained by introducing this inaccuracy, and the student is therefore advised to use the correct Formula (8), no matter how large the sample may be.

The Use of the Standard Error of the Mean with Large Samples

We are now ready to consider in terms of a fresh illustration how Formula (8) may be applied. Suppose we wish to know the mean weight of all ten-year-old boys in the state of Iowa. Let us suppose that we have selected a random sample of 50 boys from this population, and that we have found their mean weight to be 77 pounds and the S.D. of their weights to be 9.8 pounds. We then reason that if we were to continue drawing random samples of this size from this population until we had a very large number of them and were to construct a frequency distribution of the means of these samples, we would find that these means would be normally dis-

tributed and, according to Formula (8), that the S.D. of this distribution would be approximately

$$\sigma_M = \frac{\sigma}{\sqrt{N-1}} = \frac{9.8}{\sqrt{49}} = 1.4 \text{ pounds.}$$

Now we know that our obtained mean of 77 belongs somewhere in this hypothetical normal distribution of means. If our sample happens to contain an unusually large proportion of boys who are heavier than most boys of their age, our mean of 77 might be near the upper extreme of the sampling distribution, as shown in

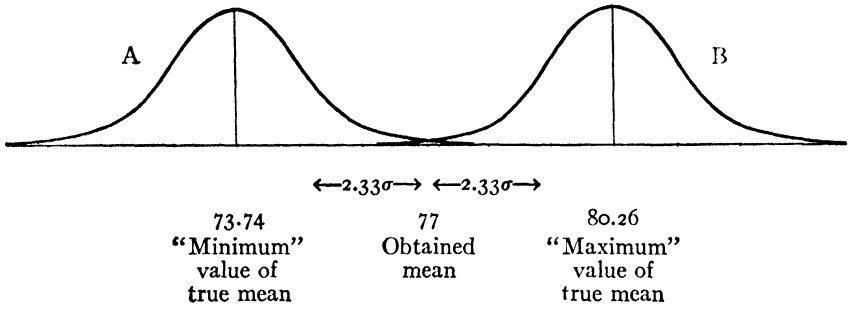


FIG. 15.
Illustrating the 2 per cent confidence interval for the true mean.

Figure 15, Curve *A*, above. On the other hand, our sample may be one which is accidentally very heavily loaded with light-weight boys of this age, in which case our obtained mean of 77 may be near the lower extreme of the sampling distribution, as shown in Curve *B* of Figure 15. Suppose, then, that we allow for the possibility that the obtained mean deviates from the true mean by an amount which would be exceeded by chance only twice in 100 times, that is, suppose we allow for the possibility that the obtained mean deviates 2.33 σ from the true mean (in either direction). If the obtained mean is 2.33 σ *above* the true mean, then the true mean is as low as 73.74 pounds. If the obtained mean is

2.33 σ *below* the true mean, then the latter is 80.26 pounds. We may thus be quite confident (2 per cent level) that the true mean lies between 73.7 and 80.3 pounds. If we prefer an interval that we may be even more highly confident contains the true mean, we may employ the 1 per cent confidence interval (73.4–80.6), or even the 0.1 per cent interval, which in this case is $77 \pm 3.3\ \sigma_M$ (or 72.4–81.6). On the other hand, if we are satisfied with a lower degree of confidence, we may employ the 5 per cent interval (74.3–79.7).

We may use this same situation to illustrate the testing of an exact hypothesis. For the sake of this illustration, let us assume now that no confidence interval for the true mean has yet been established, but that the standard error of the mean has been estimated at 1.4. Let us suppose also that it is known that the mean weight for boys of this age in the country at large is 75 pounds, and that we are therefore particularly interested in the hypothesis that this is also the true mean for Iowa boys of this age. To test this hypothesis, we reason that if it is true then our obtained mean of 77 pounds contains a sampling error of 2 pounds, which is $2/1.4 = 1.43$ times the standard error of the mean. We know, however, that sampling errors larger than this occur 15.2 per cent of the time by chance alone. While we may therefore be confident at the 15 per cent level that the hypothesis is false, this is hardly a sufficient degree of confidence to justify a categorical rejection of the hypothesis. In other words, we would admit that the hypothesis is tenable, or that it is reasonably consistent with what we know about our one sample. Had we already determined a confidence interval (conforming to whatever level of confidence we had decided to employ), we would, of course, not need to "test" the hypothesis (that the true mean is 75) in this fashion, but would only have to note whether or not the interval includes 75.

It is very important to note that the procedures just illustrated in establishing a confidence interval for, or in testing an exact hypothesis about, the true mean are not valid for small samples.

The appropriate corresponding procedures for small samples will be presented later (pages 136–139).

The Probable Error of the Mean

We have already seen (page 92) that the *probable deviation* of a randomly selected measure from the mean of a normal distribution is .6745 times the standard deviation of the distribution. The probable deviation of an obtained mean from the true mean is called the *Probable Error* (P.E.) of the mean. It may be conveniently defined as the sampling error in the mean which is exceeded half of the time, or for which the chances are even that it will be exceeded in any individual instance. Perhaps because it may be so neatly defined and presumably may therefore be more readily understood, the probable error has often been used in preference to the standard error to describe the reliability of the mean. Its formula is

$$\text{P.E.}_M = \frac{.6745\ \sigma_{(sample)}}{\sqrt{N-1}} = .6745\ \sigma_M. \qquad (9)$$

Tables of area relationships under the normal curve, based upon deviations from the mean in P.E. rather than sigma units, may be found in many statistics references. One such table is given on page 229 of the Appendix. This table may be used in very much the same way as Table 17. For example, if an obtained mean of 26.0 has a probable error of 4.0, we may be confident at the 2 per cent level [1] that the true mean lies within $3.45 \times 4.0 = 13.80$ units of the obtained mean, or in the interval 12.2–39.8.

When the P.E. is used to measure the reliability of the mean, it is customarily written immediately following the obtained mean, with a "plus or minus" sign between. For example, the statement that an obtained mean is 77 ± 2 would indicate that the probable error

[1] In the table on page 229 we see that 98 per cent of the cases lie within 3.45 P.E. of the mean (49 per cent on either side).

of the obtained mean is 2. The standard error is rarely indicated in this fashion.

While the student must become familiar with the P.E.$_M$ in order to read the research literature in which it is employed, he is advised in his own work to avoid unnecessary arithmetic by using only the standard error.

The Standard Errors of the Median, Q, and S.D.

The standard error of any statistical measure obtained from a random sample is the standard deviation of its sampling distribution, that is, it is the standard deviation of the distribution of such measures obtained from a very large number of samples of the given size. Accordingly, the *standard error of the median* is the standard deviation of a distribution of *medians* for a very large number of random samples of the same size as the given sample. The formula for the standard error of the median is as follows:

$$\text{est'd } \sigma_{\text{mdn}} = \frac{5}{4} \cdot \frac{\sigma_{(sample)}}{\sqrt{N-1}} = \frac{5}{4}\sigma_M. \qquad (10)$$

As the formula indicates, medians are somewhat less reliable than the means of the same samples.

A complete discussion of the logic underlying this formula and of its use and interpretation would parallel exactly that already given for the standard error of the mean and could be derived from the preceding discussions by simply substituting "obtained median" for "obtained mean" and "true median" for "true mean" wherever these terms are found. Formulas for computing approximate values of the standard errors of the standard deviation and the semi-interquartile range of a sample are given below:

$$\sigma_Q = \frac{.787\ \sigma_{(sample)}}{\sqrt{N-1}} = .787\ \sigma_M. \qquad (11)$$

$$\sigma_{\text{S.D.}} \text{ (or } \sigma_\sigma) = \frac{\sigma_{(sample)}}{\sqrt{2(N-1)}} = \frac{1}{\sqrt{2}} \cdot \frac{\sigma_{(sample)}}{\sqrt{N-1}} = .707\ \sigma_M. \qquad (12)$$

If the samples involved are large and drawn at random from approximately normal populations, these formulas may for most practical purposes be used and interpreted in essentially the same fashion as Formulas (8) and (10).

It will be noted that, once the standard error of the mean has been computed for the sample, the approximate standard errors of any other of these measures may be readily determined by simply multiplying by a constant.

The probable error of any measure may be found by multiplying its standard error by .6745.

In using Table 17 with any of these formulas it is assumed, of course, that the *sampling errors* are normally distributed, that is, that similar measures from a large number of random samples of the same size will form a normal distribution. If the samples are large, this assumption is likely to be sufficiently well satisfied even though the population involved is not normal. For small samples, the assumption of normality of the sampling distribution of the standard deviation or of the semi-interquartile range is definitely not satisfied, even though the population is normal. None of these formulas, particularly Formulas (11) and (12), should be employed with small samples ($N < 25$).

An exact test for the significance of a difference between the standard deviations of small samples is available,[1] but is beyond the scope of this course.

The Standard Errors of Proportions and Percentages

Very frequently we are interested in obtaining only a simple statement of the proportion (decimal fraction) or percentage of individuals in a total population that belong to a specified category. For example, we might wish to know the proportion or per cent of left-handed children among all school children in the public elementary schools of the country. To determine this proportion or percentage, we might resort to random sampling. Because of the part played by chance in determining which individuals are to

[1] See Lindquist, *Statistical Analysis in Educational Research*, pp. 60–66.

constitute the sample, we could not expect the proportion of left-handed children in our one sample to correspond exactly to the true proportion in the entire population. If we continued drawing other random samples of the same size, each sample would perhaps contain a slightly different proportion of left-handed pupils than any other. If the samples were large, these proportions (unless the "true" proportion were near zero or 1) would fall into an approximately normal distribution, the standard deviation of which would be approximately

$$\sigma_p = \sqrt{\frac{pq}{N}} = \sqrt{\frac{p(1-p)}{N}}, \tag{13}$$

in which p represents the proportion in the given category in the entire population (the "true" proportion) and $q = 1 - p$ (q is the true proportion in the remaining categories).

In practical situations, of course, the true proportions are always unknown. However, if we know the observed proportion for a single random sample, we may test any exact hypothesis concerning the true proportion by substituting the *hypothetical true proportion* (not the observed proportion) for the p in the formula. The result will be the *true* standard error of the observed proportion under the hypothesis that is being tested. Suppose, for instance, that we have found that 14 children in a sample of 100 (.14 of the sample) are left-handed, and that we wish to test the hypothesis that the true proportion is .10. Under this hypothesis, the standard error of the obtained proportion for a sample of 100 cases is

$$\sigma_p = \sqrt{\frac{.10 \times .90}{100}} = .03,$$

that is, .03 is the standard error of the obtained proportion *if* the true proportion is .10. Our obtained proportion differs from the hypothetical proportion by .04, which is only slightly more than one standard error. This discrepancy could be readily attributed to chance; hence, we are in no position to reject the hypothesis with any high degree of confidence.

The establishment of a confidence interval for the true percentage presents a more complex problem than the establishment of a confidence interval for the true mean. Our estimate of the standard error of the mean of a sample is based on the standard deviation of the sample and is the same no matter what hypothesis about the true mean we wish to test. The standard error of the obtained proportion, however, depends upon the true proportion, and hence, we must use different values of the standard error to test different hypotheses about the true proportion. How this affects the establishment of a confidence interval may best be clarified by an example.

Suppose, in the illustrative situation just considered, we tried incorrectly to establish the 2 per cent confidence interval for the true proportion by following a procedure suggested by that described on page 109. That is, suppose we estimated "the" standard error of the obtained proportion to be

$$\sqrt{\frac{.14 \times .86}{100}} = .034$$

(substituting the obtained proportion for the true proportion in the formula), and then established $.14 \pm (2.33 \times .034)$ or .061 and .219 as the limits of the 2 per cent confidence interval. That these limits are incorrect may be readily demonstrated by testing each separately in the manner described in the last paragraph on the preceding page. When this is done, we find that the hypothesis that the true proportion is .061 may be rejected at a level of confidence far beyond the 1 per cent level, rather than only at the 2 per cent level. (Under the hypothesis that the true proportion is .061, the standard error of the obtained proportion is .024. The discrepancy of $.14 - .061 = .079$ is thus over 3.2 times the standard error — a discrepancy which would occur much less than 1 per cent of the time by chance alone.)

On the other hand, the hypothesis that the true proportion is .219 may barely be rejected at the 10 per cent level, rather than at the 2 per cent level of confidence. The limits .061 and .219 are

thus incorrect because they were established by assuming that "the" standard error of the obtained proportion was .034 regardless of the hypothesis being tested, instead of recognizing that the standard error to use depends on the hypothesis to be tested.

The establishment of an exact confidence interval for the true proportion is a relatively involved process, but one which should not be beyond the typical student in this course. We have already seen that to establish a certain confidence interval is the same as to determine the "limiting" acceptable hypotheses corresponding to the selected level of confidence. For instance, to establish the 2 per cent confidence interval for the true proportion, we must find the highest and lowest hypothetical values of the true proportion which are "acceptable" each at the 2 per cent level. Since a discrepancy (between observation and hypothesis) of 2.33 standard errors will occur 2 per cent of the time by chance, we wish (in the illustrative situation already used) to know for what values (X) of the hypothetical true proportion $(X - .14)/\sigma_p$ equals $\pm$ 2.33. That is, we wish to know for what values of X the following equality holds.

$$\frac{X - .14}{\sigma_p} = \frac{X - .14}{\sqrt{\frac{X(1 - X)}{100}}} = \pm 2.33.$$

Accordingly, we must solve for X in the equation

$$X - .14 = \pm 2.33 \sqrt{\frac{X(1 - X)}{100}},$$

which reduces to

$$105.43\, X^2 - 33.43\, X + 1.96 = 0.$$

This involves the solution of a quadratic equation, for which the student may need to refer to an elementary algebra text.[1] In this

[1] The roots of the equation $ax^2 + bx + c$ are given by the formula:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}.$$

case the roots of the equation are .077 and .239. These values, accordingly, are the exact limits of the 2 per cent confidence interval for the true proportion, rather than .061 and .219 as determined by the inexact procedure earlier described. The procedure last described, incidentally, is valid only for large samples.

It may be noted that when the obtained proportion is near .5, it may be satisfactory for most practical purposes to follow the incorrect but much simpler procedure described at the middle of page 127, that is, to consider the standard error of the obtained proportion (secured by substituting the obtained proportion in the formula) as the same for any hypothesis to be tested. This procedure may not be used, however, when the obtained proportion differs markedly from .5.

The formula for the standard error of an obtained *percentage* is

$$\sigma_{\%} = \sqrt{\frac{X(100 - X)}{N}}, \qquad (14)$$

in which X is the true percentage (or the hypothetical true percentage). The use of Formula (14) is similar to that of Formula (13).

The Standard Error of a Difference

One of the most important and most frequently used of all sampling error formulas is that for the standard error of a difference. A considerable proportion of all sampling studies involve a *comparison* between measures obtained from random samples drawn from each of two populations. For example, we might wish to compare the *mean* intelligence of rural school children with that of city school children, or might wish to determine whether or not there is any difference in *variability* (S.D.) in intelligence between the two sexes, or might wish to find if there is any difference in the *percentages* of left-handed boys and left-handed girls of the same age.

The standard error of any obtained difference is the S.D. of a distribution of such differences for a large number of pairs of random samples independently drawn from the same populations

The general formula for the standard error of a difference $(X - Y)$ between uncorrelated measures is

$$\sigma_{(X-Y)} = \sqrt{\sigma_X^2 + \sigma_Y^2}, \tag{15}$$

in which X and Y represent the two measures, and σ_X and σ_Y represent their standard errors. This formula, as applied to a difference between the means of two independent random samples, becomes

$$\sigma_{(M_1-M_2)} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2}. \tag{16}$$

The standard error of a difference between any other measures derived from each of two independent random samples could be similarly found by substituting the standard errors of each of the measures in Formula (15).

The "Significance" of a Difference; Testing the Null Hypothesis

The use of the formula for the standard error of a difference involves essentially the same logic as has already been explained in connection with the standard error of the mean. However, in interpreting differences we are less often concerned with establishing confidence intervals and more often concerned with testing certain exact hypotheses. In particular, we are very often uniquely interested in testing the hypothesis that the two populations sampled are *alike* in the trait measured, or that the true difference is zero. This hypothesis (that the true difference is zero) is known as the "null" hypothesis.

When the null hypothesis may be rejected at a high level of confidence, we say that the difference is "statistically significant." Frequently, we qualify such statements, saying, for example, that a difference is "significant at the 5 per cent level" (meaning that the null hypothesis may be rejected at the 5 per cent level) or that it is "significant at the 1 per cent level" (meaning that we are confident at the 1 per cent level that the null hypothesis is false). When we say that a difference is significant, we mean that it is too large to be reasonably attributed to chance (sampling error) alone,

and that we are highly confident (or "practically certain") that the two populations *differ* in the trait measured.

Suppose, for example, that we wish to determine by sampling whether or not there is any difference in the mean weights of 10-year-old boys and 10-year-old girls in the public schools of this country. Suppose that we have selected a random sample of 226 cases from the population of boys and of 145 cases from the population of girls,[1] that we have found the mean weight for the sample of boys to be 77 pounds and that for the girls to be 75 pounds, and that the standard deviation of weights is 12 pounds for the boys and 13 pounds for the girls.

We now wish to know whether or not it is reasonable to suppose that the difference of 2 pounds in the obtained means is due entirely to chance, and that the *true* difference in mean weights is zero. We first compute the standard errors of the obtained means. According to Formula (8), the standard error of the obtained means for the boys is .8 pounds, while that for the girls is 1.08 pounds. Hence, according to Formula (16), the estimated standard error of the difference is

$$\sigma_{(M_B - M_G)} = \sqrt{\sigma_{M_B}^2 + \sigma_{M_G}^2} = \sqrt{.8^2 + 1.08^2} = 1.3 \text{ (rounded)}.$$

We may interpret this standard error in the same way that we have previously interpreted the other standard errors. In this case, our reasoning would be that if we continued drawing other random samples of 226 cases each from the population of boys and other random samples of 145 cases each from the population of girls, and that if we paired these samples at random and found the difference in the means of each pair, these differences would fall into a normal distribution, the S.D. of which would be 1.3 pounds.

Hence we know that if the true difference were zero, obtained differences as large as that found (2 pounds) in this pair of samples would be exceeded approximately 12.33 per cent of the time.[2]

[1] Ordinarily, we would select samples of the same size from each population, but different numbers are employed here to make the illustration more general.

[2] 87.66 per cent of the cases in a normal distribution would lie within $2/1.3 = 1.54\,\sigma$ of the mean (43.83 per cent on either side); hence, $100 - 87.66 = 12.33$ per cent would differ from the mean by more than $1.54\,\sigma$.

Accordingly, it is quite reasonable to suppose that the true difference *is* zero, and that our one pair of samples is one of the 12 pairs of samples in 100 that would then yield differences of at least 2 pounds. We would therefore say that our observed difference is lacking in statistical significance, meaning that it does not signify dependably that there is *any* difference in the means of the populations sampled.

The ratio between an obtained difference and its estimated standard error is often referred to as the "significance ratio." In the preceding example, for instance, the significance ratio was 1.54. To enable us to reject the null hypothesis at the 5 per cent level, the significance ratio must exceed 1.96; at the 2 per cent level it must exceed 2.33, etc. The "critical value" which the significance ratio must exceed in order that we may declare the difference "significant" depends upon the level of confidence that we choose to employ, and this in turn depends upon our temperament and other considerations. Educational and psychological research workers have in the past frequently followed the practice of requiring that the significance ratio exceed 3 before declaring a difference significant, that is, they have insisted on a very high degree of confidence (0.26 per cent level) that the null hypothesis is false. More recent practice is to utilize the 1 per cent or 2 per cent levels, with 2.58 and 2.33 as the corresponding "critical" values of the significance ratio.

It should be noted that a statistically significant difference is not necessarily a reliable difference. An obtained difference is said to be reliable to the degree that it is likely to approximate the corresponding true difference; that is, the *reliability* of an obtained difference is dependent only upon its standard error and is independent of the *magnitude* of the obtained difference or of the ratio between the difference and its standard error. A difference is said to be statistically significant if it may not reasonably be accounted for entirely in terms of chance fluctuations in random sampling. Since the significance of a difference depends upon its significance ratio, whether or not it is significant depends *both* upon its magni-

tude and upon its standard error. This means that a difference may be relatively unreliable (that is, have a large standard error), and yet be "significant," if the difference itself is sufficiently large. Again, if the obtained difference is small, it may fail to be significant even though it is very highly reliable, that is, even though its standard error is extremely small.

It is also very important to note that the fact that an obtained difference is statistically significant indicates only that the obtained difference is not entirely due to *chance* fluctuations in random sampling, but does not indicate what *does account* for the difference. The failure to take this fact into consideration and the tendency to provide only very superficial interpretations of obtained differences have been major sources of error in educational and psychological research. For example, in many "methods" experiments (in which the relative effectiveness of two methods of instruction is determined by employing the methods simultaneously with two similar samples of pupils and comparing their average achievements at the close of the period of instruction) the investigator has made the mistake of concluding, simply because the obtained difference in achievement was "significant," that he had therefore definitely established the superiority of one method over the other. Even though the samples used may have been strictly random and all of the conditions for the application of the standard error formula satisfied, the possibility remains that uncontrolled factors other than the difference between the two *methods* may be the real reason for the difference obtained, as, for example, differences in the ability of the teachers employing the methods or differences in the contemporaneous incidental learning of the pupils in other subjects. Similar difficulties arise in the interpretation of "significant differences" in many other situations. The student of statistics should consciously strive to develop a highly critical attitude in the consideration of possible cause and effect relationships in such situations.

The Standard Error of a Difference between the Means of Related Variables

The derivation of Formula (16) involves the important assumption that the samples between which the difference is found are *independent* random samples. In some instances the samples which we wish to compare may consist of individuals who may be *paired* (between samples) on some basis, and the measures obtained may be related for the individuals constituting these pairs. Suppose, for example, that we wish to compare the mean intelligence of married men with that of their wives. Suppose we select a random sample from the population of married men, and that our sample from the population of married women consists of the wives of these same men. Then suppose that we administer an intelligence test to these individuals, compute the mean score for each sample, and find the difference in these means. In this type of situation Formula (16) is not valid to describe the reliability of the difference and if used would exaggerate its unreliability.

The reason for this is that there is a definite relationship between the intelligence of husbands and wives. Men of superior intelligence tend to be married to women of superior intelligence, and men of low intelligence tend to marry women of low intelligence. Hence, if our sample of married men happened by chance to have a higher mean intelligence than most such samples, we would expect our related sample of women also to have a higher mean intelligence than most such samples of women. Two samples selected in this fashion would ordinarily be more nearly alike in mean intelligence than if the samples were independently selected, that is, if the women in the sample of women were not (except by chance in a few cases) the wives of the particular men selected. In a pair of *independent* samples, the obtained mean of the men might be above the true mean of men at the same time that the obtained mean of the women was below the true mean of the women, but in a pair of related samples this would happen much less frequently. A distribution of differences in means of related samples would therefore show less variability than a distribution of differences in means of

independent samples. In other words, the standard error of the difference for a pair of related samples would be smaller than for a pair of independent samples — how much smaller would depend upon the strength of the relationship. The formula for the standard error of the difference between means of related variables is as follows:

$$\sigma_{(M_1 - M_2)} = \sqrt{\sigma_{M_1}^2 + \sigma_{M_2}^2 - 2r_{12}\sigma_{M_1}\sigma_{M_2}}, \qquad (17)$$

in which σ_{M_1} and σ_{M_2} are the standard errors of means M_1 and M_2, and r_{12} is the coefficient of correlation between the related variables.

It will be noted that if r_{12} is 0, that is, if the variables are unrelated, Formula (17) becomes the same as Formula (16).

If r_{12} is not known, and if the sample is not very large, it may be more convenient to compute the difference for each pair of measures separately and to estimate the standard error of the mean of these differences (which is the same as the difference in means) by substituting the standard deviation of the individual differences in Formula (8), the N in the formula representing the number of differences (or pairs of measures). Suppose, for example, that each individual in a sample of 30 is weighed before and after going on a certain standard diet, and we wish to know whether the observed gain (or loss) in mean weight is significant. Suppose the initial (I) and final (F) weights, and the corresponding differences (D), are as follows:

	$I.$	$F.$	$D.$		$I.$	$F.$	$D.$		$I.$	$F.$	$D.$
1.	142	146	4	11.	147	147	0	21.	146	145	− 1
2.	140	139	− 1	12.	154	164	10	22.	131	128	− 3
3.	143	148	5	13.	106	108	2	23.	157	161	4
4.	158	161	3	14.	156	165	9	24.	165	167	2
5.	149	151	2	15.	172	176	4	25.	153	154	1
6.	140	138	− 2	16.	121	117	− 4	26.	180	178	− 2
7.	134	135	1	17.	157	159	2	27.	105	112	7
8.	124	127	3	18.	156	163	7	28.	154	153	− 1
9.	116	115	− 1	19.	149	152	3	29.	126	124	− 2
10.	157	162	5	20.	140	142	2	30.	122	121	− 1

The mean of these differences is 1.93, and their standard deviation is 3.45. Hence, the estimated standard error of the mean difference is $3.45/\sqrt{29} = .65$, and the significance ratio is $1.93/.65 = 2.96$. Since this is considerably larger than the significance ratio (2.58) required for significance at the 1 per cent level, we can reject the null hypothesis in this case with a very high degree of confidence.

In simple experiments intended to determine the relative effectiveness of two methods of instruction, the usual practice is to select two samples of pupils from the same population, to teach one group by one method and the other by the other method for a given period of time, and then to administer the same achievement test to both groups and to find the difference in their mean scores on this final test. If the samples are independently selected at random, Formula (16) may be employed to determine the reliability or significance of this difference. Very often, however, instead of selecting the samples independently, we "match" or "equate" them on some basis (for example, intelligence) at the beginning of the experiment. In other words, each pupil in one group is paired with a pupil of the same intelligence in the other group, so as to give to neither method an accidental advantage in the final comparison. In this case again Formula (16) is not strictly valid, since the samples used are not independent. The special techniques appropriate for testing the significance of the results of experiments of the "matched group" type and of other more complex types of experiments are not within the scope of this introductory course.[1]

Small Sample Theory: Establishing a Confidence Interval for the True Mean

It will be remembered that in establishing a confidence interval for the true mean the procedure (for large samples) is to: (1) estimate the standard error of the obtained mean, using Formula (8); (2) multiply the estimated standard error by the "critical value"

[1] See Lindquist, *Statistical Analysis in Educational Research*, especially chap. IV.

(for the selected level of confidence) of the significance ratio;[1] and (3) "lay off" this distance on either side of the obtained mean to determine the limits of the confidence interval. The "critical value" of the significance ratio, as determined from the normal probability integral table, is 1.28 for the 20 per cent level of confidence, and 1.96, 2.33, 2.58, and 3.33 for the 5 per cent, 2 per cent, 1 per cent, and 0.1 per cent levels respectively.

This procedure is based on the assumption that the significance ratios are normally distributed for a large number of samples of the given size — an assumption which is not valid if the sample is small. For small samples the significance ratios form a distribution that has longer tails than the normal distribution, and the form of distribution differs from one size of sample to another. Hence the "critical values" derived from the normal table are not applicable to significance ratios computed for small samples. However, the exact form of the distribution of significance ratios is known for each size of sample, and the exact "critical values" for each of the commonly used levels of confidence have been determined for each size of sample from 2 to 31. These critical values are given in Table III in the Appendix. For reasons that need not be considered here,[2] one less than the size of the sample ($N - 1$) is referred to as the number of "degrees of freedom." The numbers in the first column in Table III represent the degrees of freedom for various size samples, that is, they are equal to $N - 1$.

To show how this table may be used, suppose that the mean and S.D. of a sample of 10 cases are 11.00 and 3.60 respectively. According to Formula (8), the estimated standard error of the mean is $3.60/\sqrt{10 - 1} = 1.20$. According to Table III, the critical value of the significance ratio (t) at the 1 per cent level for a sample of 10 cases (degrees of freedom = $N - 1 = 9$) is 3.250. Hence

[1] We have heretofore used "significance ratio" to refer to an obtained difference divided by its standard error; we here use the term to refer similarly to the ratio of the difference between the hypothetical and obtained means to the estimated standard error of the obtained mean.

[2] Students desiring a more thorough explanation of small sample theory may refer to Lindquist, *Statistical Analysis in Educational Research*, pp. 18–21 and 48–75.

the limits of the 1 per cent confidence interval are ($11.00 - 3.250 \times 1.20$) and ($11.00 + 3.250 \times 1.20$) or 7.1 and 14.9.

The manner in which any exact hypothesis concerning the true mean may be tested (for small samples) by the aid of Table III should be apparent from the foregoing and from the discussion on pages 110–115.

Small Sample Theory: The Significance of a Difference in Means of Independent Samples

The procedure for testing the significance of a difference in the means of independent random samples is similar for large and small samples, except that in the latter case the significance ratio is differently calculated, and the critical value of the significance ratio is read from Table III rather than from Table 17. The formula for the significance ratio (t) for the difference in the means of two independent small random samples is

$$t = \frac{M_1 - M_2}{\sqrt{\left(\frac{N_1\sigma_1^2 + N_2\sigma_2^2}{N_1 + N_2 - 2}\right)\left(\frac{N_1 + N_2}{N_1 N_2}\right)}}, \qquad (18)$$

in which M_1 and M_2 are the obtained means, σ_1 and σ_2 the corresponding standard deviations, and N_1 and N_2 the corresponding numbers of cases.[1] The number of degrees of freedom for this t is $(N_1 + N_2 - 2)$.

For example, if a random sample of 5 cases from population A

[1] The denominator of this expression is the estimated standard error of the difference, under the hypothesis that both samples are drawn at random from the *same* population. We may see now an advantage of defining the level of confidence with which an exact hypothesis may be rejected in terms which do not consider the *direction* of the hypothetical sampling error. If the null hypothesis *is* true, then the two populations are *identical* in the trait measured, that is, they constitute a *single* population as far as that trait is concerned. Under the null hypothesis, therefore, the two samples involved are really drawn at random from the *same* population. Now if two samples drawn at random from the same population have different means, there is obviously no basis for saying that the *difference* in means is either positive or negative. We can say that the difference has a certain *magnitude*, but we cannot meaningfully attribute a definite *direction* to it. Accordingly, the second of the definitions given in the footnote on page 112 is inappropriate in testing the null hypothesis, since under that hypothesis the *direction* of the sampling error is indeterminate.

has a mean of 19.50 and a standard deviation of 2.65, and a sample of 11 cases from population B has a mean of 15.13 and a standard deviation of 3.20, then the significance ratio is

$$t = \frac{19.50 - 15.13}{\sqrt{\left(\frac{5 \times 2.65^2 + 11 \times 3.20^2}{5 + 11 - 2}\right)\left(\frac{5 + 11}{55}\right)}} = \frac{4.37}{\sqrt{3.0701}} = 2.49.$$

The number of "degrees of freedom" for this t is $N_1 + N_2 - 2 = 14$. According to Table III, for this number of degrees of freedom a t of 2.624 is required for significance at the 2 per cent level, or of 2.145 at the 5 per cent level. Hence this difference would be described as significant at the 5 per cent level.

It is important to note that the procedure just described is valid only if the *true* standard deviations of the populations involved are approximately equal. If the obtained results or other considerations suggest that this assumption is not satisfied, this formula for t is not valid. It should be noted, however, that chance alone will produce large differences in the *obtained* standard deviations. For the samples in the illustration just used, for instance, one standard deviation might easily be twice the other as the result of chance, even though the true standard deviations were equal.

The significance of a difference in the means of paired or related measures may be tested in the manner illustrated in the example on page 135, except that for small samples the "critical" values of the significance ratio should be read from Table III.

Limitations of Sampling Error Techniques Designed for Large Random Samples

The sampling error techniques that have been presented in this chapter are designed for use only with simple random samples (in most cases only for samples of considerable size). However, most samples actually employed in educational and psychological research are *not* simple random samples, and to apply to them the techniques here presented would often be more misleading than helpful. It is therefore extremely important that the student un-

derstand clearly what is meant by a random sample, and that he be able to identify any consequential departure from randomness in practical sampling situations.

A simple random sample of N cases is one so drawn that *any* set of N particular individuals is just as likely to be selected as any other set. For instance, if a sample of 100 cases from the population of ninth grade pupils in Iowa public schools is to be truly random, it must be drawn so that any combination of 100 particular high-school freshmen has as good a chance to be selected as any other particular combination. When we stress the fact that "any combination" includes samples consisting of 100 pupils from as many different schools which may be located in the most inaccessible sections of the state, we realize how impracticable is strictly random sampling in situations of this kind.

In most sampling from populations of school children, the pupils must be taken in *intact groups*, rather than independently as individuals, as would be required in simple random sampling. One of these intact groups may consist of the pupils in a single classroom, or in a single building, or in a single school system, or of the children in a given community, etc. Sampling by such intact groups is necessary in part to avoid the inconvenient geographical distribution of pupils that was suggested in the preceding illustration, and also because the things to be investigated or experimented with, such as methods of instruction or educational tests, must usually be administered simultaneously to *groups* of pupils rather than to separate individuals. Thus, if an educational research worker wanted to conduct an experiment involving 500 pupils, he would probably arrange with a number of school administrators to permit him to use whatever number of *intact classes* would total 500 pupils, instead of attempting to select 500 pupils strictly at random from the population in which he is interested. These classes would differ from one another in ability and achievement much more than would random samples of the same size, due to systematic differences in quality of instruction, previous educational experience, nature of community, etc. It is readily apparent that a sample

consisting of a small number of such intact groups (regardless of the number of pupils) cannot yield as reliable results as a simple random sample of the same size. It is very obvious, for instance, that 1000 pupils taken 500 from each of two school systems is much less likely to be representative of pupils in general than a simple random sample of 1000 cases in which the pupils would be drawn from hundreds of school systems. Consequently, it would be a serious mistake to apply Formula (8) to the mean of the first of these samples (letting $N = 1000$), since the standard error thus estimated would suggest that the mean is much more reliable than is actually the case.

There are available [1] special sampling error techniques that are valid for use with samples consisting of intact groups, if these intact groups are selected at random, but these techniques are beyond the scope of this introductory course. However, it is important that the student know that appropriate techniques are available, and that he recognize that the techniques here presented should not be used with samples of this type.

It has already been suggested that in actual research the choice of the method of sampling to be employed is often governed by factors of expediency or of administrative convenience. In actual practice we usually secure our sample from the relatively small part of the whole population that is conveniently accessible to us, and there is always the possibility that the more accessible individuals might differ systematically from the less accessible. As a result, the samples which we select, often without our being conscious of the fact, are frequently "loaded" (to an extent greater than would happen in random sampling) with individuals who are superior or inferior to the typical individual in the population that we are studying.

Whenever a sample is selected by a method that in the long run would yield samples whose obtained measures differ systematically from the corresponding true measures, we say that the sample drawn is a *biased* sample. In other words, a sample is biased if

[1] See Lindquist, *Statistical Analysis in Educational Research*, pp. 66 ff.

other samples drawn in the same manner contain sampling errors that are more often in one direction than in the other. Again, a sample may be said to be biased if drawn by a process which gives certain individuals (or individuals of a certain type) a better chance of being drawn than certain other individuals. Unfortunately, the sources of bias are frequently difficult to detect, and samples may be seriously biased without our being conscious of the fact.

Obviously, errors in sampling that are due to bias, that is, that are due to *failure* to obtain a *random* sample, are not taken into consideration by the formulas that have here been considered. Such errors, nevertheless, are among the most important of the errors which characterize actual sampling studies.

Sometimes, in order to reduce the probability of securing a biased sample, we select what may be described as a "controlled" sample. For instance, if we were studying the achievement of high-school freshmen in the state of Iowa in some school subject and recognized that there are systematic differences in average achievement between large and small schools, we might insure — by deliberate selection — that the proportion of pupils from schools in various enrollment classifications is the same in our sample as in the whole population. In other words, we might *make* our sample representative with respect to size of school, rather than allow chance to determine what proportion of pupils will be selected from schools of each enrollment classification.

It is the controlled type of sample, incidentally, that has made possible dependable polls of public opinion of the type conducted by Gallup, Roper, and others. Samples that have been properly "controlled" are considerably more reliable than random samples of the same size, and hence the techniques that have been presented in this chapter are not valid for use with such samples.

So much emphasis has here been placed upon the limitations of sampling error techniques designed for simple random samples (due either to the impracticability or the undesirability of random sampling) that the student may wonder if the time he has taken to become acquainted with these techniques was well spent. He

need, however, have no doubts on this score. In the first place, he will find some actual research situations in which these techniques are directly applicable. In the second place, while it is true that in practical research the methods of sampling and the experimental designs employed are usually of a relatively complex type that demand special sampling error techniques, these special techniques cannot possibly be understood until the student has first mastered thoroughly the simpler and more basic techniques designed for simple random samples. Any student who intends to engage extensively in educational or psychological research must acquire a more advanced statistical training than is provided in this course, and one of the principal purposes of this chapter has been to provide him with the foundation essential to such advanced training.

CHAPTER IX

STANDARD MEASURES AND METHODS OF COMBINING TEST SCORES

Standard Measures or z-Scores

ATTENTION has already been drawn (Chapter IV) to the fact that raw scores on educational and psychological tests ordinarily have little or no absolute significance and may not be directly compared from test to test. To interpret or compare such scores, we must first derive for each of them some measure of its *relative position* in the distribution to which it belongs. One of the most widely used of such derived measures is the percentile rank. The percentile rank, however, has several distinct limitations. It is a "counting" measure only, that is, it is not arithmetic in character. It may be unduly influenced by minor irregularities in the form of the distribution, and is therefore relatively unstable or unreliable. The inter-percentile distance fluctuates in magnitude throughout the scale, and may therefore not be considered as a unit. While, for administrative convenience, percentile ranks are frequently added or averaged to secure composite measures, this practice ignores the non-arithmetic character of the percentile rank and is not strictly valid.

Another derived measure, which is relatively free from the limitations just mentioned, is the standard measure or *z-score.* The z-score is algebraically defined by the formula

$$z = \frac{X - M}{S.D.} \tag{19}$$

in which z is the standard measure, X is a particular raw score in a given distribution, and M and S.D. the mean and standard deviation, respectively, of that distribution. The z-score corresponding to any given raw score indicates how many standard deviations that score deviates from the mean of the distribution. If, for example, the mean score on a test is 75 and the standard

deviation is 10 for a given group, then for that group the standard measure or z-score corresponding to a raw score of 100 is

$$z = \frac{100 - 75}{10} = 2.5$$

Likewise, a raw score of 60 in this distribution would have a z-score equivalent of -1.5. The z-score of 2.5 means that the corresponding raw score is 2.5 S.D.'s above the mean; the z-score of -1.5 means that the corresponding raw score lies 1.5 S.D.'s below the mean — the minus sign indicating that the score lies below, rather than above, the mean.

The use of the z-score does not involve any necessary assumption concerning the form of the distribution, but because of the definite relationship between the standard deviation and the normal curve (Chapter VII), the z-score may be most readily and adequately interpreted if the distribution concerned is approximately normal. If, for instance, a certain raw score has a z-score equivalent of $+2.0$ in a normal distribution, we know (Table 17) that it exceeds approximately 98 per cent of the scores in the distribution. Similarly, a z-score of -1.0 exceeds about 16 per cent of the measures. The value with which we enter Table 17, $\frac{x}{\sigma}$, is of course itself a z-score, since x represents X − M, the deviation of the raw score from the mean.

For distributions which do not approximate the form of the normal curve, the z-score is somewhat more difficult to interpret. In a distribution markedly skewed to the right, for instance, a considerable proportion of the scores might lie above the point which is 3 S.D.'s from the mean, while in a distribution markedly skewed to the left a point 3 S.D.'s above the mean might be considerably higher than the highest score in the distribution. Fortunately, however, distributions of test scores are rarely very markedly skewed, and hence z-scores above $+2.0$ may, in general, safely be considered as "very high," those between $+1.0$ and $+2.0$ as "high," those between -1.0 and -2.0 as "low," and those below -2.0 as "very low" *relative to the other scores in the same distribution.*

Transforming Raw Scores into Their z-Score Equivalents

When we transform a set of obtained scores into their z-score equivalents, we are in effect arbitrarily substituting another scale for the original raw score scale, such that the zero point on the new scale corresponds to the mean on the raw score scale, and such that the *unit* along the new scale is equal to the standard deviation of the original distribution. The arbitrary nature of this procedure may be made clear by the following. Suppose that along the raw score scale for a given frequency distribution we have marked the positions of the mean and of a point 1 S.D. above the mean. Suppose, also, that on a wide rubber band we have marked off in white ink a number of equally spaced points, have written zero opposite one of these points (near the middle of the band), and have numbered the remaining points consecutively + 1, + 2, + 3, etc., and − 1, − 2 and − 3, on either side of this zero point. If we then placed this rubber band alongside the original raw score scale and stretched and adjusted it until the zero point on the band came opposite the mean on the raw score scale and the point + 1 on the band came opposite the point 1 S.D. above the mean on the raw score scale, the scale on the rubber band would then represent the z-score scale for the distribution involved.

When all of the scores in a large distribution are to be transformed into z-scores, it may be more economical to prepare an *equivalence table* than to apply Formula (19) to each score individually. The steps in preparing a table of z-score equivalents for the raw scores in a given distribution are presented below, the statements in parentheses referring to the illustration in Table 19.

1. Compute the M and S.D. of the distribution.
   (The M and S.D. of the distribution in Table 19 are 40.10 and 5.136 respectively).
2. List, in a column or columns, the values of all possible scores in the distribution.
   (This has been done in the columns headed X in Table 19.)
3. Find the integral score just above the mean, and compute its z-score equivalent by Formula (19).

(In Table 19, the score just above the mean is 41. Its z-score equivalent is $\frac{41 - 40.10}{5.136} = .176$.)

4. Find the reciprocal of the standard deviation of the distribution. This is the difference between the z-scores corresponding to two consecutive integral raw scores.

$$\left(\frac{1}{5.136} = .195\right)$$

5. Add this reciprocal to the result of Step 3 to get the z-score corresponding to the next highest integral score. By similar consecutive additions of this reciprocal, compute the z-score equivalents of the remaining raw scores above the mean, entering each (in the z column) opposite the corresponding raw score as it is obtained.
(.176 + .195 = .371, the z-score corresponding to 42;
.371 + .195 = .566, the z-score corresponding to 43; etc.)

6. Determine the z-score corresponding to the raw score just below the mean, and compute the z-scores for the remaining raw scores by consecutive additions of the reciprocal as explained above.

TABLE 19

ILLUSTRATING THE CONSTRUCTION OF A TABLE OF Z-SCORE EQUIVALENTS FOR THE SCORES IN A GIVEN DISTRIBUTION

Frequency Distribution		z-Score Equivalents of Raw Scores					
Interval Midpoints	f	X	z	X	z	X	z
54	1	56	3.101	46	1.151	36	− .800
51	1	55	2.906	45	.956	35	− .995
48	3	54	2.711	44	.761	34	− 1.190
45	8	53	2.516	43	.566	33	− 1.385
42	17	52	2.321	42	.371	32	− 1.580
39	13	51	2.126	41	.176	31	− 1.775
36	10	50	1.931	40	− .0195	30	− 1.970
33	4	49	1.736	39	− .215	29	− 2.165
30	1	48	1.541	38	− .410	28	− 2.360
27	2	47	1.346	37	− .605	27	− 2.555
						26	− 2.750
N =	60					25	− 2.945
A.M. =	40.10					24	− 3.140
S.D. =	5.136						
1/S.D. =	.195						

The z-score corresponding to any given raw score can now be quickly read from the table; for example, the z-score equivalent of a raw score of 30 is -1.970 or -2.0 (rounded), while that for a raw score of 53 is $+2.5$. If an adding machine is available, Steps 5 and 6 in the procedure just described can be very quickly completed. The standard deviation and its reciprocal should be carried to three decimal places to avoid a large cumulative error at the extremes of the distribution, but in reading z-scores from the table it is usually well to round to one decimal place.

T-Scores

It has been noted that the z-score scale is an *arbitrary* scale adjusted to fit the raw score scale in a prescribed manner. If we wished, we could select any other reference point and any fraction or multiple of the S.D. as a unit in constructing this scale. To return to the rubber band illustration, we could, for example, divide the band into 10 equal intervals, numbering them 10, 20, 30, etc., up to 100, and then could adjust the rubber band to the raw score scale such that the point 50 would come opposite the mean of the distribution and the point 60 would come opposite the point 1 S.D. above the mean of the original distribution. In other words, we could arbitrarily set the mean of our new scale equal to 50 and the standard deviation equal to 10. This particular type of scale is ordinarily known as a T-scale, and scores expressed along this scale as T-scores. The algebraic formula for a T-score, then, would be:

$$T = \frac{10\,(X - M)}{S.D.} + 50 \qquad (20)$$

where X, M, and S.D. have the same significance as in the z-score formula. The name "T-scale" was originally applied to this scale by McCall.[1] The advantage of the T-scale lies in the fact that it does away with the necessity of dealing with negative scores and with scores expressed as decimal values. In all other respects, however, the z-score and the T-score are essentially equivalent.

[1] William McCall, *How to Measure in Education.*

Raw scores may be transformed into T-scores by the same type of procedure as has just been explained for z-scores.

Composite Measures

T-scores and z-scores and other measures based upon the mean and standard deviation are very frequently employed for the purpose of deriving *composite* measures based upon a number of scores originally expressed in different units. If we attempt to secure composites of performance on different tests for the individuals of a given group by simply averaging directly the raw scores of each individual, we automatically give to the score on each test a weight which is proportional to the variability (sigma) of the distribution of scores on that test. Suppose, for example, that on two tests, A and B, administered to a given group, the means and standard deviations are as given below.

			Scores of	
	M.	S.D.	Pupil #1	Pupil #2
Test A	120	15	135	105
Test B	85	25	60	110

Suppose that pupil number 1 makes a score 1 S.D. above the mean on Test A and 1 S.D. below the mean on Test B, while pupil number 2 makes a score 1 S.D. below the mean on Test A and 1 S.D. above the mean on Test B. The sum of the scores on the two tests would be 195 for pupil number 1 and 215 for pupil number 2. Pupil number 2 would then receive the higher composite score simply because the test on which he happened to perform the better was that with the larger standard deviation, whereas if the tests were to be considered as equally important both composites should be the same. Scores on Test B, then, are given greater weight in the composite, even though their mean value (85) is less than that of the scores on Test A. A long test, or one with a large number of items, does not necessarily carry any greater weight in a composite than a short test, or one with just a few scoring units, since the shorter test may have the larger S.D. of scores. Therefore, to insure that each test is given the same weight

in the composite, the scores on all tests should be transformed so that each distribution of transformed scores shows the same standard deviation. This, of course, is done when all scores are transformed into z-scores or T-scores. Where z-scores from different tests are added or averaged to secure a composite, each test is given the same weight. If it is desired to weigh certain tests more than others, this can be done by multiplying the z-scores for those tests by any desired number before the scores are combined.

Sometimes, when the scores from a number of tests are to be combined for the individuals in a given group to obtain measures of composite performance, and when it is desired to give each test equal weight, it may be more convenient to multiply the raw scores on each test by an integral number which is roughly proportional to the reciprocal of the S.D. on that test, and then combine the scores thus derived for each individual. Suppose, for example, that for a given group the distributions of scores on three tests, A, B, and C, show S.D.'s of 12, 21 and 9 respectively. The reciprocals of these S.D.'s are $\frac{1}{12} = .083$, $\frac{1}{21} = .047$, and $\frac{1}{9} = .11$ respectively. The smallest integral numbers closely proportional to these values are 8, 5 and 11. However, the integers 2, 1, and 3 are roughly proportional to these reciprocals, and would be sufficiently accurate for most practical purposes. If, then, all scores on Test A were multiplied by 2, the S.D. of the scores thus derived would be 2×12 or 24. Similarly, if each of the scores on Test C were multiplied by 3, the S.D. of the derived scores would be 27. These S.D.'s are so nearly equal to the S.D. of the original raw scores on Test B (21) that a fairly satisfactory composite could be secured by adding to the B scores the derived scores for Tests A and C. (If a more equitable weighting were desired, all of the scores on Test A should be multiplied by 8, all scores on Test B by 5, and all scores on Test C by 11, which would result in S.D.'s of 96, 105, and 99 respectively.)

The conditions under which, and the purposes for which, z-scores

and T-scores and other similar derived scores may be validly applied are left to the student to discover for himself with the aid of the suggestions offered in the study exercises. It is particularly important that the student become thoroughly familiar with the z-score technique, not only because of its frequent application in practical work, but more especially because a thorough understanding of z-scores will lead to a better appreciation of other statistical techniques. A thorough understanding of z-scores is particularly essential in the study of simple correlation theory in the following chapter.

CHAPTER X

CORRELATION THEORY

The Meaning of Correlation

WHEN measures of each of two traits are secured for each individual in a given group, it may frequently be noted that the two measures for any individual tend to have roughly the same relative position in their respective distributions; that is, individuals far above average in one trait tend also to be well above average in the other, those below average in one tend to be correspondingly below average in the other, and those at or near the average in one tend also to be at or near the average in the other. When this is true, we say that the two traits (or measures) are "positively related" for the group in question, or that they show a "positive correlation." Height and weight, for example, are positively related for almost any group; that is, the tall individuals tend also to be the heavy and the short individuals to be the light.

Sometimes traits may be found such that measures of these traits for the individuals in a given group are "negatively related." By this we mean that individuals above average in one tend to be below average in the other, while those below average in the first tend to be above average in the second. For the children in the seventh grade in almost any public elementary school, for example, chronological age and scholastic ability are likely to be negatively related; that is, the over-age children in the grade are usually among the dullest, while the youngest children are usually among the brightest. The reason is that the dull children have been retarded and the bright children accelerated in their school progress.

The nature of the relationship between two variables can be most readily studied by one not technically trained in statistics by preparing a "scatter-diagram" for the measures obtained. Suppose, for example, that we wish to study the nature of the re-

lationship between the scores on an arithmetic test and on a reading test for a given group of seventh grade pupils. To do this graphically, we could subdivide a square or rectangle into a large number of "cells" by drawing equally spaced and parallel horizontal and vertical lines through it as in Figure 16. Each horizontal row of cells could then be made to correspond to a given interval along the scale of arithmetic scores, and each vertical column to an interval along the scale of reading scores. For example, in Figure 16 the upper row represents the interval 43–47 along the arithmetic scale, while the third column represents the interval 30–39 on the reading test scale. The scores of any pupil could then be represented on this diagram by a single tally mark placed so that its position with reference to the vertical scale represents his arithmetic score and so that its position with reference to the horizontal scale represents his reading score. For example, if a pupil made a score of 36 on the arithmetic and 63 on the reading test, we would place the tally mark for him in the cell which is both in the 33–37 row and in the 60–69 column, that is, in the sixth cell (from the left) in the third row (from the top).

The tally marks in Figure 16 represent the reading and arithmetic test scores for a group of 62 pupils. Each number along the bottom of the figure represents the total frequency in the column above it, while the numbers along the right-hand margin represent the frequencies in the individual rows. For example, 14 pupils made scores of 40–49 on the reading test, while 12 pupils made scores of 23–27 in arithmetic.

For the pupils tallied in any single column we could, if we wished, compute the mean score made by them on the arithmetic test. For example, the 5 pupils tallied in the first column (who made scores of from 10 to 19 on the reading test) made a mean score of 19 on the arithmetic test (computed by using as the arithmetic score of each pupil the midpoint of the arithmetic interval in which he is tallied). The position of this mean along the vertical scale is represented by the small circle in the first column. Similarly, the circle in each of the other columns represents the value of the

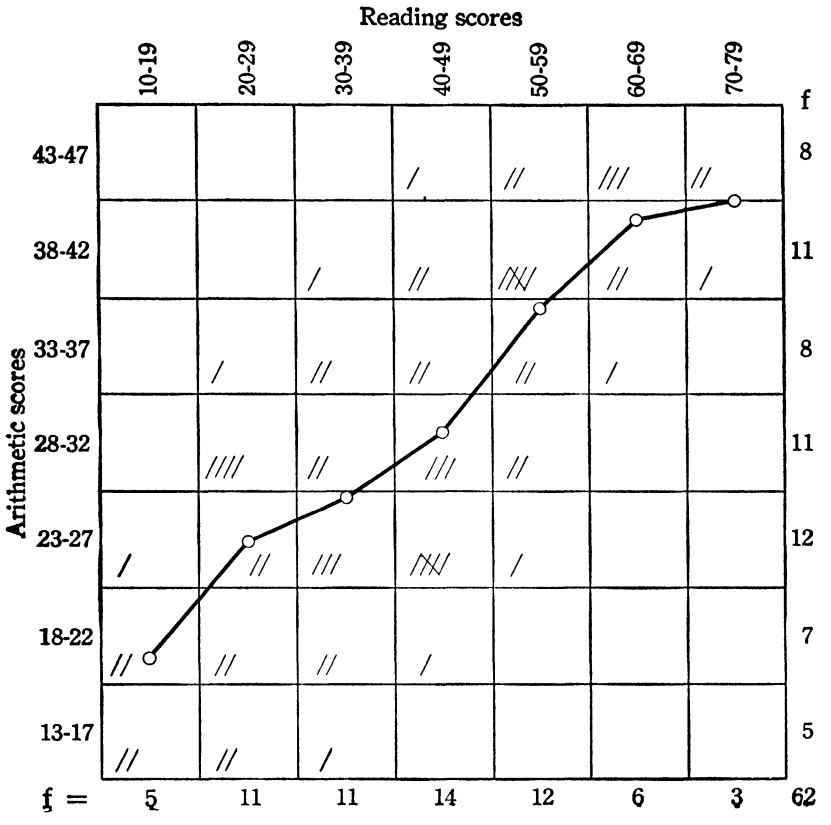


FIG. 16.
Scatter-diagram of reading and arithmetic test scores for a group of 62 pupils.

mean score on the arithmetic test for the pupils tallied in that column.

It is at once apparent that these means tend to fall along a straight line running from the lower left-hand to the upper right-hand corner of the diagram. It is probable that the only reason that they do not lie exactly on a straight line is that each mean is based upon such a very small number of cases and is therefore unstable because of sampling error. Had enough pupils been tested so that the frequency in each column had been large, it is likely that the column means would much more closely approximate a straight line pattern than did the means represented in Figure 16.

In the fashion already described, we could calculate also the mean reading score for the pupils tallied in each horizontal *row* and mark the position of the mean in each row individually. If this were done, we would find that these means also would tend to lie along a straight line, although the position of this straight line would not correspond to that which best fits the means of the columns.

Whenever the relationship between measures of two variables is such that the means of the rows and the means of the columns on the scatter-diagram each tend to lie along a straight line, we say that these variables are "rectilinearly" related, or that they represent an instance of rectilinear correlation. Not all variables, however, are related in this way. Sometimes we find that the means of the rows or of the columns lie along a curved line. For example, if we were to plot on a scatter-diagram the age and some measure of bodily strength for each individual in a group which includes all age levels from infancy to extreme old age, we would find that the mean strength for individuals of a given age increases during the periods of childhood and adolescence, that it remains relatively stable from early adulthood until well past middle age, and that it decreases at the higher age levels, dropping quite rapidly when the age of senescence is reached. If ages were plotted along the horizontal scale and strength measures along the vertical scale, the smooth line which would best fit the means of the columns would be a curve running upward from left to right, gradually flattening out until the maximum was reached near the middle of the age range, after which it would drop, at first slowly and then rapidly, to the end. Variables which show a curved pattern of tally marks in a scatter diagram are said to be "curvilinearly" related. Such variables cannot always be described as being positively or negatively related, since the relationship may be positive along certain portions of the scale and negative along other portions.

If the means of the rows on a scatter-diagram tended to lie along a straight *vertical* line, while the means of the columns tended to lie along a straight *horizontal* line, we would say that the two

variables were entirely unrelated; that is, individuals high in one measure would tend to be neither high nor low in the other. For any adult group, for example, height and intelligence would probably show zero relationship, since the mean intelligence of persons of any given height would tend to be the same as for persons of any other height.

Because of the variety of ways in which two variables may be related, it is difficult to describe in a single statement what is meant by a "relationship" between two variables. Perhaps the best general definition of related variables would be as follows: measures of two traits for a given group of individuals may be said to be related if all individuals in the group who have the *same* measure of one trait show less *variability* in the second trait than do the individuals in the entire group. For example, according to this definition, height and weight may be said to be related because, if from any large group we select a number of individuals all of whom are of the same height, these individuals will be more alike with respect to weight than are all individuals (of differing heights) in the entire group. This definition would apply equally well whether the relationship were curvilinear or rectilinear, positive or negative.

Measures of one pair of traits, of course, may show a different *degree* of relationship for the individuals in a given group than do the measures of a different pair of traits for the same group, or the same pair of traits may show different degrees of relationship for the individuals in different groups. Using the approach suggested by the definition just given, we may say that measures of two traits are *highly* related if individuals who are exactly alike in the measures of the first trait tend also to be very much alike in the measures of the second. Measures of two traits may be said to show low relationship if individuals alike in the first trait show wide variations in the second. Height and weight, for example, are not highly related, since we know that individuals of the same height may show wide variations in weight. Arm span (distance from fingertip to fingertip when both arms are outstretched horizontally)

is highly related to height, since this span tends to be very nearly the same for all individuals of the same height and to differ proportionately for individuals of different heights. Height and intelligence are unrelated for most adult groups, since individuals of the same height are just as variable in intelligence as are individuals of differing heights.

Employing the approach suggested by the scatter-diagram, two variables may be said to be highly related if the measures in each row (or column) cluster closely about the line (either curved or straight) which most closely fits the means of the rows (or columns). On a scatter-diagram representing high positive rectilinear relationship, therefore, the more heavily concentrated tally marks (or the larger cell frequencies) would tend to fall into a pattern represented by a very narrow oval running from the lower left-hand to the upper right-hand corner of the diagram. If the relationship were medium and negative, the larger cell frequencies would lie inside a relatively broad oval whose axis would run from the upper left-hand to the lower right-hand corner of the chart. If there were no relationship between the two variables, about the same number of tally marks would be found in each quadrant [1] of the chart, and the larger cell frequencies would lie within a circle whose center would lie at the intersection of the lines fitting the means in the rows and in the columns. (See Figure 19.)

The Significance of Correlation

The nature of the relationship and the degree of relationship between measures of two traits for the individuals in a given group may be of significance in education and psychology for a number of different purposes, among the most important of which are prediction of future success, the description of the reliability and validity of measurement, and the study of cause and effect. These and other ways in which correlation is important will be considered

[1] The chart could be divided into four parts or "quadrants" by drawing a horizontal line across it through the general mean along the vertical scale and a vertical line through the mean on the horizontal scale.

in greater detail later, and will therefore be only very briefly illustrated here.

To illustrate the significance of correlation in prediction, suppose that a special examination designed to measure "scholastic aptitude" was administered last year to each member of the freshman class upon entrance to a certain university, and that at the end of the academic year a scatter-diagram was prepared showing the relationship between the scores made on this examination and the grade-point averages earned by the freshmen during that year. Let us suppose that this relationship is fairly high and positive. Assuming that the freshman class studied is fairly representative of succeeding freshman classes, this examination could then be used in subsequent years to predict, at the time of entrance, which students would later succeed or fail in their freshman courses. On the basis of these predictions, certain students could be advised to alter their plans, or could be placed in sections in which instruction is specially adapted to the level of ability of the group taught. If more than one examination designed for this purpose had been administered to the freshmen at the beginning of the year, and if it was later shown that the scores on one of these examinations were more highly related to grade-point averages than the scores on the other examinations, then this examination would, of course, be the best to use later for purposes of prediction. Through the study of correlations, then, a selection may be made from a number of possible different bases for predicting success, not only in scholastic work, but also in many other types of activity.

To illustrate the second of the purposes mentioned, let us suppose that in an attempt to estimate the general spelling ability of persons in a given group, two lists of 100 words each have been independently selected at random from the words in a certain abbreviated dictionary. Suppose that each of these lists is administered as a "list-dictation" test to the given group and that the number of words spelled correctly in each list is obtained for each student. The reliance which could be placed upon the score obtained from either test as a measure of general spelling ability

would be dependent upon the degree of relationship existing between the two sets of scores. If there were no relationship between these scores — that is, if individuals making a high score on one test were just as likely to make a low as a high score on the other — then no reliance could be placed upon either score as a measure of the ability of the individual student. If, on the other hand, there were close agreement between the two sets of scores, this would indicate that both tests are measuring the same ability with high dependability. The degree of correlation between scores on equivalent tests, therefore, constitutes a measure of the reliability of the measures provided by either test.

To illustrate the third of the purposes mentioned, let us assume that for a given group of readers the mean number of "eye fixations" per line made in reading a given printed selection is determined for each individual. Suppose, also, that for each individual there has been secured a measure of *rate* of reading. If, then, it can be shown that a high negative relationship exists between mean number of eye fixations per line and reading rate, this fact would suggest, although it would not prove, that the character of eye movements is an important factor in determining reading ability. It would suggest further, although again it would not prove, that an individual's reading rate might be improved by specific training intended to increase his eye span or to decrease his number of fixations per line. If, again, a higher relationship could be shown to exist between number of fixations per line and reading rate than exists between some other characteristic of the person's reading habits and his reading rate, this would suggest (but again not prove) that the first factor is more important than the second in determining an individual's speed of reading.

The Need for a Quantitative Measure of Relationship

The preceding illustrations are only suggestive of the many ways in which a study of correlation between obtained measures may be of assistance in attacks upon many educational and psychological problems. For most of these purposes, it is essential

that the description of relationship be reduced to a single numerical index which can be conveniently interpreted and readily compared with other similar indices. While it is possible to secure a rough notion of the degree of relationship between two sets of measures by simply inspecting the scatter-diagram prepared for them, just as it is possible to estimate the central tendency and the variability of a frequency distribution by inspection, the notions thus secured are not sufficiently objective or quantitative for comparative purposes. Our problem, then, given two sets of related measures for a given group of individuals, is to derive from these measures a single number or index which is proportional to the degree of relationship, and which is comparable to other measures similarly obtained.

The Selection of an Index of Relationship

Suppose, then, that for each of the individuals in a given group we have the scores made on each of two school examinations and that we wish to obtain a quantitative measure of the degree of relationship between these scores for that group. The arbitrary character of the procedure which we shall finally adopt (the Pearson product-moment coefficient of correlation) may best be made clear by first considering and rejecting a number of other equally arbitrary but less satisfactory solutions.

Since the scores on these tests are expressed in different units, it should at once be apparent that we cannot readily derive from them any measure of relationship until they have first been expressed in comparable terms. One way of doing this would be to express each score in terms of its *rank position* when the scores on each test are arranged in order of magnitude. If this were done and if it were found that there was, in general, a close agreement in the two ranks for each individual, then we could say that a high relationship existed. If, on the other hand, large differences between the two ranks characterized most individuals, then we would say that a low or perhaps even a negative relationship existed, depending upon the magnitude of the differences. This

suggests that we could secure a quantitative index of the degree of relationship by determining the difference in the two ranks for each individual and then averaging these differences for the entire group. If the mean value of these differences (all differences being considered as positive) were very small, we would say that a high relationship existed. If the mean difference in rank were large, the relationship would be low or negative. The magnitude of this mean difference in rank, however, obviously would depend upon the number of individuals in the group. A difference between a rank of 3 and a rank of 7 would have quite a different meaning in a group of 10 individuals than in a group of 100. Hence, this type of index would not be comparable for groups of different sizes.

This objection could be overcome by expressing each score as a *percentile rank*, and finding the mean difference in percentile ranks for the various individuals. This index would be comparable for groups of different sizes, but it would be inversely proportional to the degree of relationship (the smaller the mean difference, the higher the degree of relationship), and would continue to increase as the relationship changed from positive to negative. It would therefore be difficult to determine any point along the scale of possible values of the mean difference in percentile ranks that would correspond to zero relationship. Furthermore, as we have already learned, percentile ranks are not directly proportional to the original raw scores, and the variations in inter-percentile distances from point to point throughout the scale would introduce ambiguities into the measures obtained.

Another possibility which has some advantages over the preceding suggestions would be to express the scores in each set as standard measures or z-scores, to find the difference between the two z-scores for each individual and compute the mean of these differences (all differences being considered as positive). This measure would provide a dependable basis for comparing the degrees of relationship between two sets of variables, but again would be difficult to interpret because it would be inversely proportional to

the degree of relationship, and would remain positive in cases of negative correlation.

The Mean z-score Product

There are many other ways in which z-scores (or other derived measures) corresponding to two related sets of measures may be combined arithmetically so as to produce a single number or index that is indicative of the degree of relationship existing. One of the most promising of these consists of determining the (algebraic) *product* of the two z-scores for each individual and finding the mean of these products for all individuals concerned. Let us consider some characteristics of the index thus derived.

Suppose, first, that the relationship between the two sets of measures for the group considered is high, positive, and rectilinear. This is equivalent to saying that most individuals above average in one trait are also above average in the other, or that only a relatively small number of individuals are above average in one measure and below average in the other. If this is the case, then the majority of individuals in the group will either have two positive z-scores or two negative z-scores. In either case, the algebraic product of the z-scores for such individuals will be positive in sign. For the relatively small number of individuals with a positive z-score in one distribution and a negative z-score in the other, the z-score products will be negative. Many of the positive products, furthermore, will be quite large, since high z-scores in one distribution will usually be paired with high z-scores in the other, and low (large negative) z-scores in one distribution will be associated with low z-scores in the other. For the entire group, then, the sum of the positive z-score products will greatly exceed the sum of the negative z-score products, so that the mean of the z-score products for all individuals will be positive.

Suppose, next, that the relationship considered is positive but low. This would mean that, while again most individuals above average in one measure would also be above average in the other, and vice versa, there would be a larger number of instances than

before in which individuals above average in one measure would be below average in the other. There would also be fewer large products than in the first instance, since individuals with extreme z-scores (either high or low) in one distribution would seldom also have extreme z-scores in the other. In this case, then, the sum of the positive z-score products would not exceed the sum of the negative products by as great an amount as in the first instance, and, while the mean of the products for the entire group would still be positive, we would not expect it to be as large as before. In other words, we would expect the mean z-score product to be larger for high than for low degrees of relationship.

Now let us consider the case of unrelated measures. To say that two sets of measures are entirely unrelated for a given group is to say that individuals above average in one measure are equally likely to be above average or below average in the other. For the whole group, then, the number of positive z-score products (except for chance) would be equal to the number of negative z-score products. The individual products would also tend to be small, since two extreme z-scores would seldom be paired together. The negative products, furthermore, would tend to be about the same size as the positive products. The algebraic sum of these products for the entire group would therefore approximate zero, as would the mean of the products.

It should now be apparent that if the relationship were negative — that is, if most individuals above average in one measure were below average in the other — then the z-score product for most individuals would be negative in sign. The algebraic sum, and hence the mean of all z-score products, would therefore be negative, while the absolute magnitude of the mean product would depend upon the degree of relationship.

Now let us consider finally the case of *perfect* rectilinear relationship. To say that two sets of measures are perfectly related (rectilinearly) for a given group is to say that each individual has exactly the *same* relative status in both distributions of measures. This again is equivalent to saying that every individual has

exactly the *same z-score* in both distributions. This being the case, the product of the two z-scores for any one individual must be the same as the *square* of either of his z-scores taken alone. Hence, the *sum* of the z-score products for all individuals is the same as the sum of the *squared* z-scores for the first distribution alone (or for the second distribution alone). From this it follows that the mean z-score product would be the same as the mean of the squared z-scores in either distribution.

At this point, we may remind ourselves that the standard deviation of any distribution of measures is the "square root of the mean of the squared deviations from the mean." Since each z-score is itself a deviation from the mean, the standard deviation of any distribution of z-scores is equal to the square root of the mean of the squared z-scores, that is:

$$\text{S.D.}_{(of\ z\text{-}scores)} = \sqrt{\frac{\Sigma z^2}{N}}$$

But the standard deviation of any complete distribution of z-scores is 1.00 by definition. Hence,

$$\sqrt{\frac{\Sigma z^2}{N}} = 1.00$$

Squaring both sides of this expression, we get

$$\frac{\Sigma z^2}{N} = 1.00$$

since the square of 1.00 is still 1.00. The mean of the squared z-scores, then, is always equal to unity for any complete distribution of z-scores.

We have already pointed out, however, that when two sets of measures are perfectly related rectilinearly the mean of the z-score products will be the same as the mean of the squared z-scores for either distribution. Since the mean of the squared z-scores is always 1.00, it follows that the mean of the z-score products is, in the case of perfect rectilinear relationship, always equal to unity. If the relationship is perfect and positive, the mean of the z-score products will be + 1.00. If the relationship is perfect and negative, the mean z-score product will be − 1.00.

If, then, we have two sets of measures that are rectilinearly related for a given group, if we transform the measures in each distribution into their z-score equivalents and obtain the product of the two z-scores for each individual, the mean of these products for all individuals will have the following characteristics:

Its value will be positive when the relationship is positive.

Its value will be zero when there is no relationship.

Its value will be negative when the relationship is negative.

Its value will be + 1.00 when the relationship is perfect and positive.

Its value will be − 1.00 when the relationship is perfect and negative.

Its value will lie between + 1.00 and − 1.00 for intermediate degrees of relationship, and will be larger for high than for low degrees of relationship. (For reasons that will be given later, the mean z-score product is *not directly proportional* to the degree of relationship. For example, a mean product of .8 does not indicate twice as close a relationship as a mean product of .4.)

Because of these characteristics, the mean z-score product is an excellent index for the quantitative description of degrees of relationship *when the relationship is known to be rectilinear.* The use of the mean z-score product for this purpose was first proposed by the English statistician Karl Pearson and is therefore called the *Pearson product-moment coefficient of correlation.* The universal notation for this coefficient is r. It may be algebraically defined as follows:

$$r_{xy} = \frac{\Sigma z_x z_y}{N} \tag{21}$$

in which r_{xy} is the coefficient of correlation between the x and y measures, in which $\Sigma z_x z_y$ means "the sum of the products of the z-scores for variables x and y," and in which N represents the number of products or the number of individuals in the group studied. Other subscripts may be used to identify the variables. For example, r_{12} would be read "r sub one, two," and would mean "the coefficient of correlation between variables 1 and 2,"

while r_{46} would be read "r sub four, six," and would have a similar significance.

The Computation of r

We have already seen that the coefficient of correlation between two rectilinearly related sets of measures for any group of individuals can be computed by (1) transforming each measure into its z-score equivalent in its respective distribution, (2) multiplying the two z-scores for each individual in the group, and (3) finding the mean of these products.

While this computational procedure is easily explained and readily understood, it is rendered impracticable by the amount of time required for the first step, particularly where the number of cases is large. In the practical situation, it is much more economical to work directly with the raw score values. A formula for this purpose may be derived by substituting the following values of z_x and z_y in Formula (21).

$$z_x = \frac{X - M_x}{\sigma_x} \qquad z_y = \frac{Y - M_y}{\sigma_y}$$

(where X equals a raw score in the X distribution, M_x equals the mean of the X's, σ_x equals the standard deviation of the X distribution, and where Y, M_y, and σ_y have a similar meaning.)

The result of this substitution is

$$r_{xy} = \frac{\frac{\Sigma XY}{N} - M_x M_y}{\sigma_x \sigma_y} \tag{22}$$

(While the mathematics required to understand the derivation of this formula from the simpler z-score expression is not beyond the average high-school graduate, in the interests of economy of time the beginning student in statistics is advised to take Formula (22) for granted and not to concern himself with the algebra of its derivation.)

According to this formula, the coefficient of correlation between two sets of scores or measures may be obtained from an unordered list of paired scores, as follows:

1. Compute the mean and standard deviation of each set of measures.[1]
2. Secure the product of the two raw scores for each individual, add the products and divide the sum by N, the number of cases.
3. Subtract from the mean of these products the product of the means of the two distributions.
4. Divide the result by the product of the two standard deviations.

If a multiplying type of computing machine is available, if a scatter-diagram of the measures is not desired, and if the number of cases is small, this procedure is perhaps as good as any other. When a computing machine is not available, and when a large number of scores are to be correlated, a more economical procedure is to compute r from a scatter-diagram by a "short" method (analogous to that used in computing the mean and the standard deviation of a frequency distribution) in which each score is expressed as a deviation from an arbitrary reference point (or guessed mean) in its own distribution.

The formula employed for this purpose is:

$$r_{xy} = \frac{\frac{\Sigma x'y'}{N} - \left(\frac{\Sigma x'}{N}\right)\left(\frac{\Sigma y'}{N}\right)}{\sqrt{\frac{\Sigma x'^2}{N} - \left(\frac{\Sigma x'}{N}\right)^2} \cdot \sqrt{\frac{\Sigma y'^2}{N} - \left(\frac{\Sigma y'}{N}\right)^2}} \qquad (23)$$

in which x' represents the deviation of an X score from the arbitrary reference point (A.R.$_x$) in the X distribution (that is, $x' = X - \text{A.R.}_x$), and y' represents the deviation of a Y score

[1] The mean of each set of measures may in this case be computed by simply adding the raw scores and dividing the sum by the number of cases. The standard deviation of each series can similarly be computed without preparing a frequency distribution by (1) squaring each raw score; (2) securing the sum of the squared scores for the whole series; (3) dividing this sum by N; (4) subtracting from this quotient the square of the mean score; (5) extracting the square root of the result. The formula for the standard deviation employed in the preceding steps is

$$\text{S.D.} = \sqrt{\frac{\Sigma X^2}{N} - M_x^2}$$

This method of computing the standard deviation is, of course, not restricted to the present application.

from the arbitrary reference point (A.R.$_y$) in the Y distribution. $\Sigma x'$ means "the sum of the x''s," $\Sigma y'$ "the sum of the y''s," $\Sigma x'y'$ "the sum of the $x'y'$ products," and $\Sigma x'^2$ and $\Sigma y'^2$ "the sums of the squared x''s and the squared y''s" respectively.[1]

The application of this formula may be greatly facilitated by employing a specially prepared "correlation chart." There are many such printed forms available. A copy of the one recommended for use with this text is attached to the inside back cover of this book (Figure 17). The directions for the use of this chart are given on pages 169 to 174. The statements in brackets illustrate the application of this procedure in computing the coefficient of correlation between the scores given in Table 20, as shown in Figure 17.

Directions for Using the Correlation Chart

1. Let one of the sets of measures to be correlated be known as the X series, the other as the Y series.

   [In the illustrative problem, the E.T. scores are taken as the X series, the M.A. scores as the Y series.]

2. Find the range of the measures in the Y series. Determine the appropriate interval for grouping these measures, as explained in Chapter II. (Be sure not to let the number of intervals exceed 21.) Write the integral limits of the intervals in the extreme left-hand column on the chart just as you would write them in the score column of a frequency distribution. Try to arrange the intervals on the scale so that the interval most likely to contain the mean of the Y measures will fall between the heavy lines in the middle of the chart.

   [An interval of 2 is used for the M.A. scores, and the integral limits are written in the column at the extreme left of the chart (see Figure 17, on the inside back cover), leaving two blank rows at the bottom and one at the top of the chart.]

[1] While Formula (23) may be derived from Formula (21) without the use of complicated mathematics, the student of elementary statistics is again advised to take this formula for granted and to be content with the assurance that it is the exact algebraic equivalent of Formula (21).

[2] Copies of this chart may be obtained from Houghton Mifflin Company.

TABLE 20

SCORES MADE ON THE ENGLISH TRAINING (E.T.) AND MATHEMATICS APTITUDE (M.A.) TESTS OF THE IOWA PLACEMENT EXAMINATIONS BY 50 UNIVERSITY OF IOWA FRESHMEN

Student Number	Score (E.T.) X	Score (M.A.) Y	Student Number	Score (E.T.) X	Score (M.A.) Y
1	71	17	26	125	15
2	79	33	27	137	17
3	122	10	28	86	21
4	92	19	29	151	14
5	99	33	30	137	23
6	76	9	31	46	11
7	74	14	32	100	23
8	129	9	33	157	27
9	56	11	34	153	21
10	84	29	35	63	16
11	163	22	36	74	23
12	50	11	37	151	27
13	117	24	38	58	14
14	129	42	39	101	31
15	64	26	40	91	14
16	25	10	41	86	43
17	45	35	42	54	19
18	111	18	43	55	14
19	28	13	44	81	9
20	99	32	45	80	30
21	150	30	46	115	26
22	61	12	47	42	14
23	60	11	48	127	38
24	132	31	49	89	19
25	72	21	50	113	20

3. Now find the range of the X series. Determine the interval and write the integral limits of the intervals (beginning with the lowest) from left to right along the upper row on the chart. Again arrange the intervals so that the one most likely to contain the mean will fall at the center of the chart.

   [The range of the X series is $163 - 25 = 138$, and an interval of 10 units is employed. The integral limits 20–29, 30–39, etc., are written in the row at the top of the chart. The limits are so entered that the middle of the range comes between the heavy vertical lines, leaving three blank columns at each end of the chart.]

4. Now we are ready to tabulate the measures. Each tally mark on the chart is to denote two things: Its vertical

position on the chart will denote the Y measure of a pair; its horizontal position, the X measure of the same pair. There will be one tally mark for each pair of measures or for each individual. Begin with the first pair in the list. Find the interval along the Y scale in which the Y measure of the pair will fall. The tally mark for the first individual will fall somewhere in the horizontal row determined by this interval. Now go along this row to the right until you come to the vertical column corresponding to the interval along the X scale in which the X measure of the pair will fall. Place a tally mark at this point. The tally mark for the first individual will then fall in the horizontal row determined by his Y measure and in the vertical column determined by his X measure. In the same way locate the tally mark for the second individual, for the third, etc., until you have made a tally mark on the chart for each individual (or for each pair of measures) in the group.

[The first pair of measures in Table 20 is 17 (Y) and 71 (X). The tally mark for this individual will therefore lie in the 16-17 row and in the 70–79 column, that is, in the ninth cell from the left in the row labeled 16–17. Similarly, the tally mark for the second individual is placed in the ninth cell of the row labeled 32–33, and that for the third individual in the fourteenth cell in the row labeled 10–11.]

5. Now count the number of tally marks in each horizontal row and place the result for each row in the f column at the right of the chart. Next count the total number of tally marks in each vertical column and place the result for each column in the f row at the bottom of the chart.

[The number of tally marks in the 42–43 row is 2, in the 38–39 row is 1, in the 34–35 row is 1, etc. The numbers are entered in the f column at the right of the chart. Similarly, the number of tally marks in the 20–29 column is 2, in the 40–49 column is 3, etc., as is shown in the f row at the bottom of the chart.]

6. Total the frequencies in the f column and write the result in the square at the bottom of the column. Also total the frequencies in the f row at the bottom of the chart. The two sums should agree and should be equal to N, the total number of cases.

   [The sum of the frequencies in the *f column* is 50, which checks with the sum of the frequencies in the *f row*.]

7. Multiply each frequency in the f column by its corresponding deviation in the d column. Write the product in each case in the adjoining y' column. The y' column here corresponds to the fd column in an ordinary frequency distribution. Do the same for the frequencies on the scale at the bottom of the chart, writing the products in the x' row.

   [In the *y′ column* at the right of the chart the numbers entered are $2 \times 9 = 18$, $1 \times 7 = 7$, $1 \times 5 = 5$, $3 \times 4 = 12$, etc. Similarly, in the *x′ row* at the bottom of the chart, $2 \times -7 = -14$, $3 \times -5 = -15$, etc.]

8. Multiply each product in the y' column by its corresponding deviation in the d column. Write the resulting product in the y'^2 column. The y'^2 column here corresponds to the fd^2 column in an ordinary frequency distribution. Repeat the process for the scale at the bottom of the chart, writing the products in the x'^2 row.

   [In the y'^2 column, $9 \times 18 = 162$, $7 \times 7 = 49$, etc. Similarly, in the x'^2 row, $7 \times -14 = -98$, $5 \times -15 = -75$, etc.]

9. Find the algebraic sum (taking account of signs) of the numbers in the y' column, and write the sum in the cell at the bottom of the column. The sum is denoted in the formulas by $\Sigma y'$. Also total the numbers in the y'^2 column and write the result at the bottom of the column. The symbol for this sum is $\Sigma y'^2$. In the same way find $\Sigma x'$ and $\Sigma x'^2$ along the bottom of the chart.

   [The algebraic sum of the y' column is $\Sigma y' = -89$. The algebraic sum of the x' row is $\Sigma x' = -5$. Similarly $\Sigma y'^2 = 1165$ and $\Sigma x'^2 = 659$.]

10. Now multiply the small number in the upper right-hand corner of each cell by the number of tally marks in the cell, and write the result in the upper left-hand corner of the cell. This result is the "product-moment" for the frequencies in the cell. Find this product for each cell that contains a tally mark, taking account of signs.

    [For example, there are two frequencies in the cell common to the 14–15 row and the 50–59 column. The small number in this cell is 20, hence the "product-moment" for this cell is $2 \times 20 = 40$.]

11. Now find the sum of all positive product-moments in each horizontal row and write the result in the $(+)\, x'y'$ column. Then find the sum of all negative product-moments in each row and write the result in the $(-)\, x'y'$ column.

    [For example, the sum of the positive moments in row 14–15 is $25 + 40 + 10 = 75$, and the sum of the negative moments is $15 + 30 = 45$.]

12. Total the $(-)$ and $(+)$ $x'y'$ columns and find the algebraic sum of the results. This final sum is denoted in the formulas by $\Sigma x'y'$.

    [The sum of the numbers in the $+ x'y'$ column is 458, and in the $- x'y'$ column is 190. The total of the $x'y'$ products is $458 - 190 = 268$. Hence, $\Sigma x'y' = 268$.]

13. Now you have found the values of $\Sigma x'$, $\Sigma y'$, $\Sigma x'^2$, $\Sigma y'^2$, and of $\Sigma x'y'$. Each of these values must be divided by N to give

$$\frac{\Sigma x'}{N}, \frac{\Sigma y'}{N}, \frac{\Sigma x'^2}{N}, \frac{\Sigma y'^2}{N}, \text{ and } \frac{\Sigma x'y'}{N}$$

    Spaces for the computation of these values are provided along the right-hand margin of the chart.

$$\left[\frac{\Sigma x'}{N} = \frac{-5}{50} = -.1; \quad \frac{\Sigma x'^2}{N} = \frac{659}{50} = 13.18;\right.$$

$$\frac{\Sigma y'}{N} = \frac{-89}{50} = -1.78; \quad \frac{\Sigma y'^2}{N} = \frac{1165}{50} = 23.30;$$

$$\left.\frac{\Sigma x'y'}{N} = \frac{268}{50} = 5.36\right]$$

14. Now square the values of $\left(\frac{\Sigma x'}{N}\right)$ and $\left(\frac{\Sigma y'}{N}\right)$ to give the values of $\left(\frac{\Sigma x'}{N}\right)^2$ and $\left(\frac{\Sigma y'}{N}\right)^2$, and write the results in the appropriate spaces in the right-hand margin also.

$$\left[\left(\frac{\Sigma x'}{N}\right)^2 = (-.10)^2 = .01; \quad \left(\frac{\Sigma y'}{N}\right)^2 = (-1.78)^2 = 3.17\right]$$

15. Now compute the values of the standard deviations, using the formulas given in the middle of the right-hand margin of the chart.

$$\left[\begin{aligned} \sigma_x &= \sqrt{13.18 - .01} = \sqrt{13.17} = 3.63 \\ \sigma_y &= \sqrt{23.30 - 3.17} = \sqrt{20.13} = 4.47 \end{aligned}\right]$$

Note: It is very important to note that these standard deviations are expressed in units of the interval and not in raw score values. The S.D.'s should be expressed in interval units in the formula for r given on the chart, but if the S.D.'s in raw *score* units are needed for any other purpose, the values here obtained must be multiplied by the size of the interval in each case.

[The S.D. of the X measures is 3.63, expressed in interval units. Since an interval of 10 was used along the X scale, the S.D. of the English Training scores is $3.63 \times 10 = 36.3$ in raw score units. The S.D. of the Mathematics Aptitude scores is similarly $4.47 \times 2 = 8.94$.]

16. You will then have all the values needed for substitution in the formula to give the value of r — the coefficient of correlation between the paired series. Substitute these values in the formula, reduce the expression to a simple decimal number, and write the result as the value of r.

$$\left[r = \frac{5.36 - (-.10)(-1.78)}{3.63 \times 4.47} = \frac{5.36 - .178}{3.63 \times 4.47} = \frac{5.182}{16.23} = .32\right]$$

While the student of statistics must necessarily become acquainted with economical procedures for computing r from raw score data, he is strongly advised to make no attempt to base his *interpretation* of r upon the relatively complicated formulas, such as Formulas (22) and (23), which are used in such computation. These formulas are extremely difficult to interpret directly and are likely to create the false impression that the correlation coefficient is a much more complicated concept than it really is. The Pearson product-moment coefficient of correlation is nothing more than a simple average (mean) of a number of z-score products, and the student should do all of his thinking about the correlation coefficient in terms of this relatively simple z-score definition, or in terms of Formula (21).

The Phenomenon of Regression

An interesting characteristic of the frequency distributions of measures of two rectilinearly related traits for any group of individuals is the fact that if, from the total group, a number of individuals are selected all of whom are exactly alike with reference to the first trait, these individuals will, on the average, lie closer to the general average of the second trait than they do to the general average of the first. Suppose, for example, we consider the relationship between height and weight for any large group of adults. If, from the total group, we selected a number of persons all of whom were 6 feet 3 inches tall, we would find these individuals, on the average, less extreme in weight than in height. Similarly, if we selected from the total group those individuals who were, say, 275 pounds in weight, we would not expect the average individual in this group to be as extreme in height as in weight. A few of these heavyweight individuals would also be unusually tall, but many of them would be only moderately tall or even below the general average in height. Again, if from the total group we selected a number of individuals all of whom were extremely short, we would find that on the average these individuals would be less extreme in weight than in height, since

many of these short persons would be near or above the general average in weight.

Again, suppose that a test in general mathematics and one in world history were administered to all freshmen in a given university. If from the total group we selected a number of individuals who were outstanding in their performance on the mathematics test, we would find that, while most of these individuals would also be above average on the history test, only a few of them would be as far above average in history as in mathematics. For these selected individuals the mean score on the history test would be lower (when expressed in comparable terms, such as z-scores) than their scores on the mathematics test. Similarly, most individuals making very low scores on the history test would make better scores on the mathematics test.

This phenomenon is one which we have all noticed, but which we have seldom expressed in quantitative terms or referred to as the "phenomenon of regression." We have all observed, however, that individuals selected because they show a certain degree of superiority in one trait (whether the superiority is marked or slight) are seldom equally superior in other related traits, and that individuals inferior in one trait are seldom equally inferior in others.

A graphic representation of this phenomenon will be helpful in arriving at a more exact understanding of its character. The two frequency curves in Figure 18 represent the distributions of measures of height and weight for the same group of adults. The X distribution represents the distribution of height; the Y distribution, that of weight. Both distributions are plotted along comparable (z-score) scales. In the height distribution there has been marked off a shaded area including individuals who are between two high values on the height scale. Let us suppose that there are 12 individuals in this interval. For each of these individuals there has been located the position of his weight in the distribution of weights, and a line has been drawn from the midpoint of the shaded interval to each of these positions. We note

that three of these individuals are higher up along the weight scale than along the height scale, but that the others are less extreme in weight than in height. This is consistent with what we have already observed about the nature of the relationship between height and weight. The mean weight of these 12 individuals is indicated by the point M'_y, and a heavy line has been drawn from the shaded area to this point. We note that this line points inward toward the middle of the weight distribution, that is, that the mean weight of these 12 individuals is only about half as far from the general mean, M_y, of the weight distribution

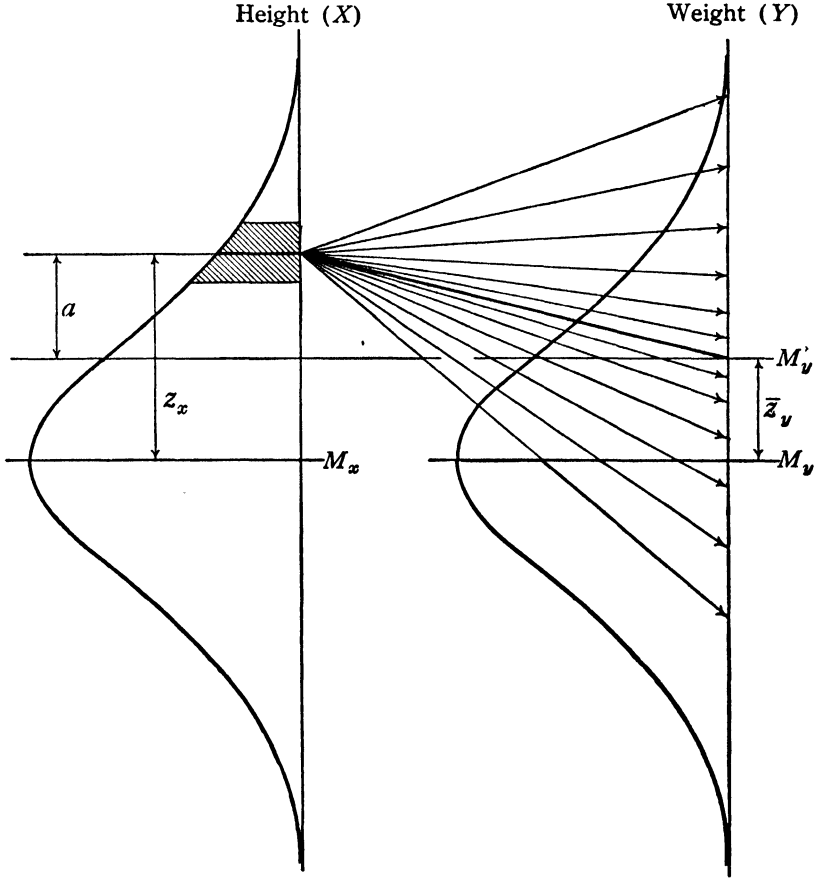


FIG. 18.

Illustrating the phenomenon of regression in terms of the distributions of height and weight for a given sample.

as the shaded interval is from the general mean, M_x, of the height distribution. In other words, $\bar{z}_y$ is less than z_x.

This picture suggests what would be found in the distributions of any two positively related traits for any group. If the relationship between the two traits were perfect, then the lines from any interval in the left-hand distribution would all run exactly horizontally across to the other distribution, that is, every measure in the one distribution would be paired with another with the *same* relative status in the second distribution. If the relationship were high but not perfect, these lines would spread apart, but would form a relatively narrow "fan," and the heavy line (to the mean value of the second variable for the selected group) would be deflected only slightly toward the middle of the second distribution. If the relationship were very low but positive, the lines would spread to nearly all parts of the second distribution, and the heavy line would point more sharply into the middle of that distribution. If the traits were entirely unrelated, the lines would spread throughout the whole of the second distribution, and the mean of the selected cases (M'_y) would coincide with the mean of the entire unselected group. For example, if height and intelligence were the measures concerned, and if lines were drawn from an interval near the lower end of the height scale to the positions of the corresponding measures on the intelligence scale, these lines would spread throughout the entire intelligence distribution around a mean which coincided with the general mean in intelligence. This is the same as saying that short persons are just as variable in intelligence and show the same mean intelligence as tall persons, or as the persons in a group of individuals of differing heights.

If the relationship were negative, the majority of the lines from any one interval in the first distribution would go to the *opposite* half of the second distribution, as would the heavy line (at the mean of the selected cases).

In general, then, the higher the degree of correlation, the narrower will be the fan-shaped pattern of lines drawn from scores in

a certain interval of one distribution to the corresponding measures in the other, and the more nearly horizontal will be the heavy line drawn to the mean of these selected cases. As long as the relationship is not perfect, this heavy line will point inward, however slightly. The lower the correlation, the wider will be the spread of these lines, and the more nearly will the heavy line point to the middle or general mean of the second distribution. In other words, for individuals selected from a given group because they are alike in one trait, the *mean* value of a second related trait will *regress* toward the general mean of the second trait for the entire group. The amount of this regression (equal to the distance *a* in Figure 18) can be shown to be inversely related to the coefficient of correlation between the two measures. With perfect correlation, there is no regression. With zero correlation, the regression is complete, that is, the mean of the selected cases will coincide with the general mean of the second distribution and $\bar{z}_y$ will become zero.

A more complete understanding of the exact nature of the phenomenon of regression may perhaps be acquired by considering further just what it means graphically in terms of the scatter-diagram showing the relationship between two sets of measures. The oval in Figure 19 represents the pattern of the distribution of tally marks on a scatter-diagram showing the relationship between height and weight for a given group of individuals. For the sake of simplicity in illustration, the tally marks themselves have been omitted from the chart, but the student can visualize them as being distributed over the area included inside this oval. This, then, would represent a case of fairly high positive relationship. The scales employed in plotting this figure are the *z*-score scales corresponding to the original scales of height and weight. These scales have been used so that the deviations from the mean along either scale may be directly compared. Again, for the sake of simplicity, only one column (*C*) and one row (*R*) are shown. The point *A* represents the mean weight of the individuals tallied in column *C*, that is, of a group of individuals all of the same height

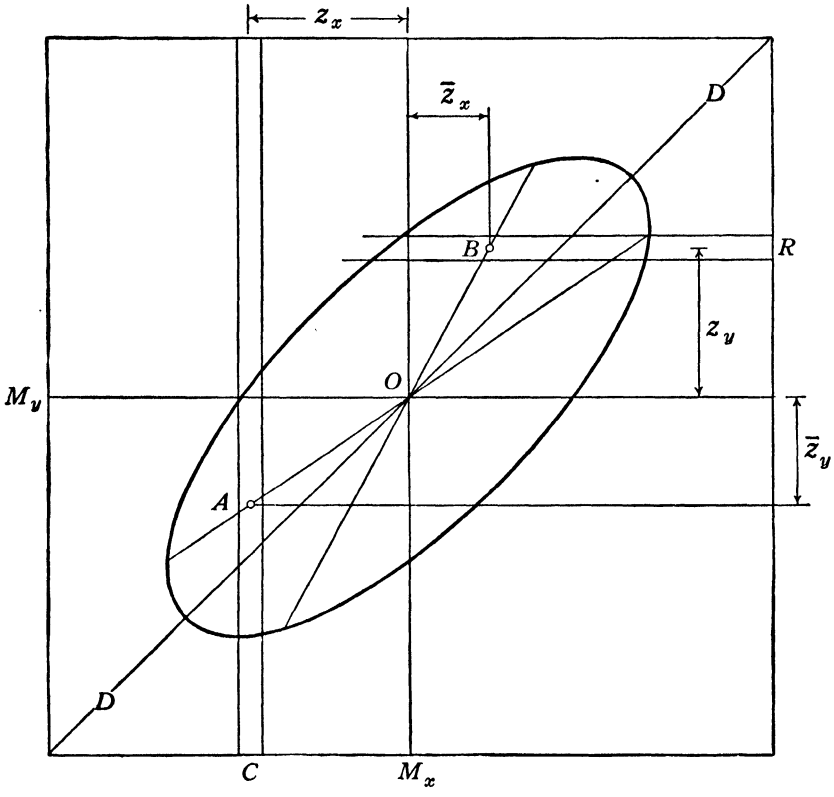


FIG. 19.
Illustrating the phenomenon of regression in terms of the scatter-diagram of height and weight measures for a given sample.

and all of whom deviate from the general mean (M_x) of the height distribution by an amount equal to z_x (at top of chart). It will now be noted that the point A lies closer to the horizontal line M_y representing the general mean in weight than it does to the vertical line M_x representing the general mean in height. (This is apparent because the point A is above the diagonal DD, which is equidistant from the two axes.) In other words, individuals all of the same height are, on the average, nearer ($\bar{z}_y$) to the general mean in weight than they are (z_x) to the general mean in height, that is, $\bar{z}_y$ is less than z_x. (Do not confuse z_x with $\bar{z}_x$, or z_y with $\bar{z}_y$). Still another way of saying this is to say that the ratio

$\bar{z}_y/z_x$ is less than 1.00, since the numerator in this expression is smaller than the denominator. This ratio, furthermore, may be shown to be proportional to the coefficient of correlation, that is:

$$\frac{\bar{z}_y}{z_x} = r_{xy}$$

If the relationship were perfect, then all of the tally marks would lie along the diagonal *DD*, and the point *A* would therefore lie on this diagonal and be equidistant from the two axes. In this case, $\bar{z}_y$ would be equal to z_x, and the ratio $\bar{z}_y/z_x$ would be equal to 1.00, as would the coefficient of correlation. If the variables were unrelated, then all of the tally marks would be considered as lying within a circle with center at *O*. The point *A* would then lie on the line M_y and $\bar{z}_y$ would be equal to zero. The ratio $\bar{z}_y/z_x$ would then be equal to zero, as would the coefficient of correlation.

Similarly, if we consider only the individuals in a single row *R*, all of whom are of the same weight (z_y), we find that their mean height (represented by point *B*) lies closer to M_x than to M_y, that is, we find that $\bar{z}_x$ is less than z_y. Again, the ratio between $\bar{z}_x$ and z_y would be proportional to the coefficient of correlation, that is:

$$r_{xy} = \frac{\bar{z}_x}{\bar{z}_y}$$

On any chart of this kind, representing a positive rectilinear relationship between two variables, the mean, *A*, of *any* column would lie closer to the horizontal than to the vertical axis. The means of the other columns in Figure 19 (if they were shown) would lie approximately along a straight line drawn through *A* and *O*. Similarly, the means of the other rows would lie near the line drawn through *B* and *O*. These two lines are known as the "regression lines." If the relationship were perfect, these two lines would coincide along the diagonal *DD*. The amount by which they would diverge would depend upon the degree of relationship. If the relationship were zero, they would coincide with the vertical and horizontal axes, and would therefore be at right angles to one another. (Note that these statements apply only

when both measures are plotted along comparable z-score scales.)

We have already noted that

$$\frac{\bar{z}_y}{z_x} = \frac{\bar{z}_x}{z_y} = r_{xy}$$

Hence

$$\bar{z}_x = r_{xy} z_y \tag{24}$$

in which z_y is any given z-score in the Y distribution, and $\bar{z}_x$ is the *mean* z-score in the X distribution for those individuals with the given z-score in the Y distribution. Similarly,

$$\bar{z}_y = r_{xy} z_x \tag{25}$$

in which z_x is any given z-score in the X distribution, and $\bar{z}_y$ is the mean of the corresponding z-scores in the Y distribution.

These equations (24 and 25) are known as the "regression equations" (in z-score form). Their application may be illustrated as follows: Suppose that the coefficient of correlation between height and weight for a given group of adults is $r = .6$. If from the total group we selected all individuals with a z-score of 1.4 in the height distribution, we would find that their mean z-score in the weight distribution would be $1.4 \times .6 = .84$. Again, if from the weight distribution we selected all individuals with a z-score of -2.0, we would find that their mean z-score in the height distribution would be -1.2. If a correlation of $-.75$ existed between two sets of measures, the individuals who were 1.4 of a standard deviation above average in one distribution would, on the average, be $1.4 \times -.75 = (-)\ 1.05$ of a standard deviation *below* (because of negative sign) the general mean of the other distribution. These regression equations, then, are simply a way of saying in algebraic language that the amount of regression is dependent upon the degree of relationship.

The Use of the Regression Equations in Prediction

The significance of the regression equations for practical work in education and psychology lies in the fact that they constitute an objective means of estimating the value of one variable for an individual when the value of another related variable is known for

that individual, and when the degree of correlation between the two variables is known for the group to which he belongs. If, for example, we know that an individual's height is 1.2 S.D.'s above the mean of the group to which he belongs, and if we know that the coefficient of correlation between height and weight for that group is .5, and if the relationship is rectilinear, then the best estimate that we can make of his z-score in weight is $.5 \times 1.2 = .6$, because that is the mean z-score in weight of all individuals (from the total group) who are of the same height as the given individual.

Again, suppose that a test of scholastic aptitude was given a year ago last fall to all entering freshmen in a certain university and that at the end of the academic year it was discovered that a correlation of .7 existed between the scores on this examination and the grade-point averages earned by the students during their freshman year. Suppose that in the fall of the present academic year the same test was administered to the entering freshmen and that on this test a certain freshman earned a z-score of + 1.8. Let us now assume that the frequency distributions of test scores and of grade-point averages for the present freshman class will each show the same central tendency and variability as the corresponding distribution for the preceding class, and that the coefficient of correlation between these test scores and grade-point averages will be approximately the same this year as it was last year. According to the principle of regression, the individuals who make a z-score of + 1.8 on the scholastic aptitude test this year will, on the average, earn a grade-point average that is 1.26 S.D.'s above the general mean of the distribution of grade-point averages for the entire class ($.7 \times 1.8$). Accordingly, the best prediction that we can make for any one of these individuals is that he will make a grade-point average with a z-score equivalent of + 1.26.

One further illustration of this use of the regression equations might be considered here, this time in the field of industrial psychology. Suppose that the sales organization of a large corpora-

tion had followed the practice of administering to each applicant for a position as salesman a test of "salesmanship ability." Suppose that on the basis of past experience it had been shown that there was a coefficient of correlation of .8 between the scores on this examination and the sales records later made by the applicants. The sales manager could then use the regression equations to predict, at the time that application is made, how well any applicant will succeed on the job, and could select his new salesman from available applicants accordingly.

In general, then, whenever the coefficient of correlation between two related traits is known for a sample selected for a given population, if we know only the z-score equivalent of an individual in one of these traits, we can predict (by means of the regression equations) his probable status in the general distribution of the other trait.

The Raw Score Form of the Regression Equations

In practical work, when predictions are to be made of one variable from known values of a related variable, it is not convenient first to transform each of the known measures into its z-score equivalent, then to determine (by means of the regression equations) the expected value of the related z-score in the second trait, and then in turn to transform this estimated z-score into its equivalent raw score value. To save the time required for these transformations of known raw scores into z-scores and the estimated z-scores into raw scores, the predictions are usually computed directly from and expressed in raw score values. The raw score form of the regression equations may be derived by substituting in Formulas (23) and (24) the following equivalents of z_x and z_y.

$$z_x = \frac{X - M_x}{\sigma_x}$$

$$z_y = \frac{Y - M_y}{\sigma_y}$$

Upon substitution of these values and simplification and transposition of terms, Formulas (24) and (25) become:

$$X = r_{xy}\frac{\sigma_x}{\sigma_y}(Y - M_y) + M_x \tag{26}$$

$$Y = r_{xy}\frac{\sigma_y}{\sigma_x}(X - M_x) + M_y \tag{27}$$

The following examples will illustrate how these formulas are applied. Suppose that to a given number or random sample of individuals in a given school population a test of general intelligence and a test of silent reading comprehension are administered. Let us refer to the intelligence test as the X test, and to the reading comprehension test as the Y test. Suppose that the following measures are derived from the distributions of scores on these two tests:

$$M_x = 102;\ \sigma_x = 12;\ M_y = 80;\ \sigma_y = 9;\ r_{xy} = .8.$$

Suppose then that some other individual from the population in question made a score of 126 on the intelligence test, but that he did not take the reading test, and that we wish to estimate as accurately as possible what score he would be likely to make on it. We could do this by substituting the given values in Formula (27) as follows:

$$Y = .8 \times \frac{9}{12}(126 - 102) + 80 = 94.4, \text{ or } 94 \text{ (rounded value)}$$

If similar predictions were to be made for a large number of individuals, it would be more convenient to reduce the general expression

$$Y = .8 \times \frac{9}{12}(X - 102) + 80$$

to the simpler form

$$Y = .6\,X + 18.8.$$

Using this expression, then, if an individual were known to have made a score of 90 on the intelligence test, we would estimate for him a score of $.6 \times 90 + 18.8 = 72.8$ on the reading test.

Estimates of probable X scores could similarly be made from known Y scores, using Formula (26).

The Reliability of Prediction; The Standard Error of Estimate

In the preceding sections we have seen that when the regression equations for two related variables have been determined for a given group we can, by means of these equations, estimate for any individual the probable value of one variable from a known value of the other. Predictions based upon these regression equations, however, are never (except in the case of perfect correlation) perfectly reliable. These equations only indicate, for individuals with a given measure of one trait, what is the *mean* of their measures in a second trait. The *actual* measures of the second trait for these individuals are scattered on either side of this mean, so that the estimate of the second trait for any particular individual would seldom coincide with his actual measure of that trait. For example, if it is known that the mean weight of persons 6 feet tall is 150 pounds, then 150 pounds is the best *estimate* of the weight of any particular six-footer, but we would know that his actual weight would probably differ considerably from this estimate. The reliability of this estimate, then, would depend upon the variability in weight for six-footers in general. If the actual weights of six-footers in general were known to cluster closely around the mean of 150 pounds, then this mean would be a close approximation to the weight of any six-footer. If, on the other hand, the actual weights of six-footers in general were known to show a wide spread on either side of 150 pounds, then we could not consider this mean as a dependable or reliable estimate of the weight of any single individual of this height; that is, the actual weight of the particular six-footer involved would be likely to differ considerably from this estimate. The reliability of any prediction of this kind would, then, depend upon the variability in weight for persons of the same height. More specifically, the reliability of these estimates would depend upon, or would be measured by, the *standard deviation* of weights for persons of the same height

When thus used to describe the reliability of prediction, this S.D. of the one variable for individuals with a given value of another variable is known as the *standard error of estimate.* Accordingly, the standard error of estimate in predicting reading test scores from arithmetic test scores for a given group would be the S.D. in reading test scores for individuals (selected from the given group) all of whom have the same arithmetic score.

Let us now consider some of the more important characteristics of this standard error of estimate. Suppose that we have measures of two rectilinearly related traits, A and B, for a given group of individuals, that σ_A and σ_B represent the S.D.'s of distributions of the measures for the entire group, and that r_{AB} represents the coefficient of correlation between these measures for this group. Suppose further that from the total group we select a group of individuals all of whom have the same measure of trait A, that we make up a frequency distribution of the B measures for these selected individuals, and that we compute the S.D. of this distribution. Let this S.D. be represented by $\sigma_{B \cdot A}$. This expression may be variously interpreted as "the S.D. of the B measures which are paired with a given value of A," or "the S.D. of B when A is held constant," or "the standard error of estimating B from A."

One interesting characteristic of $\sigma_{B \cdot A}$ is that its value, in most cases of rectilinear relationship, is independent of the given value of A, that is, of the value at which A is held constant. For example, the S.D. of weights for adult individuals of a given height is about the same regardless of the value of the given height. A group of six-footers will show about the same variability in weight as a group of five-footers. This is again equivalent to saying that in the scatter-diagram the variability of the measures in any one column is about the same as the variability of the measures in any other column on the chart. Or again, it is equivalent to saying, with reference to Figure 18, that the "spread" of the lines drawn from any point in the left-hand distribution is about the same regardless of the position of the point in that distribution from which the lines were drawn. In all of our subsequent discussions

of the standard error of estimate, we will assume that the relationship between the two variables is of this character. (This assumption is frequently known as the assumption of *homoscedacity.*)

Another characteristic of $\sigma_{B \cdot A}$ is that it is always less than σ_B unless, of course, the traits are entirely unrelated, in which case $\sigma_{B \cdot A}$ is equal to σ_B. This is only equivalent to saying that individuals alike with reference to one trait will be more alike with reference to a related trait than will individuals with differing measures of the first trait. Again, it is equivalent to saying, with reference to Figure 18, that the fan-shaped pattern of lines will not spread throughout the whole distribution, or, with reference to the scatter-diagram, that the S.D. of measures in any one column is less than the S.D. of measures for all columns combined.

Another significant characteristic of $\sigma_{B \cdot A}$ is the fact that its ratio to σ_B, that is, the ratio $\frac{\sigma_{B \cdot A}}{\sigma_B}$, will be large if the correlation between A and B is low, and will be small if the correlation is high. It can be shown that this ratio bears a definite relationship to r_{AB}, as follows:

$$\frac{\sigma_{B \cdot A}}{\sigma_B} = \sqrt{1 - r_{AB}^2} \tag{28}$$

Similarly:

$$\frac{\sigma_{A \cdot B}}{\sigma_A} = \sqrt{1 - r_{AB}^2} \tag{29}$$

We may note at once that these algebraic expressions are entirely consistent with what has just been said. If r_{AB} equals zero, then the ratio $\frac{\sigma_{B \cdot A}}{\sigma_B}$ becomes equal to 1.00, that is, $\sigma_{B \cdot A}$ becomes equal to σ_B. If r_{AB} equals 1.00, then $\frac{\sigma_{B \cdot A}}{\sigma_B}$ equals zero, which means that $\sigma_{B \cdot A}$ equals zero, or that individuals alike in A are also all exactly alike in B. Similarly, if $r_{AB} = .8$, then $\frac{\sigma_{B \cdot A}}{\sigma_B} = \sqrt{1 - .8^2} = .6$, which means that $\sigma_{B \cdot A}$ is .6 or 60 per cent as large as σ_B.

Muitiplying both sides of Equation (28) by σ_B and canceling σ_B in the left-hand term, we get the following formula for the standard error of estimate (in estimating B from A).

$$\sigma_{B \cdot A} = \sigma_B \sqrt{1 - r_{AB}^2} \tag{30}$$

Similarly, the standard error of estimating A from B is

$$\sigma_{A \cdot B} = \sigma_A \sqrt{1 - r_{AB}^2} \tag{31}$$

The following illustrations will indicate how these formulas are applied. Suppose that for a given group of adults the coefficient of correlation between height and weight is $r_{HW} = .6$, that $\sigma_H = 3$ inches, $\sigma_W = 12$ lbs., that the means in height and weight for the entire group are 69 inches and 145 lbs., respectively. By means of the raw score form of the regression equation (Equation 26 or 27), it could then be shown that the best estimate of the weight of an individual who is 72 inches tall would be 152.2 lbs. The reliability of this estimate, as we have already noted, would depend upon the variability (standard deviation) in weight of all persons (in the given group) who are 72 inches tall. This standard deviation according to Formula (30) or (31) is

$$\sigma_{W \cdot H} = \sigma_W \sqrt{1 - r_{WH}^2} = 12 \sqrt{1 - .6^2} = 9.6 \text{ lbs.}$$

This standard error of estimate may be interpreted in much the same fashion as we have previously interpreted the standard error of the mean in sampling error theory. Assuming that the distribution of weights for a large number of persons 72 inches tall would approximate the form of the normal curve, we may say that approximately 68 per cent of all persons of this height will be within 9.6 lbs. of the mean weight (152.2 lbs) of all persons of this height. We may say, then, that the chances are about 68 out of 100 that the actual weight of any one individual of this height is within 9.6 lbs. of the best estimate (152.2 lbs.) that we can make of his weight. Again applying the known area relationships under the normal curve, we may say that the chances are 95 out of 100 that the actual weight of this individual is within two standard errors of the estimate; that is, the chances are about 95 out of 100 that

his actual weight is within 19.2 lbs. of 152.2 lbs., or that it is between 171.4 lbs. and 133.0 lbs. Again we may say that we are "practically certain" that his actual weight is within three standard errors of his estimated weight, or that we are practically certain that his actual weight lies somewhere between 180.8 and 123.2 lbs. [$152 + (3 \times 9.6)$ and $152 - (3 \times 9.6)$].

The Assumption of Rectilinearity

In the foregoing discussions, attention has been repeatedly drawn to the fact that the Pearson product-moment correlation coefficient, as well as the regression equations and standard errors of estimate based upon it, is intended for use only with measures that are *rectilinearly* related. This fact deserves greater emphasis than it has yet been given. If the relationship between two sets of measures departs markedly from rectilinearity, r not only becomes a poor measure of the degree of relationship, but its use may even lead to serious misinterpretations. Instances may be found, for example, in which $r = 0$ but in which the measures are nevertheless very closely or even perfectly related. Fortunately for the student of statistics, instances of data showing markedly curvilinear relationship are relatively rare in educational and psychological research, so rare that it is hardly worth while to burden the beginning student with any consideration of the special correlation methods that are available for the treatment of such data. It is sufficient for him to know that such methods do exist and may be referred to if the occasion demands. It is extremely important, however, that in all instances in which he makes use of r or of the regression equations based upon it, the student demonstrate conclusively that the relationship involved is at least approximately rectilinear in form. Certain mathematical tests of rectilinearity or curvilinearity are available, but the application of such tests is rarely necessary in practical work and need not be considered here. The most practicable test of rectilinearity is that based simply upon an inspection of the scatter-diagram. If the curvilinearity is not so marked that it is not immediately

apparent upon inspection of the scatter-diagram, the student need have no fear that the use of product-moment correlation techniques will lead to any serious error. It is well, therefore, to construct a scatter-diagram whenever a correlation coefficient is to be computed, even though the scatter-diagram is not needed in the computation, as when certain machine methods of computation are employed.

Sampling Errors in r

If two samples of the same size were selected strictly at random from the same population and a scatter-diagram showing the relationship between the measures of two given traits were prepared for each sample, we would almost invariably find that the distribution of the tally marks would not be exactly the same for both samples. The operation of chance in selecting the individuals constituting each sample would practically guarantee differences in these scatter-diagrams. The sum of the z-score products, and hence the value of r, would therefore differ somewhat for the two samples. The r for either sample could therefore not be taken as a perfectly reliable indication of the r that would be obtained if the entire population were considered. If a large number of random samples of the same size were selected from the population, very few samples would show the same r as any other. These obtained r's would be distributed on either side of the true r, and (if the true r were not too high) the form of this distribution would be that of the normal curve. The standard deviation of this distribution may be shown to be

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}} \tag{32}$$

in which the r represents the *true* r for the population and N the number of cases in each sample. If this standard deviation is small, that is, if all obtained r's cluster closely around the *true* r, then the r obtained from any single sample is, of course, unlikely to deviate far from the true r, and may be accepted as a close approximation to it. If this standard deviation is large, then

the r for any one sample is likely to differ considerably from the true r, that is, it is likely to contain a large sampling error and must be considered as unreliable. This standard deviation is therefore a measure of the reliability of the r for a single sample, and is known as the "standard error of r." For a large sample drawn from a population in which the true r does not closely approach 1.00, the sampling distribution of r is approximately normal. When the true r is high, the distribution of obtained r's is markedly skewed, even though the samples are large. The reason for this is readily apparent. Suppose, for example, that a large number of random samples are drawn at random from a population for which the true r is .96, and that the obtained r is independently computed for each sample. Obviously, none of the sample r's could deviate from the true r by more than .04 in one direction, while sampling errors very much larger than this could readily occur in the other direction. The result would be a sampling distribution markedly skewed to the left — the smaller the sample, the more extreme the skew.

Accordingly, it is only for large random samples in which the obtained r is low or only moderately high that one may interpret the standard error of r with the aid of Table 17. With what maximum value of r the standard error may safely be thus interpreted depends upon the size of the sample. A safe rule to follow is never to use Formula (32) at all with small samples (say, $N < 60$), and to use it with large samples only if the obtained r is less than .80. (If the sample is very large, consisting of several hundred cases, the formula and Table 17 may perhaps be safely used for r's as large as .9.) Other techniques, beyond the scope of this course, for dealing with small samples and high r's are elsewhere available.[1]

For samples that satisfy the preceding conditions, the standard error of r may be interpreted in much the same fashion as the standard error of the mean (see pages 106–123). Suppose, for example, that the correlation between x and y is $r_{xy} = .60$ for a

[1] See Lindquist, E. F., *Statistical Analysis in Educational Research*, pp. 210–218. Houghton Mifflin Company, 1940.

given random sample of 100 cases. To compute the standard error of this r by means of Formula (32), we should know the value of the true r for the population concerned. Not knowing this, we substitute for the true r in the formula the obtained r from our sample and thus secure a useful approximation [1] to the standard error desired, as follows:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}} = \frac{1 - .6^2}{\sqrt{100}} = .064$$

This means that if r's were similarly obtained from a large number of random samples of 100 cases each, these obtained r's would fall into an approximately normal distribution with a standard deviation of approximately .064. About 95 per cent of these r's would then lie within $.064 \times 1.96 = .125$ of the true r. Accordingly,[2] we may be confident at the 5 per cent level that the true r lies somewhere between $.60 - .125 = .475$ and $.60 + .125 = .725$. Similarly, the 2 per cent confidence interval for the true r is equal to $.60 \pm (0.64 \times 2.33)$ or .451 to .749, and the 1 per cent confidence interval is .435 to .765.

Suppose, now, that for a certain population the true r between x and y is zero. For any random sample drawn from this population, we could, nevertheless, hardly expect (because of chance fluctuations) that the sum of the positive z-score products would exactly cancel the sum of the negative z-score products. In other words, the r obtained from the sample would almost certainly differ from zero. For example, while the true correlation between height and intelligence for a given population of adults might be zero, in any particular random sample from this population we might by chance find that the tall individuals are, on the average, slightly more intelligent than the short individuals. In another

[1] Because of this substitution, this procedure will yield a close approximation to the standard error of the obtained r only when that r is itself highly reliable. Hence, this is an added reason for not using Formula (32) with small samples.

[2] Strictly, for reasons similar to those explained on page 127, the procedure here suggested is inexact and somewhat biased. An exact procedure for establishing confidence intervals for r's of any size and for samples of any size is described in Lindquist's *Statistical Analysis in Educational Research*, pp. 211–214.

sample, again by pure chance, the reverse might be found. The fact that a correlation not equal to zero is obtained for a *sample*, therefore, does not constitute conclusive evidence that a true r other than zero exists for the whole population. Before accepting an obtained r as evidence of a real relationship, we must show that it cannot reasonably be accounted for by chance fluctuations in random sampling.

If the true r for a given population is zero, then according to Formula (32) the S.D. of obtained r's for a large number of random samples of N cases each will be

$$\sigma_r = \frac{1 - 0^2}{\sqrt{N}} = \frac{1}{\sqrt{N}}$$

If, then, a sample of N cases is drawn at random from a certain population and the obtained r is found to exceed $2.58/\sqrt{N}$, we may be confident at the 1 per cent level that the true r is greater than zero, or that the obtained r does not represent a chance deviation from a true r of 0.00. In other words, an obtained r greater than $2.58/\sqrt{N}$ is significant at the 1 per cent level. Similarly, an r is significant at the 2 per cent level if it exceeds $2.33/\sqrt{N}$ or at the 5 per cent level if it exceeds $1.96/\sqrt{N}$.

Suppose, then, that we selected a sample of 100 cases from a population for which the true r between two variables is unknown and find that for this sample the obtained r is .24. We could then be confident at the 2 per cent level that the true r for this population is *not* zero; in other words, we could be confident at the 2 per cent level that there is *some* relationship between these variables as far as the entire population is concerned. On the other hand, if the obtained r had been .15, the hypothesis would be quite tenable that the true r for the population is 0.0, and that the obtained r of .15 is entirely due to chance fluctuations in random sampling.

For the convenience of the student in determining whether or not an obtained r is statistically significant and at what level, Table 21 has been prepared. This table shows the maximum

value of r that is required for significance at each of the three commonly employed levels of significance. For example, an r obtained from a sample of 125 cases must exceed .175 to be significant at the 5 per cent level, .208 to be significant at the 2 per cent level, and .230 to be significant at the 1 per cent level.

TABLE 21

MINIMUM VALUES OF CORRELATION COEFFICIENT REQUIRED FOR SIGNIFICANCE AT VARIOUS LEVELS FOR VARIOUS SIZES OF SAMPLES

Number of Cases (N)	Level of Confidence 5%	2%	1%
50	.277	.329	.364
60	.253	.300	.333
70	.234	.278	.308
80	.219	.260	.288
90	.207	.245	.271
100	.196	.233	.258
125	.175	.208	.230
150	.160	.190	.210
175	.148	.176	.195
200	.139	.164	.182
250	.124	.147	.163
300	.113	.134	.149
400	.098	.116	.129
500	.088	.104	.115
1000	.061	.071	.083

The significance of a difference between r's obtained from independent random samples may be tested in much the same fashion as a difference in means. If the samples are large and the r's are not high, a close approximation to the standard error of the difference may be obtained by substituting the standard errors of the separate r's in Formula (15). The ratio between the difference and its standard error may then be interpreted by means of Table 17. It is very important to note that this procedure is not valid, in general, if the r's being compared are both obtained from the *same* sample.

Influence of the Variability of Measures upon the Magnitude of r

If, in a study of the relationship between measures of two traits, we selected two groups of individuals such that one group showed greater variability in these measures than the other, we would find that the coefficient of correlation r between the measures

would be greater for the more variable than for the more homogeneous group. Suppose, for example, that we administered a reading test and an arithmetic test to a group of *sixth* grade pupils, and that the oval marked "VI" in Figure 20 represents the pattern of the tally marks for these scores on a scatter-diagram. That is, this oval might contain all of the larger cell frequencies, although a few scattered tally marks might lie outside this oval. This, then, would represent only a "moderately high" positive correlation, since the oval is so broad in proportion to its length, that is, since the variability of the measures in any single column (or row) would be only slightly less than the variability of the measures in all columns (or rows) combined for the entire sixth grade group.[1] Now suppose that we administered the same tests to a group of seventh grade pupils. These pupils would, in general, earn higher scores than the sixth graders on both tests, and the pattern of their tally marks, when plotted on the same scatter-diagram, might be represented by the oval marked "VII." Similarly, ovals IV, V and VIII might represent the patterns of tally marks, plotted on the same scatter-diagram, for groups of fourth, fifth and eighth graders, respectively. Again we may note, so far as any one of these groups alone is concerned, that the relationship is only moderately high, since in each case the oval is short and broad. However, when we consider *all* groups together, we note that an oval including all of the larger cell frequencies would be quite narrow in proportion to its length. We would therefore expect the coefficient of correlation between these scores to be considerably higher for the total group (all grades) than for the sixth (or any other) grade group alone.[2] In one sense, how-

[1] That is to say, σ_c (the standard deviation of the measures in column *C*) would be almost as large as σ_{VI} (the standard deviation of reading scores for *all* sixth grades).

[2] The standard deviation of the measures in any column (such as column *C*) would be quite small in proportion to the standard deviation (σ_T) of the measures from all columns for the combined groups. It follows, then, that $\frac{\sigma_c}{\sigma_T}$ would be considerably smaller than $\frac{\sigma_c}{\sigma_{VI}}$. From this it follows (see Formula (28)) that the correlation, r_{xy}, would be considerably larger for the heterogeneous total group than for the more homogeneous sixth grade group.

ever, the real degree of relationship is the *same* in either case, since the reliability of estimate (which depends upon the standard deviation of the measures in individual rows or columns) is the same whether we consider the sixth grade alone or all grades combined.

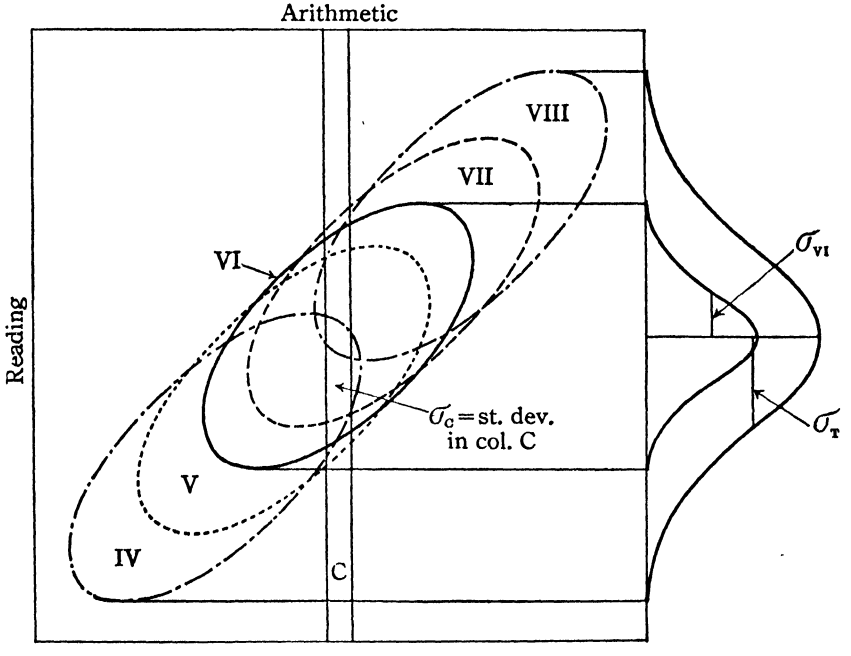


FIG. 20.
Showing influence of range of talent upon *r*.

The magnitude of the coefficient of correlation between measures of two traits for a given group will then depend upon the variability of these measures for the given group, or, as the same idea is frequently expressed, it will depend upon the "range of talent" of the group. The correlation between measures of the *same* two traits may therefore have one magnitude for one group of individuals, and quite a different magnitude for still another group. It follows from this that it is not meaningful to speak of *the* correlation between any two traits, apart from any description of the group for which the correlation is determined. Statements

such as, "The correlation between height and weight is .52," or "There is only a low correlation between intelligence and spelling ability," are therefore indicative of loose thinking. Such statements should *always* be accompanied by a description of the particular group involved, including a description of its variability in the measures concerned. To say, for example, that the correlation between achievement in two school subjects is .60 for a group of fifth grade pupils is quite another thing than to say that it is .60 for a group which includes pupils from all grades from the first to the eighth. Comparisons of the closeness of relationship should therefore not be based on comparisons of r's unless they are established for groups that are at least approximately alike in "range of talent."

The Meaning of a Given Value of r

We have already noted that while the coefficient of correlation r (because of the characteristics noted on page 150) is a convenient *index* of relationship, it may not be considered as *directly* proportional to the *degree* of relationship. A coefficient of correlation of .80, for example, may not be said to represent exactly twice as close a relationship as one of .40, even though both are established for the same range of talent. To be able to make such a statement, we would have to be able to describe, independently of r, just exactly what we mean by closeness of or degree of relationship, and no such description or definition that is generally acceptable has yet been proposed. Because of our inability to define "degree of relationship," we are unable to state in general how r changes in value for given changes in that degree.

It may be well to remind ourselves that r, after all, is simply one of a number of equally arbitrary mathematical procedures which, when applied to sets of related measures, will yield a single number somehow indicative of the degree of relationship. The coefficient of correlation r is based on z-score products; other indices could be derived from *differences* in z-scores for the individuals concerned, or from the *ratios* between their z-scores, or from the

squared differences in z-scores, or from similar measures based on percentiles and ranks, and so on almost without limit. Few of these other indices would have the characteristics that would make them as *convenient* to use and interpret as *r*, but which of them is most nearly *linearly* [1] related to the degree of relationship we cannot say, since this would depend upon how we defined degree of relationship. For the same reason, we cannot say in general that *r* is any better than many other available indices in this regard.

Numerous schemes and devices have nevertheless been suggested to assist the student of statistics to appreciate the significance of a given value of *r*. Some of these devices are quite helpful in certain restricted types of situations, but all of them may be seriously misleading in other situations or in general, and must be used with extreme caution.

One of the most common and most misleading of these practices has been that of classifying *r*'s of certain values as "high," "medium," and "low." For example, an *r* of .30 or less has been said to be "low," one of from .30 to .70 "medium," one of from .70 to .90 "high," and one of above .90 "very high." The numerical values of *r* corresponding to each of these categories has, of course, differed for various classifiers. Such classifications are invariably misleading, since what constitutes a "high" or a "low" correlation is a relative matter, and differs markedly for different types of situations. Coefficients of correlation of as high as .50 between measures of a physical and a mental trait are extremely rare, and a correlation of .60 between two such traits would be considered as phenomenally high for almost any group. Correlations of this magnitude between reliable measures of two mental traits, however, are quite common, and in this instance would be considered as only medium for most populations in which we are interested. Again, a correlation of .90 between two independent measures of

[1] Two variables are linearly related if a given amount of change in the value of one is always accompanied by a constant amount of change in the value of the other.

the *same* mental trait — for example, between the scores on two equivalent tests of spelling ability [1] — might be considered as only medium or low, particularly if the tests were very long and comprehensive. In this situation, an r of .60 would be considered as extremely low. There is no single classification, then, that is applicable in all situations, and because of the danger that they will be applied in situations in which they are not valid, it is best that any and all such classifications be disregarded entirely by the beginning student.[2]

Another device for the interpretation of r is that which is concerned with the improvement over a "best guess" in predictions based on the regression equations. Suppose, for example, that an individual is selected at random from a given group whose mean and standard deviation of a given measure (X) are M_x and σ_x respectively. The "best guess" that we could make of this individual's x measure would then be M_x, and the "standard error" of this guess or estimate would be σ_x. Suppose, however, that we knew the measure of another trait (Y) for this individual, and that variables X and Y were rectilinearly related for the group in question. We could then, by means of the appropriate regression equation, make a better estimate than before of this individual's x measure. The standard error of this estimate would be $\sigma_{x \cdot y} = \sigma_x \sqrt{1 - r_{xy}^2}$. The difference between this latter standard error and the first, that is, the *reduction* or improvement in the standard error of estimate, would then be

$$\sigma_x - \sigma_x \sqrt{1 - r_{xy}^2}$$

[1] In this case the correlation coefficient would also be the coefficient of reliability.

[2] The student will have noted that the adjectives "high," "low," and "medium" have been applied several times in this chapter to correlation coefficients and degrees of correlation. This may appear inconsistent with what has just been said. These adjectives, however, have been used to refer only to the absolute mathematical magnitude of the correlation coefficient; that is, a high correlation in these discussions means one high up along the scale of possible value (near 1.00), a low correlation means one near zero, and a medium correlation means one near .50. Used in this sense, "high" does not imply "important" or "consequential," nor does "low" mean "of no importance" or "of no consequence." The student must distinguish carefully between this use of these adjectives and their use in interpretation or evaluation of correlation coefficients.

This expression could be expressed as a *per cent* of σ_x (the standard error of a "best guess") by dividing by σ_x and multiplying by 100, as follows

$$100 \cdot \frac{\sigma_x - \sigma_x \sqrt{1 - r_{xy}^2}}{\sigma_x} = 100 \left(1 - \sqrt{1 - r_{xy}^2}\right)$$

If, then, the coefficient of correlation between X and Y were $r_{xy} = .80$, the standard error of an estimate based on the regression equation would be less than the standard error of a best guess by an amount equal to 40 per cent of the latter. An r of .60, similarly, would represent a 20 per cent improvement over a best guess. The nature of the relationship between r and this per cent improvement over a best guess is shown in Figure 21. From this figure we note, for example, that an r of .50 represents an improvement of about 14 per cent over a best guess, that an r of .86 represents a 50 per cent improvement, etc. We see, then, that the reduction in the standard error of estimate remains small and increases very slowly for low values of r and that marked improvements come only with very high values of r. For the purposes of prediction, then, an r of .40 is not much better than an r of 0, while the difference between an r of .80 and one of .90 is very much greater than between an r of .50 and one of .60.

Figure 21 is quite helpful in the interpretation of r's used for *purposes of prediction*, but, like all other devices of this type, may be seriously misleading when applied in other situations. It would be a grave mistake, for example, to reason that, because an r of .40 is very little better than an r of 0 in prediction, it is therefore to be considered as "very low" or as of no consequence in other situations. An r of .40 between measures of two mental traits for a given group might have very important implications to the educational psychologist with reference to a theory of learning, even though it would be practically useless in estimating the measure of one trait from that of another for any individual.

Many other devices for the interpretation of r have been suggested in the literature of education and psychology. Except when used by persons highly trained in statistics, however, all

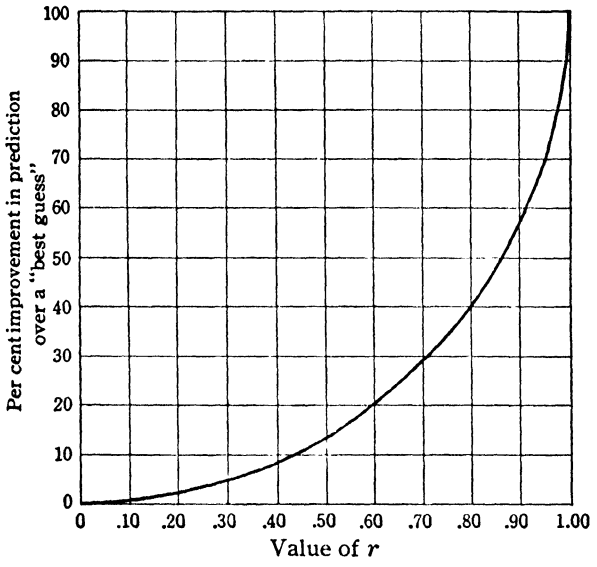


FIG. 21.
Improvement in accuracy of prediction for increasing values of r.

of these devices are more likely to be misleading or confusing than helpful. It is therefore recommended that the beginning student in statistics make no attempt to arrive at any absolute interpretation of r. He should look upon it simply as an arbitrarily selected index which happens to be indicative of (although not linearly related to) the degree of relationship. When comparing r's of different magnitude, he should avoid trying to estimate "how much" closer the relationship is in one case than in another, but should be content with the knowledge that there *is* a difference of some indeterminate amount. He should be careful, also, never to compare r's except when the relationships are known to be rectilinear and when the groups involved are comparable in "range of talent," and should take sampling errors into consideration in all such comparisons. If he wishes to secure a more definite notion of what an r of a given magnitude really means, he can do no better than to study the distribution of the tally marks on the scatter-diagram from which it is computed.

Causal vs. Casual Relationship

One other very important admonition remains to be made. No more serious blunder in the interpretation of correlation coefficients can be made than that of assuming that the correlation between two traits is a measure of the extent to which an individual's status in one trait is *caused by* or due to his status in the other. It is indefensible, for example, to argue that, *because* a high correlation exists between measures of silent reading comprehension and arithmetic-problem-solving ability for the individuals in a given group, problem solving is therefore *dependent* upon reading comprehension or *vice versa*, or that a given student does well in arithmetic *because* he is a good reader. All of this may be true, but it does not necessarily follow from the statistical evidence of correlation.

The observed correlation between measures of two traits is *sometimes* due to a cause-and-effect relationship between them, but there is nothing in the statistical evidence to indicate which is the cause and which the effect. For example, there is a fairly high correlation between age and grade status of elementary school children. In this case we know, of course, that we cannot increase a pupil's age simply by promoting him from one grade to the next — that age is not *due* to grade status — but we know this because of logical considerations which are quite independent of the statistical correlation.

Again, correlations are *sometimes* observed between traits that have no cause-and-effect connection whatever, the observed correlation being due entirely to a third factor (or several factors) which is (or are) related to each of the traits in question. For example, there is a positive correlation for the general population between ages of mothers at parturition and the intelligence of their offspring, but this is because women of high intellectual standards and ability tend, for economic and cultural reasons, to be married later in life, and not because middle age is the best time to bear intelligent children. Again, however, we arrive at this interpretation on the basis of logic which is quite independent of the direction or magnitude of the observed correlation.

Finally, the observed correlation between two traits may sometimes be in just the *opposite* direction from a cause-and-effect relationship which really exists. For example, in almost any high school or college course there is a *negative* correlation (of usually about $-.30$) between grades earned and number of hours spent in study. The students who make the highest grades are in general those who spend the least time studying, while those who make low grades in general spend more than the average amount of time in study. It would obviously be absurd, however, to contend on the basis of this evidence that anyone can make higher grades by studying less. The negative correlation is largely due to the fact that intelligence is positively related to grades and negatively related to time spent in studying — that the less able students *must* study more to even approach, though not equal, the achievements of their more able classmates. The causal connection between grades earned and time spent in study is positive, even though the observed correlation is negative.

Whenever a significant correlation is found between two sets of measures, there are always the possibilities: (*a*) that there is *no* cause-and-effect connection; (*b*) that a cause-and-effect connection is present in the same direction as the observed correlation; (*c*) that there is a cause-and-effect connection, but in the opposite direction from the observed correlation. Which of these possibilities exists, and what is the strength of the cause-and-effect connection (if any) *cannot be determined from the observed correlation.* Any interpretations concerning cause and effect must be based on *logical* considerations, not based on the observed correlation. The observed correlation may *suggest* a cause-and-effect relationship, but can never *prove* that it exists, or show in what degree it exists.

CHAPTER XI

CORRELATION TECHNIQUES APPLIED IN THE EVALUATION OF TEST MATERIALS

A VERY large proportion of all educational and psychological research takes as its basic data the measures or scores obtained through educational and psychological tests, such as intelligence tests, tests of educational achievement, tests of special aptitudes and abilities, and scales for rating personality traits. The dependability and meaningfulness of any conclusions drawn from such research must, of course, depend upon the dependability and meaningfulness of the original data upon which the conclusions are based. Obvious as this statement may seem, it expresses a truth which has been very frequently neglected in past research. Investigators in education and psychology have tended to be seriously uncritical of their original data. They have too often taken it for granted that educational and psychological tests really measure the things which the titles of the tests imply that they measure. They frequently have allowed themselves to become overly intrigued with statistical techniques for their own sake, and to become so impressed by the *method* or *technique* of analysis employed as to overlook the lack of meaning in the data analyzed. The "jingle fallacy" — the mistake of failing to distinguish between the name of a thing and the true nature of the thing named, of failing to differentiate between the name of a test and that which it actually measures — has characterized many reports of educational and psychological research.

If, then, the student of statistics in education and psychology is to develop a sound statistical judgment, it is essential that he acquire a thorough appreciation of the limitations of the original data with which he will have to work. It is extremely important that he recognize how seriously measures of mental traits [1] are

[1] The term "trait," as used in this discussion, is broadly defined to include skills, abilities, aptitudes, attitudes, and educational achievements.

characterized by ambiguity and error, and how inadequately we are able to control these errors or to describe their nature and magnitude by means of available statistical and research techniques. The purpose of the following discussion, accordingly, is to develop in the student a better appreciation of the nature of measurement in education and psychology.

THE NATURE OF MEASUREMENT IN EDUCATION AND PSYCHOLOGY

Mental traits or abilities, unlike height and weight, are intangible in character and, in general, can be measured only *indirectly* in terms of their manifestations in the overt behavior of individuals. Let us consider, for example, the nature of a test of general intelligence. The measurement of "general intelligence" consists essentially of noting how many of a number of selected mental tasks of varying difficulty an individual can complete successfully under certain standard conditions. To construct such a test, the test author would first make a collection of problems, puzzles, questions, or other mental tasks each of which, in his opinion or in the opinion of other competent observers, requires the exercise of intelligence. He would try to include a variety of types of tasks involving various aspects of general intelligence, and would attempt to secure a wide range in difficulty, including some tasks intended to test the very stupid and some to challenge the very intelligent individuals. He might then administer these potential test items experimentally to a group of individuals, some of whom are *generally considered* to be "bright" and some to be dull mentally. He would then discard any item not successfully completed by a larger *proportion* of bright than of dull individuals, since such items would not contribute to the purpose of the whole test, which is to reveal *differences* in intelligence. On the basis of the assumption that mental ability increases with chronological age, he might also discard any items which do not show an increasing proportion of successes at succeeding age levels. He would then

assemble the remaining tasks or items into a "test," in which the person tested is to be given one point of credit for each task completed. He would then devise a set of standard directions for administering the test, and finally would administer it to a large and representative group of individuals for the purpose of establishing "norms" of performance in terms of which mental ages could be computed.

The important thing to note about this whole procedure is that at no stage in the process, either in making the original selection of tasks or in their final assembly into a "test," would the test author be able to describe *exactly* what he means by general intelligence. Certainly, no test author has yet been able to provide a meaningful definition which has proved entirely acceptable to other equally competent psychologists. He can only claim for his test that it does, on a more objective, reliable, and *comparable* basis, what each of us does when we subjectively estimate the intelligence of our acquaintances by noting what things they are able to do. In the last analysis, the only unambiguous definition of general intelligence is that it is what is measured by a general intelligence test. Intelligence, like nearly all other mental traits, is both defined and measured in terms of the concrete situations in which it is overtly manifested.[1]

It should be apparent from the foregoing discussion that the number of behavior situations in which any given mental trait or ability may manifest itself is almost unlimited. There is no practical limit, for example, to the number of mental tasks which might be employed in the construction of a general intelligence test, or to the number of different problems which might be devised for use in a test of arithmetic reasoning. These behavior situations, furthermore, are in general quite complex; that is,

[1] The description here given of the nature of a general intelligence test does not do justice to the most recent work in mental testing, in which attempts have been made to identify the basic "factors" in general intelligence through objective, mathematical analyses of test data. This description, however, is in all essential respects valid with reference to most existing tests of "general" intelligence, as well as to available tests of other more specific mental abilities and aptitudes and to tests of educational achievement.

the individual's behavior in any one of them may depend upon many other traits and abilities than the one in which we are interested. The individual's score on an arithmetic reasoning test, for example, might depend in part upon his understanding of the vernacular in which the problems were stated, upon his rates of reading and writing, upon his ability to follow the directions for taking the test, and upon many other factors, some of which may be quite irrelevant to his ability in arithmetic reasoning. Situations representing a "pure" manifestation of any single trait in isolation are virtually impossible to find. Most of the traits in which we are interested, furthermore, are in themselves quite complex in character; that is, they may consist of combinations or hierarchies of still simpler skills and abilities. Ability in arithmetic, for instance, consists, among other things, of ability in addition, in subtraction, and in multiplication, while ability in addition of whole numbers may not be identical with ability in addition of fractions or of denominate numbers, and even ability in addition of whole numbers may be further analyzed into other simpler skills, resting finally upon the 100 basic addition facts (the possible combinations of two one-digit numbers). Not only are nearly all traits, abilities and achievements measured by educators or psychologists of this complex type, but for few if any of them do authorities agree upon the nature and relative significance of the elements constituting the complex total.

All Measurement Involves Sampling

The *definition* of any mental trait, then, involves the identification or description of situations or types of situations in which an individual's behavior is partly or primarily dependent upon the amount of the given trait which he possesses. Since the number of such situations is practically unlimited, the *measurement* of the trait involves the selection of a *sample* of these situations — a sample small enough to make it practicable for us to observe the individual's behavior in each situation. Ideally, the sample of situations used for the measurement of the individuals in any

given group should, for fairly obvious reasons, satisfy each of the following conditions.

1. The sample must be *representative* of all the various types of situations in which the trait may be manifested, or of all the various elements constituting the complex total trait, or of all the various aspects of that trait.

   For example, an adequate test of general ability or achievement in arithmetic must contain some problems in addition, some in multiplication, some in subtraction, etc., and among the problems in addition there must be some in addition of whole numbers, some in addition of decimals, and some in addition of denominate numbers, while among the problems in addition of denominate numbers different types of denominate numbers and different degrees of complexity must be represented.

2. The sample must be *large* enough to yield a *stable* or dependable measure of the individual's general ability.

   For example, in a test of spelling ability consisting of only 20 words, two individuals who differ in general spelling ability might make the same score, since the small sample of words used might by chance contain a relatively large proportion of the few words which the one can spell and a relatively small proportion of the many that the other can spell. For similar reasons, two persons of the same general ability might make markedly different scores. The longer the test, that is, the more extensive the sampling, the less serious will be these chance fluctuations in obtained scores.

3. The individual's behavior in each situation must be relatively uninfluenced by traits or factors irrelevant to the trait being measured.

   For example, the usual self-administering type of general intelligence test would not be satisfactory for measuring the intelligence of a group of recent immigrants to this country, because of the high premium placed in such tests upon knowledge of the English language. Again, the grade received by

a pupil on an essay examination may depend in part upon the legibility of his writing, or upon the speed with which he writes, or upon his ability to infer the teacher's intended meaning from ambiguously stated questions, or upon his ability to reproduce the sterotyped phrasing of the textbook, or upon other factors not closely related to his true achievement in the field tested.

(2) Each situation must *in itself* differentiate between individuals possessing different amounts of the total trait or representing different degrees of development of that trait.[1] The whole sample of situations, furthermore, must show a sufficient *range* of difficulty to discriminate between individuals above and below all *levels* of ability found in the group being measured.

The first element in this condition has been partially illustrated in the description of the manner in which an intelligence test author discards certain items. Some items in a test may not differentiate because they are either too easy or too difficult, so that all individuals tested may succeed or all may fail on the item. If the response (whether right or wrong) to an item is identical for all persons tested, then clearly that item cannot help to reveal any *differences* between these individuals. Other items may be answered differently by different individuals, but those who respond correctly may, on the average, possess the same amount of the total trait as those who respond incorrectly. Certain words, for instance, may be misspelled as frequently by good spellers as by poor spellers.

To illustrate this condition further, suppose we wished to rank a number of 15-year-old boys in the order of their ability in the high jump. To do this, we would need a number of hurdles of varying heights. The range of heights required in the hurdles used would depend upon the range of ability in the boys being

[1] A more adequate discussion of the differentiating power of individual test items may be found on pages 39 ff. in *The Construction and Use of Achievement Examinations*, Hawkes, Lindquist, Mann and others. Houghton Mifflin Co.

measured. A hurdle which *none* or which *all* of these boys could jump would not help us in *ranking* them. Similarly, in the construction of a vocabulary test for a seventh grade group, it would be futile to include words known to *all* individuals in the group or words known to *none* of them, since such words obviously could not *discriminate* between pupils (in this group) whose vocabulary is broad or limited. Furthermore, the test must contain some very difficult words to discriminate between the superior and the very superior pupils, some very easy words to discriminate between the very inferior and the inferior pupils, and some words of intermediate difficulty to discriminate between pupils at other levels of ability.

These ideal requirements, of course, can never be completely satisfied in actual test construction. A major obstacle to satisfying the first requirement, for example, is the failure of authorities to agree upon a specific and meaningful analysis and description of the trait to be measured. Without any such analysis, it is impossible to say when the content of a test assigns proper weight to or duly represents each of the constituent elements of the complex total trait being measured. A test frequently can be made *long* enough to satisfy the second condition, but length avails little if the content of the test is biased or if test performance is unduly influenced by irrelevant factors, and as has already been noted, it is rarely possible to discover or devise situations that are entirely free from irrelevant factors.

All Mental Measures Are Uncertain as to Meaning

It should be apparent from a consideration of the preceding requirements that the selection of a sample of behavior-situations (or test exercises) for the measurement of any mental trait depends in large part upon arbitrary subjective opinion. For most traits it would be virtually impossible to select a sample upon which complete agreement in authoritative opinion could be secured. For this reason, if for no other, the measures obtained are inevitably ambiguous to some degree, and often very seriously so. This

ambiguity is accentuated by the character of the process of "scoring" an educational or psychological test, and by the nature of the *scale* along which the scores are expressed. The common practice in mental testing is to assign an arbitrary number of points of credit for the desired response to each situation (or test item), the individual's score being the sum of such points earned. There is, however, no way of demonstrating conclusively that these arbitrary weights are in proportion to the "true" values of the items, or of determining with complete objectivity what these weights *should* properly be. Hence, the linear scale of values along which these scores are expressed is not comparable to those employed in physical measurement. The "unit" employed is unique to each scale and cannot be described or defined in more fundamental terms, while its value fluctuates from point to point even within the same scale. Furthermore, the zero point on each scale is merely an arbitrary reference point whose relation to the absolute zero is not known. (See pages 29–31 of this text and questions 1 and 2 on pages 8 and 9 of the manual.)

Still further ambiguity in mental measurement results from the facts that the traits which we wish to measure are themselves dynamic and fluctuating within the same individual, and that the measures obtained of them are partly dependent upon attending (and often accidental) circumstances. An individual's performance on a mental test always depends to some degree upon the manner in which and the circumstances under which the test is administered. Even though these external factors are controlled or held constant, variations in the individual's own physiological or emotional status may influence his responses.

For these reasons, the scores obtained on educational and psychological tests must always be very cautiously interpreted. Such scores must never be accepted at their face value, but must always be considered as only *approximate* indications of the true relative status of the individuals tested, or as likely to contain "errors" of various types, any of which may be of considerable magnitude.

The sources of these "errors of measurement" may be briefly summarized as follows:

1) The *indirect* character of all mental measurement
2) The lack of generally accepted, objective and meaningful *definitions* of the things to be measured
3) The *limited sampling* of behavior-situations upon which the measures are based
4) The unintentional measurement of *irrelevant* factors
5) The nature of the measuring *scales* employed
6) The fluctuating character of the individual's mental, emotional, or physiological state

THE MEASUREMENT OF ERRORS IN MEASUREMENT

Test Validity

The *validity* of a test may be defined as the accuracy with which it measures that which it is *intended* to measure, or as the degree to which it approaches infallibility in measuring what it purports to measure. The degree of validity of a test, therefore, depends upon the magnitude of the "errors" (due to any and all of the causes just considered) which are present in the measures obtained from it. The actual magnitude of the errors in a set of fallible measures of any trait could, of course, be determined directly only if we had available the corresponding infallible or "perfect" measures of the same trait for the same individuals. We could then describe the validity of the fallible measures in terms of the average or median difference between the fallible and infallible measures (that is, in terms of the average or median error), or in terms of the coefficient of correlation between the two sets of measures, or in terms of the probable error of estimating the infallible from the fallible measures. In this case, the coefficient of correlation between the infallible and fallible measures would be considered as the true *coefficient of validity* of the latter for the group of individuals involved. For reasons already given, however, it is impossible to secure perfect or infallible measures of any mental trait for any group of individuals, and hence it is impossible

to provide exact descriptions of the true validity of any of the fallible measures which we are able to obtain. The true validity of an educational or psychological test *must always remain a hypothetical concept*, since there is never available an infallible "criterion" measure against which the fallible obtained measures may be evaluated.

In some situations, however, a partial indication of the validity of a given test may be secured by a study of the correlations between the scores obtained on the given test and on other tests of the same trait. If a number of different tests are available for the measurement of the same trait, and if it is the consensus of authorities that one of these tests is definitely better than any of the others, then this test may be used as a "criterion" against which the others may be evaluated. Suppose, for example, that we wish to determine which of two given intelligence tests (Tests A and B) is the more valid for use at the seventh grade level, and that each of these tests is of the type which may be conveniently administered to a large group of pupils in a short testing period — let us say, 30 minutes. Now it would be generally admitted by test authorities that no 30-minute test of the "self-administering" type can yield results as dependable as those which may be secured from an "individual" intelligence test such as the *New Stanford Revision of the Binet-Simon Scale for the Measurement of Intelligence*, which can be administered only to one pupil at a time and in a relatively long period. Suppose then that we administer all three of these tests under standard conditions to the pupils in a random sample of seventh graders, and that we compute the coefficient of correlation between the *New Stanford* I.Q.'s and those obtained from each of tests A and B. If, then, we find that this correlation is significantly higher for Test A than for Test B, we might consider this fact as strong evidence that, *for pupils like those in the sample*, Test A is the more valid in determining individual I.Q.'s. How convincing this evidence would be to us would depend, of course, upon our confidence in the *New Stanford Revision* as a criterion test.

The principal limitation of this method of securing objective evidence of test validity is that very often the tests which we wish to evaluate are themselves the best instruments that we know how to build for measuring the traits in question. With reference to the preceding illustration, for example, it would be extremely difficult to devise an intelligence test which would be generally conceded to be definitely superior to the *New Stanford Revision* and which could be used as a criterion to describe the validity of that test. Only when we feel certain (on the basis of *subjective opinion*) that one method of measurement is definitely superior to another may we reasonably use the one as a criterion in the evaluation of the other. Even in this case, the criterion itself would still be fallible, and hence the correlation coefficient obtained would not be a true coefficient of validity, but would only be indicative of the *amount* of *agreement* between one fallible measure and another which is perhaps somewhat less fallible. The student is warned that in the literature of education and psychology, he will find presented as "validity coefficients" many correlation coefficients which, because of the questionable character of the criterion, should not be thus described.

Test Reliability

An important characteristic of any test, a characteristic which is essential to but not a guarantee of validity, is *self*-consistency or *reliability* in measurement. The individual items or behavior situations constituting any mental test always represent only a very limited *sample* selected from a very much larger number of possible or available items. Any two such samples, even though similarly selected, are almost certain to present differences in difficulty as far as any given individual is concerned. Suppose, for example, that two samples of 50 words each are selected in the same way from the same "master list" of spelling words, and that each list is administered as a list-dictation spelling test to the same large group of high-school juniors. While the two distributions of scores may be practically identical as far as the groups

are concerned, very few individuals will make exactly the same score on both tests. As a result of chance differences in the sample, any one pupil will almost certainly find more words that he can spell in one list than in the other. If the differences in the two scores are large for most pupils — that is, if there is a low relationship between the two sets of scores for the entire group — then either of the scores made by a given pupil would have to be considered as largely due to chance, and very little reliance could be placed in it as a measure of his ability. Close agreement in the scores, however, would not prove the test to be valid as a measure of general spelling ability, since close agreement could be found even though each list represented a very *biased* sample, or even though irrelevant factors had seriously influenced pupil performance. If, for instance, each list had been unduly weighted with words of Latin origin, those pupils who had studied Latin might have an unfair advantage, but this fact would not lead to *inconsistency* of performance if both lists were of this character. Again, if both lists were dictated too rapidly, each pupil's score might depend in part upon how fast he could write, but since the slow writers would be equally handicapped in both tests, higher rather than lower agreement in the two sets of scores might result. Consistency in measurement is therefore *an essential but not a sufficient condition* for test validity.

The Coefficient of Reliability

The coefficient of reliability of a test for a given group is defined as the coefficient of correlation between the scores made by that group on two equivalent forms of the test successively administered under the same conditions. Two forms of a test are said to be "equivalent" if both contain similar content, that is, if the samples of items constituting them were similarly selected from the same materials, and if both forms show the same distribution of scores (equal means and equal variability) for the same group. Since strictly equivalent forms of a test are seldom available, a more satisfactory practicable definition of the coefficient of reliability of

a test is that it is the *average* intercorrelation between scores on a number of forms that have been made *closely* equivalent. By definition, then, all equivalent forms of the same test are equally reliable.

The coefficient of reliability is then simply a special application of the coefficient of correlation. Whatever was said in the preceding chapters concerning the interpretation of correlation coefficients in general is equally applicable to coefficients of reliability. Particular consideration should be given to the influence of the range of talent upon the coefficient of reliability. When computed for a group that is widely variable in the trait measured, the coefficient of reliability of a given test may be considerably higher than when computed for a group that is relatively homogeneous in the same trait. An achievement test, for example, might show a reliability of .95 for a group of third to eighth grade pupils and of only .80 for a group selected from grade five alone. We therefore cannot speak meaningfully of *the* coefficient of reliability of any test. The same test will show different reliability coefficients for different groups. Each reliability coefficient must be accompanied by a description of the group upon which it is based to be meaningfully interpreted. For this reason, coefficients of reliability of different tests may be directly compared only if computed for the same group or for groups of comparable ranges of talent.

As is true of the coefficient of correlation in general, coefficients of reliability are also subject to fluctuations in random sampling, and little dependence can be placed in them unless they are based upon reasonably large groups of individuals. Again, as is true of correlation coefficients in general, it is dangerous to attempt to set up any arbitrary standards for the evaluation of reliability coefficients. What may be considered as a "high" or "satisfactory" coefficient of reliability in one situation may be considered as "low" or "unsatisfactory" in another, depending upon the nature of the thing measured, upon the length of the test, upon the range of talent involved, and upon the purpose for which the scores are used. A reliability as low as .40 may be adequate for

comparisons of average scores for large groups of individuals, while a coefficient even as high as .95 may in some situations be considered inadequate where very accurate descriptions of individuals are desired. The student is therefore advised to make no attempt to set up any single classification of reliability coefficients as "high," "medium," or "low," but to evaluate the reliability of each test on a relative basis in comparisons with coefficients similarly obtained for other available tests of the same trait.

Ways of Estimating Coefficients of Reliability

The coefficient of reliability of a test for a given group can, of course, be computed in the manner implied in the definition only if two or more equivalent forms of the test are available. The majority of the tests whose reliability we wish to describe exist in only one form, and the labor involved in constructing an equivalent form makes it impracticable to do so simply for the sake of computing a reliability coefficient. For such a test, we can sometimes obtain a useful approximation to its true reliability for a given group by splitting the single test by chance into halves, assuming that these halves are "equivalent" to one another, and scoring each half separately for the individuals in the given group. Such "chance halves" are usually obtained by letting the odd-numbered items constitute one half and the even-numbered items the other. If the two halves are truly equivalent, the coefficient of correlation between the scores on them would, by definition, be the coefficient of reliability of either half alone. We can then estimate the reliability of the whole test by means of the Spearman Brown Prophecy Formula, which indicates the relationship between the reliability of a test and its length. The general form of this formula is as follows:

$$r_n = \frac{nr_{12}}{1 + (n - 1)r_{12}} \tag{33}$$

where r_{12} represents the coefficient of reliability of a given test (the correlation between scores on equivalent forms 1 and 2), and r_n represents the coefficient of reliability of a test n times as long

as the given test but in all other respects comparable to it. (The longer test may best be thought of as consisting of n equivalent forms of the given test.) In the case in which we wish to estimate the coefficient of reliability of a whole test from the coefficient of reliability of one of its halves, n in this formula would be equal to 2, and the formula would become

$$r_{12} = \frac{2r_{\frac{1}{2}\frac{1}{2}}}{1 + r_{\frac{1}{2}\frac{1}{2}}} \tag{34}$$

in which $r_{\frac{1}{2}\frac{1}{2}}$ is the correlation between scores on the chance halves, and r_{12} the *estimated* reliability of the whole test, or the estimate of the correlation that would be found between scores on equivalent forms 1 and 2 if such forms were available.

The principal shortcoming of this method of estimating test reliability is that chance halves of a test are rarely closely equivalent, and hence the coefficient of correlation between scores on the halves only roughly approximates the coefficient of reliability of either half. Furthermore, the juxtaposition of the items constituting the two halves and the fact that the individual's responses to certain items may be influenced by the responses he has already made to others, together with other factors, may result in a closer agreement in the scores on the two chance halves of the same test than would be found if these two halves were independently administered as separate tests. Whatever the reason, it has been well established that coefficients of reliability estimated by the chance halves method are usually higher than those computed for the same test by correlating scores on independently administered equivalent forms. Coefficients of reliability estimated in this manner, then, are not only less dependable than those computed directly, but also are likely to be spuriously high, and must be interpreted accordingly.

Another estimate of the coefficient of reliability of a test existing in only one form may be made by finding the coefficient of correlation between the scores obtained by administering the *same* test twice to the same group. This is in general a very unsatisfactory method, since it almost invariably results in spuriously

high coefficients. The correlation between scores on successive administrations of the same test is essentially an index of the individual's consistency of performance on the *same items*, rather than of the adequacy of these items as a sample of what he can do in general. If, for example, an individual were given 50 words to spell and sometime shortly thereafter were given the same 50 words to spell again, in the latter situation he probably would simply reproduce without variation the same spellings previously given. Certainly we would expect his score on the second test to be much more like that on the first than if the second test had consisted of an entirely different set of 50 words. The method of repeated administration perhaps may be safely employed only when the individual's responses in the second testing are not a function of his *memory* of specific information or of his ability to recall the responses made by him in the first testing. This means that this method should never be employed to determine the reliability of a test of educational achievement.

The Reliability of a Single Score

Suppose we had available a large number of equivalent forms of the same test, and that we administered all of these forms to the same individual under the same conditions. Because of the differences in the samples of items constituting the various forms, we would, of course, expect the individual to make higher scores on some forms than on others. If the test were highly reliable, we would expect most of his obtained scores to have very nearly the same value, but if the test were low in reliability we would expect wide variations in his obtained scores. The standard deviation of the distribution of these obtained scores would then describe the reliability of a single score obtained on one form of the test and hence would also describe the reliability of the test. If the number of obtained scores (or equivalent forms) were very large, the mean score in this distribution would be known as the individual's "true score" on the test, and the standard deviation of the distribution would be the *standard error* of a single obtained

score. Assuming that the obtained scores would be distributed in the form of the normal curve, we could then interpret this standard error in the same fashion in which we previously interpreted the standard error of the mean of a random sample (as is illustrated in the following paragraph).

Since we rarely have even as many as two equivalent forms available for a test, and never a very large number, the standard error of a single score can never be computed empirically by the method just described. It may be shown, however, that

$$\sigma_m = \sigma \sqrt{1 - r_{12}} \qquad (35)$$

in which σ represents the standard deviation of obtained scores on a single form of the test administered to a large group of individuals, r_{12} represents the coefficient of reliability of that test for that group,[1] and σ_m represents the standard error of a single score or the "standard error of measurement." To illustrate the application of this formula, suppose that on a given achievement test in United States history administered to a large group of tenth grade pupils the standard deviation of obtained scores was 24. Suppose also that the correlation between scores on two equivalent forms (that is, the coefficient of reliability) of this test is .84 for the group in question. In this case, the standard error of a single score would be $\sigma_m = 24 \sqrt{1 - .84} = 9.6$. The *probable* error of a single score accordingly would be $.6745\, \sigma_m = 6.5$. If, then, a very large number of equivalent forms of this test were actually administered to one of these pupils, we would expect his obtained

[1] The student will note that this formula is much like that of the standard error of estimate (see page 189). In fact, σ_m may be considered as a standard error of estimating an obtained score on a test from the corresponding true score. If for each of the individuals in a large group we knew both the "true score" on a test and the obtained score for a single form of the test, the coefficient of correlation, r_{ot}, between these obtained and true scores could be shown to be equal to $\sqrt{r_{12}}$, that is, to the square root of the coefficient of reliability of the test. Hence the standard error of estimating obtained scores from true scores is equal to

$$\sigma_{o.t} = \sigma_o \sqrt{1 - r^2_{ot}} = \sigma \sqrt{1 - r_{12}}$$

since $r^2_{ot} = r_{12}$. The standard error of measurement, σ_m, may then be considered as the standard deviation of obtained scores for a group of individuals all of whom have the same true score. The student should guard carefully against any tendency to confuse the standard error of measurement with the standard error of estimate.

scores to fall into a distribution whose standard deviation would be 9.6 score units. Assuming that this distribution would be normal, we could then say that on approximately 68 out of every 100 forms of this test the pupil's obtained score would be within 9.6 units of his true score, or that the chances are 68 in 100 that his score on any single form would be within 9.6 units of his true score. Similarly, the chances are 50 in 100 that his obtained score on any single form is within 6.5 units of his true score. Again, since only a negligible proportion of the measures in a normal distribution deviate from the mean by three standard deviations, we can say that it is practically certain that any single obtained score will be within $3 \times 9.6 = 28.8$ score units of the corresponding true score.

The standard error of measurement, like the standard error of estimate, has the advantage that it is presumably independent of the range of talent in the group for which it was determined. The standard error of measurement for an achievement test in arithmetic, for example, would have nearly the same value if computed for a group of third to eighth grade pupils or for a group of fifth graders only. The standard error of measurement, however, has the disadvantage that it is expressed in terms of the unique unit in which the scores are expressed. Unlike the coefficient of reliability, which is an abstract index independent of the size of unit employed in measurement, the standard error of measurement may not be compared for different tests and is difficult to interpret for a single test because of the uncertainty as to the meaning or absolute magnitude of the "unit" employed. The standard error of measurement is, nevertheless, an extremely important statistical concept, and should be much more widely employed in educational and psychological research than it has been in the past. Even though the standard error of measurement is difficult to interpret because of the nature of the measuring scales employed, its use does serve to emphasize the very important fact that test scores may never be accepted at their face value but must always be considered as only approximate indications of

the true relative status of the individuals measured. If those concerned with the interpretation of test scores, whether in educational research or in the practical school situation, followed the practice of writing after each score the value of its probable error, the mistake would be less frequently made of attributing real significance to what are often only accidental variations in test performance. In the interpretation, for example, of educational profiles of individual pupils based upon achievement test batteries such as the Stanford Achievement Test, it should always be remembered that minor "peaks" and "sags" in the profile can readily be explained in terms of the unreliability of the tests and should not be taken too seriously.

The Significance of Measures of Reliability

It has already been noted that tests intended for the measurement of certain abilities or achievements for a given group of individuals often, in spite of the best efforts of the test author, actually measure other abilities than those which they are intended to measure. In other words, what a test is intended to measure and what it actually does measure may be and often are quite far apart. The reliability of a test, however, does not give any indication of how far apart these two things may be. The reliability of a test indicates only how consistently it measures that which it actually does measure. As long as a test measures *anything* consistently, it is reliable, no matter how much what it does measure differs from what it is intended to measure. If a test is unreliable, that is, if it is not measuring *anything* consistently, it of course cannot be valid, that is, it cannot be measuring accurately what it is intended to measure. The coefficient of reliability theoretically sets an upper limit to the validity of a test,[1] but it does not indicate how far below that limit the true

[1] The coefficient of reliability of a test, as well as the standard error of measurement, takes into consideration only those fluctuations of obtained scores that are found between equivalent forms when administered under the same conditions. In other words, it takes into consideration only those variations or "errors" that are due to *chance* differences in the samples of items constituting the various forms, and disregards any *constant* error in the scores that may be due to any systematic

validity lies. The coefficient of reliability is, therefore, most useful for identifying *poor* tests or for demonstrating that a test cannot possibly be high in validity. The fact that a test has a high reliability coefficient, however, never constitutes proof that the test is highly valid. Reliability coefficients are, therefore, of very little if any value in demonstrating which of a number of tests is *most* valid.

If a test happens in part to be measuring some trait or ability other than that which it is intended to measure, this undesirable characteristic of the test may in itself contribute to high reliability, even though, of course, it would tend to lower the validity of the test. Suppose, for example, that the students in a freshman course in college English are given forty-five minutes to write an exposition on a certain designated topic and that grades are subjectively assigned to these expositions, to be used as a measure of what is vaguely described as "the ability to organize ideas and to express them effectively in writing." In rating these papers, the instructor may be unconsciously influenced by such factors as the legibility of the student's writing, the sheer length of his exposition, and the number of mechanical errors that he has made in spelling, capitalization, punctuation, and grammar. Because of the influence of these irrelevant factors, the grades assigned may be more consistent from one situation of this kind to another than if the instructor had succeeded in entirely disregarding them in rating the papers. The reason is that such things as errors in the mechanics of correct writing, legibility, and length are much more readily and objectively recognized than are weaknesses in "organization

bias which characterizes all forms alike. If there is no such bias in the equivalent forms, that is, if the only errors present are chance errors, then the "true score," which is theoretically a perfectly reliable measure of whatever the test is actually measuring, becomes also a perfectly reliable measure of what the test is intended to measure. In this case (which is hypothetical only, and would never be found in actual practice) the "true score" would be a perfectly valid as well as a perfectly reliable score. Hence, in this instance the coefficient of correlation between obtained scores and true scores would be a true coefficient of validity of the test. We have already seen that the correlation between obtained and true scores is equal to the square root of the reliability coefficient of the test. Hence, $\sqrt{r_{12}}$, which is known as the *index* of reliability, theoretically represents the upper limit of the coefficient of validity for a test. For example, if a test has a reliability of .81, theoretically it cannot have a validity coefficient higher than .90.

and expression." Whenever performance on a test is influenced by irrelevant factors, then, and when these irrelevant factors are highly consistent in their influence upon performance, their presence will frequently tend to raise rather than lower the reliability of the scores obtained.

The coefficient of reliability is particularly restricted in usefulness in the evaluation of available standardized tests of educational achievement. In most of the common school subjects, the available published tests nearly all have fairly high coefficients of reliability for the groups for which they are intended; that is, they are not characterized by any very large differences in reliability. For the range of reliability coefficients which have been reported for the standardized tests in a given school subject, it is probably true that there is very little relationship between test validity and test reliability, and it is even conceivable that in some instances there may be a tendency toward a negative relationship. In other words, it may sometimes happen that for a number of tests, all of which are fairly high in reliability, the most reliable test is among the least valid and the least reliable is among the most valid. This happens because within any field of instruction certain outcomes of teaching are much more difficult to measure objectively than certain others, and because there appears to be some tendency to give undue prominence in tests to those outcomes that may be most readily and most objectively measured, regardless of their relative significance. For example, in the field of United States history it is relatively easy to measure with high reliability the amount of descriptive information the pupil has acquired, but it is comparatively difficult to measure the extent to which he has integrated this information, has appreciated its significance, and can use it in the interpretation of contemporary institutions and practices. Again, it is easier to measure the ability to recall stereotyped textbook statements than to measure true understanding of the ideas that they contain. Tests which are primarily informational in character, therefore, or which place an undue premium upon lesson learning of the

verbal type, tend to be more reliable than those in which a sincere effort is made to measure the less tangible objectives of teaching.

For these and similar reasons, it is conceivable that the efforts of test constructors to obtain high reliability in their tests has in some instances resulted in less valid measurement than if they had not depended so much on the reliability coefficient as an index of test quality.

The concept of test *reliability* has been given undue prominence in the research literature of educational and psychological measurement during the past ten or fifteen years, probably because quantitative descriptions of validity are so difficult to secure and coefficients of reliability so easy to secure for most educational and psychological tests. For this reason, a special effort has been made in the preceding discussion to draw to the student's attention the limitations of the coefficient of reliability as a measure of test quality. It is hoped, however, that the student will not derive from this discussion the idea that the coefficient of reliability is of *no* value in test evaluation or that a measure of a test's reliability is necessarily misleading as to its quality. High reliability is an essential characteristic of a good test, and reliability data are extremely useful for the identification and elimination of unpromising techniques of measurement. Tests that differ very widely in reliability will usually differ in the same direction in validity, but small differences in reliability, particularly at the upper limits of the range of possible values, are probably seldom indicative of either the direction or magnitude of the corresponding differences in validity.

There are many other statistical procedures, in addition to those considered here, that may be used in the evaluation of test material. In general, however, the true quality of a test can rarely be adequately described objectively in statistical terms. In most instances, particularly in educational achievement testing, final judgment as to the validity of a test must be based primarily on a *subjective* appraisal of the detailed content of the test in relation to an authoritative description and competent logical analysis of

the trait or objectives to be measured. Lacking the extensive technical training and experience necessary to make an adequate appraisal of this type himself, the ordinary test user must depend almost entirely upon authoritative or expert opinion in the selection of test materials.

APPENDIX

TABLE I

PER CENT OF TOTAL AREA UNDER THE NORMAL CURVE BETWEEN MEAN ORDINATE AND ORDINATE AT ANY GIVEN P.E. DISTANCE FROM THE MEAN[1]

$\frac{x}{P.E.}$	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.00	.27	.54	.81	1.08	1.35	1.61	1.88	2.15	2.42
.1	2.69	2.96	3.23	3.49	3 76	4.03	4.30	4.56	4.83	5.10
.2	5 37	5.63	5 90	6 16	6 43	6.70	6 96	7.23	7 49	7.75
.3	8 02	8.28	8.54	8.81	9.07	9.33	9 59	9.85	10.11	10.37
.4	10.63	10.89	11 15	11.41	11.67	11.93	12.18	12.44	12 69	12.95
.5	13.20	13 46	13 71	13.96	14.22	14.47	14.72	14 97	15.22	15.47
.6	15.71	15 96	16.21	16.46	16.70	16 95	17.19	17.43	17.68	17.92
.7	18.16	18 40	18.64	18 88	19.12	19 35	19.59	19 82	20 06	20 29
.8	20 53	20 76	20.99	21.22	21.45	21 68	21.91	22.13	22 36	22.58
.9	22.81	23 03	23.25	23.48	23.70	23.92	24 13	24.35	24.57	24 79
1.0	25.00	25.21	25.43	25.64	25.85	26.06	26.27	26.48	26.68	26 89
1.1	27.09	27.30	27.50	27.70	27.90	28.10	28.30	28.50	28 70	28 89
1.2	29.09	29 28	29.47	29 66	29.85	30.04	30.23	30.42	30.60	30 79
1.3	30.97	31 15	31.34	31.52	31.70	31 87	32 05	32.23	32 40	32.58
1.4	32.75	32 92	33.09	33.26	33.43	33.60	33.76	33 93	34 09	34 25
1.5	34.42	34 58	34.74	34.90	35.05	35.21	35.36	35.52	35.67	35.82
1.6	35.97	36 12	36.27	36.42	36.57	36.71	36.86	37.00	37 14	37 28
1.7	37.42	37.56	37.70	37.84	37.97	38.11	38.24	38 37	38.50	38.63
1.8	38.76	38.89	39.02	39.15	39.27	39.39	39.52	39.64	39.76	39.88
1.9	40.00	40.12	40.23	40.35	40.46	40.58	40.69	40 80	40.91	41.02
2.0	41.13	41 24	41.35	41.45	41.56	41 66	41.77	41.87	41.97	42.07
2.1	42 17	42.27	42.36	42.46	42.55	42.65	42.74	42.84	42.93	43.02
2.2	43.11	43.20	43 29	43 37	43 46	43.54	43.63	43.71	43.80	43.88
2.3	43.96	44 04	44.12	44.20	44.28	44.35	44.43	44.50	44.58	44.65
2.4	44.73	44.80	44.87	44.94	45.01	45.08	45.15	45.21	45.28	45.35
2.5	45.41	45 48	45.54	45.60	45.67	45.73	45.79	45.85	45.91	45.97
2.6	46 03	46.08	46.14	46.20	46 25	46.31	46.36	46.41	46.47	46.52
2.7	46.57	46.62	46.67	46 72	46 77	46.82	46.87	46.91	46.96	47.01
2.8	47.05	47.10	47.14	47.19	47.23	47.27	47 31	47.36	47.40	47.44
2.9	47.48	47.52	47.56	47.59	47.63	47.67	47.71	47.74	47.78	47.81
3.0	47.85	47.88	47.92	47.95	47.98	48.02	48.05	48.08	48.11	48.14
3.1	48.17	48.20	48.23	48.26	48.29	48.32	48.35	48.37	48.40	48.43
3.2	48.46	48 48	48.51	48.53	48.56	48.58	48.61	48.63	48.65	48.68
3.3	48.70	48.72	48.74	48.76	48.79	48.81	48.83	48.85	48.87	48.89
3.4	48.91	48.93	48.95	48.97	48 98	49.00	49.02	49.04	49.05	49.07
3.5	49 09	49.10	49.12	49 14	49.15	49.17	49 18	49.20	49.21	49.23
3.6	49.24	49.26	49.27	49.28	49.30	49.31	49.32	49.33	49.35	49.36
3.7	49.37	49.38	49.39	49.41	49.42	49.43	49.44	49.45	49.46	49.47
3.8	49.48	49.49	49.50	49.51	49.52	49.53	49.54	49.55	49.56	49.57
3.9	49.57	49.58	49.59	49.60	49.61	49.61	49.62	49.63	49.64	49.64
4.0	49.65	49.66	49.67	49.67	49.68	49.68	49.69	49.70	49.70	49.71
4.5	49.88									
5.0	49.963									
5.5	49.9896									
6.0	49.9974									
7.0	49.99988									
8.0	49.9999966									

[1] Adapted from R. H. Krause and H. S. Conrad, "A Seven-Decimal Table of the Area (a) under the Unit Normal Curve, etc." *Psychometrika*, 1937, 2:55–66.

TABLE II

TABLE OF SQUARES AND SQUARE ROOTS OF THE NUMBERS FROM 1 TO 1000

Number	Square	Square Root	Number	Square	Square Root
1	1	1.000	51	26 01	7.141
2	4	1.414	52	27 04	7.211
3	9	1.732	53	28 09	7.280
4	16	2.000	54	29 16	7.348
5	25	2.236	55	30 25	7.416
6	36	2.449	56	31 36	7.483
7	49	2.646	57	32 49	7.550
8	64	2.828	58	33 64	7.616
9	81	3 000	59	34 81	7.681
10	1 00	3.162	60	36 00	7.746
11	1 21	3.317	61	37 21	7.810
12	1 44	3.464	62	38 44	7.874
13	1 69	3 606	63	39 69	7.937
14	1 96	3.742	64	40 96	8.000
15	2 25	3 873	65	42 25	8.062
16	2 56	4 000	66	43 56	8.124
17	2 89	4.123	67	44 89	8.185
18	3 24	4.243	68	46 24	8.246
19	3 61	4.359	69	47 61	8.307
20	4 00	4 472	70	49 00	8.367
21	4 41	4.583	71	50 41	8.426
22	4 84	4.690	72	51 84	8.485
23	5 29	4.796	73	53 29	8.544
24	5 76	4.899	74	54 76	8.602
25	6 25	5.000	75	56 25	8.660
26	6 76	5.099	76	57 76	8.718
27	7 29	5.196	77	59 29	8.775
28	7 84	5.292	78	60 84	8.832
29	8 41	5.385	79	62 41	8.888
30	9 00	5.477	80	64 00	8.944
31	9 61	5.568	81	65 61	9.000
32	10 24	5.657	82	67 24	9.055
33	10 89	5.745	83	68 89	9.110
34	11 56	5.831	84	70 56	9.165
35	12 25	5.916	85	72 25	9.220
36	12 96	6.000	86	73 96	9.274
37	13 69	6.083	87	75 69	9.327
38	14 44	6.164	88	77 44	9.381
39	15 21	6.245	89	79 21	9.434
40	16 00	6.325	90	81 00	9.487
41	16 81	6.403	91	82 81	9.539
42	17 64	6.481	92	84 64	9.592
43	18 49	6.557	93	86 49	9.644
44	19 36	6.633	94	88 36	9.695
45	20 25	6.708	95	90 25	9.747
46	21 16	6.782	96	92 16	9.798
47	22 09	6.856	97	94 09	9.849
48	23 04	6.928	98	96 04	9.899
49	24 01	7.000	99	98 01	9.950
50	25 00	7.071	100	1 00 00	10.000

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
101	1 02 01	10.050	151	2 28 01	12.288
102	1 04 04	10.100	152	2 31 04	12.329
103	1 06 09	10.149	153	2 34 09	12.369
104	1 08 16	10 198	154	2 37 16	12.410
105	1 10 25	10.247	155	2 40 25	12.450
106	1 12 36	10.296	156	2 43 36	12.490
107	1 14 49	10.344	157	2 46 49	12.530
108	1 16 64	10 392	158	2 49 64	12.570
109	1 18 81	10 440	159	2 52 81	12.610
110	1 21 00	10.488	160	2 56 00	12.649
111	1 23 21	10.536	161	2 59 21	12.689
112	1 25 44	10 583	162	2 62 44	12.728
113	1 27 69	10 630	163	2 65 69	12.767
114	1 29 96	10 677	164	2 68 96	12.806
115	1 32 25	10 724	165	2 72 25	12.845
116	1 34 56	10 770	166	2 75 56	12 884
117	1 36 89	10 817	167	2 78 89	12.923
118	1 39 24	10 863	168	2 82 24	12.961
119	1 41 61	10 909	169	2 85 61	13.000
120	1 44 00	10 954	170	2 89 00	13.038
121	1 46 41	11 000	171	2 92 41	13.077
122	1 48 84	11.045	172	2 95 84	13.115
123	1 51 29	11.091	173	2 99 29	13.153
124	1 53 76	11 136	174	3 02 76	13.191
125	1 56 25	11.180	175	3 06 25	13.229
126	1 58 76	11 225	176	3 09 76	13 266
127	1 61 29	11 269	177	3 13 29	13.304
128	1 63 84	11 314	178	3 16 84	13.342
129	1 66 41	11 358	179	3 20 41	13.379
130	1 69 00	11.402	180	3 24 00	13.416
131	1 71 61	11.446	181	3 27 61	13.454
132	1 74 24	11 489	182	3 31 24	13.491
133	1 76 89	11.533	183	3 34 89	13.528
134	1 79 56	11.576	184	3 38 56	13.565
135	1 82 25	11.619	185	3 42 25	13.601
136	1 84 96	11.662	186	3 45 96	13.638
137	1 87 69	11.705	187	3 49 69	13.675
138	1 90 44	11.747	188	3 53 44	13.711
139	1 93 21	11.790	189	3 57 21	13.748
140	1 96 00	11.832	190	3 61 00	13.784
141	1 98 81	11.874	191	3 64 81	13.820
142	2 01 64	11.916	192	3 68 64	13.856
143	2 04 49	11.958	193	3 72 49	13.892
144	2 07 36	12.000	194	3 76 36	13.928
145	2 10 25	12.042	195	3 80 25	13.964
146	2 13 16	12.083	196	3 84 16	14.000
147	2 16 09	12.124	197	3 88 09	14.036
148	2 19 04	12.166	198	3 92 04	14.071
149	2 22 01	12.207	199	3 96 01	14.107
150	2 25 00	12.247	200	4 00 00	14.142

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
201	4 04 01	14.177	251	6 30 01	15.843
202	4 08 04	14.213	252	6 35 04	15.875
203	4 12 09	14.248	253	6 40 09	15.906
204	4 16 16	14.283	254	6 45 16	15.937
205	4 20 25	14.318	255	6 50 25	15.969
206	4 24 36	14.353	256	6 55 36	16.000
207	4 28 49	14.387	257	6 60 49	16 031
208	4 32 64	14.422	258	6 65 64	16.062
209	4 36 81	14.457	259	6 70 81	16.093
210	4 41 00	14.491	260	6 76 00	16.125
211	4 45 21	14.526	261	6 81 21	16.155
212	4 49 44	14.560	262	6 86 44	16.186
213	4 53 69	14.595	263	6 91 69	16.217
214	4 57 96	14.629	264	6 96 96	16.248
215	4 62 25	14.663	265	7 02 25	16.279
216	4 66 56	14.697	266	7 07 56	16.310
217	4 70 89	14 731	267	7 12 89	16.340
218	4 75 24	14.765	268	7 18 24	16.371
219	4 79 61	14.799	269	7 23 61	16.401
220	4 84 00	14.832	270	7 29 00	16.432
221	4 88 41	14.866	271	7 34 41	16.462
222	4 92 84	14.900	272	7 39 84	16.492
223	4 97 29	14.933	273	7 45 29	16.523
224	5 01 76	14.967	274	7 50 76	16.553
225	5 06 25	15 000	275	7 56 25	16.583
226	5 10 76	15.033	276	7 61 76	16.613
227	5 15 29	15.067	277	7 67 29	16.643
228	5 19 84	15.100	278	7 72 84	16.673
229	5 24 41	15.133	279	7 78 41	16.703
230	5 29 00	15.166	280	7 84 00	16.733
231	5 33 61	15.199	281	7 89 61	16.763
232	5 38 24	15.232	282	7 95 24	16 793
233	5 42 89	15.264	283	8 00 89	16.823
234	5 47 56	15.297	284	8 06 56	16.852
235	5 52 25	15.330	285	8 12 25	16.882
236	5 56 96	15.362	286	8 17 96	16.912
237	5 61 69	15.395	287	8 23 69	16.941
238	5 66 44	15.427	288	8 29 44	16.971
239	5 71 21	15.460	289	8 35 21	17.000
240	5 76 00	15.492	290	8 41 00	17.029
241	5 80 81	15.524	291	8 46 81	17.059
242	5 85 64	15.556	292	8 52 64	17.088
243	5 90 49	15.588	293	8 58 49	17.117
244	5 95 36	15.620	294	8 64 36	17.146
245	6 00 25	15.652	295	8 70 25	17.176
246	6 05 16	15.684	296	8 76 16	17.205
247	6 10 09	15.716	297	8 82 09	17.234
248	6 15 04	15.748	298	8 88 04	17.263
249	6 20 01	15.780	299	8 94 01	17.292
250	6 25 00	15.811	300	9 00 00	17.321

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
301	9 06 01	17.349	351	12 32 01	18.735
302	9 12 04	17.378	352	12 39 04	18.762
303	9 18 09	17.407	353	12 46 09	18.788
304	9 24 16	17.436	354	12 53 16	18.815
305	9 30 25	17.464	355	12 60 25	18.841
306	9 36 36	17.493	356	12 67 36	18.868
307	9 42 49	17.521	357	12 74 49	18.894
308	9 48 64	17.550	358	12 81 64	18.921
309	9 54 81	17.578	359	12 88 81	18.947
310	9 61 00	17.607	360	12 96 00	18.974
311	9 67 21	17.635	361	13 03 21	19.000
312	9 73 44	17.664	362	13 10 44	19.026
313	9 79 69	17 692	363	13 17 69	19.053
314	9 85 96	17.720	364	13 24 96	19.079
315	9 92 25	17.748	365	13 32 25	19.105
316	9 98 56	17.776	366	13 39 56	19.131
317	10 04 89	17.804	367	13 46 89	19.157
318	10 11 24	17.833	368	13 54 24	19.183
319	10 17 61	17.861	369	13 61 61	19.209
320	10 24 00	17.889	370	13 69 00	19.235
321	10 30 41	17.916	371	13 76 41	19.261
322	10 36 84	17.944	372	13 83 84	19.287
323	10 43 29	17.972	373	13 91 29	19.313
324	10 49 76	18.000	374	13 98 76	19 339
325	10 56 25	18.028	375	14 06 25	19.363
326	10 62 76	18.055	376	14 13 76	19.391
327	10 69 29	18.083	377	14 21 29	19.416
328	10 75 84	18 111	378	14 28 84	19.442
329	10 82 41	18.138	379	14 36 41	19.468
330	10 89 00	18.166	380	14 44 00	19.494
331	10 95 61	18.193	381	14 51 61	19.519
332	11 02 24	18.221	382	14 59 24	19.545
333	11 08 89	18.248	383	14 66 89	19.570
334	11 15 56	18.276	384	14 74 56	19.596
335	11 22 25	18.303	385	14 82 25	19.621
336	11 28 96	18 330	386	14 89 96	19.647
337	11 35 69	18.358	387	14 97 69	19.672
338	11 42 44	18.385	388	15 05 44	19.698
339	11 49 21	18.412	389	15 13 21	19.723
340	11 56 00	18.439	390	15 21 00	19.748
341	11 62 81	18.466	391	15 28 81	19.774
342	11 69 64	18.493	392	15 36 64	19.799
343	11 76 49	18.520	393	15 44 49	19.824
344	11 83 36	18.547	394	15 52 36	19.849
345	11 90 25	18.574	395	15 60 25	19.875
346	11 97 16	18.601	396	15 68 16	19.900
347	12 04 09	18.628	397	15 76 09	19.925
348	12 11 04	18.655	398	15 84 04	19.950
349	12 18 01	18.682	399	15 92 01	19.975
350	12 25 00	18.708	400	16 00 00	20.000

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
401	16 08 01	20.025	451	20 34 01	21.237
402	16 16 04	20.050	452	20 43 04	21.260
403	16 24 09	20.075	453	20 52 09	21.284
404	16 32 16	20.100	454	20 61 16	21.307
405	16 40 25	20.125	455	20 70 25	21.331
406	16 48 36	20 149	456	20 79 36	21.354
407	16 56 49	20 174	457	20 88 49	21.378
408	16 64 64	20.199	458	20 97 64	21.401
409	16 72 81	20.224	459	21 06 81	21.424
410	16 81 00	20.248	460	21 16 00	21.448
411	16 89 21	20.273	461	21 25 21	21.471
412	16 97 44	20.298	462	21 34 44	21.494
413	17 05 69	20 322	463	21 43 69	21.517
414	17 13 96	20.347	464	21 52 96	21.541
415	17 22 25	20 372	465	21 62 25	21.564
416	17 30 56	20.396	466	21 71 56	21.587
417	17 38 89	20.421	467	21 80 89	21.610
418	17 47 24	20.445	468	21 90 24	21.633
419	17 55 61	20.469	469	21 99 61	21.656
420	17 64 00	20.494	470	22 09 00	21.679
421	17 72 41	20.518	471	22 18 41	21.703
422	17 80 84	20 543	472	22 27 84	21.726
423	17 89 29	20.567	473	22 37 29	21.749
424	17 97 76	20.591	474	22 46 76	21.772
425	18 06 25	20.616	475	22 56 25	21.794
426	18 14 76	20 640	476	22 65 76	21.817
427	18 23 29	20.664	477	22 75 29	21.840
428	18 31 84	20.688	478	22 84 84	21.863
429	18 40 41	20 712	479	22 94 41	21.886
430	18 49 00	20.736	480	23 04 00	21.909
431	18 57 61	20.761	481	23 13 61	21.932
432	18 66 24	20.785	482	23 23 24	21.954
433	18 74 89	20.809	483	23 32 89	21.977
434	18 83 56	20.833	484	23 42 56	22.000
435	18 92 25	20.857	485	23 52 25	22.023
436	19 00 96	20.881	486	23 61 96	22.045
437	19 09 69	20.905	487	23 71 69	22.068
438	19 18 44	20.928	488	23 81 44	22.091
439	19 27 21	20.952	489	23 91 21	22.113
440	19 36 00	20.976	490	24 01 00	22.136
441	19 44 81	21.000	491	24 10 81	22.159
442	19 53 64	21.024	492	24 20 64	22.181
443	19 62 49	21.048	493	24 30 49	22.204
444	19 71 36	21.071	494	24 40 36	22.226
445	19 80 25	21.095	495	24 50 25	22.249
446	19 89 16	21.119	496	24 60 16	22.271
447	19 98 09	21.142	497	24 70 09	22.293
448	20 07 04	21.166	498	24 80 04	22.316
449	20 16 01	21.190	499	24 90 01	22.338
450	20 25 00	21.213	500	25 00 00	22.361

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
501	25 10 01	22.383	551	30 36 01	23.473
502	25 20 04	22.405	552	30 47 04	23.495
503	25 30 09	22.428	553	30 58 09	23.516
504	25 40 16	22.450	554	30 69 16	23.537
505	25 50 25	22.472	555	30 80 25	23.558
506	25 60 36	22.494	556	30 91 36	23.580
507	25 70 49	22.517	557	31 02 49	23.601
508	25 80 64	22.539	558	31 13 64	23.622
509	25 90 81	22.561	559	31 24 81	23.643
510	26 01 00	22.583	560	31 36 00	23.664
511	26 11 21	22.605	561	31 47 21	23.685
512	26 21 44	22.627	562	31 58 44	23.707
513	26 31 69	22 650	563	31 69 69	23.728
514	26 41 96	22.672	564	31 80 96	23 749
515	26 52 25	22 694	565	31 92 25	23.770
516	26 62 56	22 716	566	32 03 56	23.791
517	26 72 89	22.738	567	32 14 89	23.812
518	26 83 24	22 760	568	32 26 24	23.833
519	26 93 61	22.782	569	32 37 61	23.854
520	27 04 00	22.804	570	32 49 00	23.875
521	27 14 41	22 825	571	32 60 41	23.896
522	27 24 84	22.847	572	32 71 84	23.917
523	27 35 29	22.869	573	32 83 29	23.937
524	27 45 76	22.891	574	32 94 76	23.958
525	27 56 25	22 913	575	33 06 25	23 979
526	27 66 76	22.935	576	33 17 76	24.000
527	27 77 29	22.956	577	33 29 29	24.021
528	27 87 84	22.978	578	33 40 84	24.042
529	27 98 41	23.000	579	33 52 41	24.062
530	28 09 00	23 022	580	33 64 00	24.083
531	28 19 61	23.043	581	33 75 61	24.104
532	28 30 24	23 065	582	33 87 24	24.125
533	28 40 89	23.087	583	33 98 89	24.145
534	28 51 56	23.108	584	34 10 56	24.166
535	28 62 25	23.130	585	34 22 25	24.187
536	28 72 96	23 152	586	34 33 96	24.207
537	28 83 69	23.173	587	34 45 69	24.228
538	28 94 44	23.195	588	34 57 44	24.249
539	29 05 21	23.216	589	34 69 21	24.269
540	29 16 00	23.238	590	34 81 00	24.290
541	29 26 81	23.259	591	34 92 81	24.310
542	29 37 64	23.281	592	35 04 64	24.331
543	29 48 49	23.302	593	35 16 49	24.352
544	29 59 36	23.324	594	35 28 36	24.372
545	29 70 25	23.345	595	35 40 25	24.393
546	29 81 16	23.367	596	35 52 16	24.413
547	29 92 09	23.388	597	35 64 09	24.434
548	30 03 04	23.409	598	35 76 04	24.454
549	30 14 01	23.431	599	35 88 01	24.474
550	30 25 00	23.452	600	36 00 00	24.495

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
601	36 12 01	24.515	651	42 38 01	25.515
602	36 24 04	24.536	652	42 51 04	25.534
603	36 36 09	24.556	653	42 64 09	25.554
604	36 48 16	24.576	654	42 77 16	25.573
605	36 60 25	24.597	655	42 90 25	25.593
606	36 72 36	24.617	656	43 03 36	25.612
607	36 84 49	24.637	657	43 16 49	25.632
608	36 96 64	24.658	658	43 29 64	25.652
609	37 08 81	24.678	659	43 42 81	25.671
610	37 21 00	24.698	660	43 56 00	25.690
611	37 33 21	24.718	661	43 69 21	25.710
612	37 45 44	24.739	662	43 82 44	25.729
613	37 57 69	24.759	663	43 95 69	25.749
614	37 69 96	24.779	664	44 08 96	25.768
615	37 82 25	24.799	665	44 22 25	25.788
616	37 94 56	24.819	666	44 35 56	25.807
617	38 06 89	24.839	667	44 48 89	25.826
618	38 19 24	24 860	668	44 62 24	25.846
619	38 31 61	24.880	669	44 75 61	25 865
620	38 44 00	24.900	670	44 89 00	25 884
621	38 56 41	24.920	671	45 02 41	25.904
622	38 68 84	24.940	672	45 15 84	25.923
623	38 81 29	24.960	673	45 29 29	25.942
624	38 93 76	24.980	674	45 42 76	25.962
625	39 06 25	25.000	675	45 56 25	25.981
626	39 18 76	25.020	676	45 69 76	26.000
627	39 31 29	25.040	677	45 83 29	26.019
628	39 43 84	25.060	678	45 96 84	26 038
629	39 56 41	25.080	679	46 10 41	26.058
630	39 69 00	25.100	680	46 24 00	26.077
631	39 81 61	25.120	681	46 37 61	26.096
632	39 94 24	25.140	682	46 51 24	26.115
633	40 06 89	25.159	683	46 64 89	26.134
634	40 19 56	25.179	684	46 78 56	26.153
635	40 32 25	25.199	685	46 92 25	26.173
636	40 44 96	25.219	686	47 05 96	26.192
637	40 57 69	25.239	687	47 19 69	26.211
638	40 70 44	25.259	688	47 33 44	26.230
639	40 83 21	25.278	689	47 47 21	26.249
640	40 96 00	25.298	690	47 61 00	26.268
641	41 08 81	25.318	691	47 74 81	26.287
642	41 21 64	25.338	692	47 88 64	26.306
643	41 34 49	25.357	693	48 02 49	26.325
644	41 47 36	25.377	694	48 16 36	26.344
645	41 60 25	25.397	695	48 30 25	26.363
646	41 73 16	25.417	696	48 44 16	26.382
647	41 86 09	25.436	697	48 58 09	26.401
648	41 99 04	25.456	698	48 72 04	26.420
649	42 12 01	25.475	699	48 86 01	26.439
650	42 25 00	25.495	700	49 00 00	26.458

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
701	49 14 01	26.476	751	56 40 01	27.404
702	49 28 04	26.495	752	56 55 04	27.423
703	49 42 09	26.514	753	56 70 09	27.441
704	49 56 16	26.533	754	56 85 16	27.459
705	49 70 25	26.552	755	57 00 25	27.477
706	49 84 36	26.571	756	57 15 36	27.495
707	49 98 49	26.589	757	57 30 49	27.514
708	50 12 64	26 608	758	57 45 64	27.532
709	50 26 81	26.627	759	57 60 81	27.550
710	50 41 00	26.646	760	57 76 00	27.568
711	50 55 21	26.665	761	57 91 21	27.586
712	50 69 44	26.683	762	58 06 44	27.604
713	50 83 69	26.702	763	58 21 69	27.622
714	50 97 96	26 721	764	58 36 96	27.641
715	51 12 25	26.739	765	58 52 25	27.659
716	51 26 56	26 758	766	58 67 56	27.677
717	51 40 89	26 777	767	58 82 89	27.695
718	51 55 24	26.796	768	58 98 24	27.713
719	51 69 61	26.814	769	59 13 61	27.731
720	51 84 00	26.833	770	59 29 00	27.749
721	51 98 41	26.851	771	59 44 41	27.767
722	52 12 84	26.870	772	59 59 84	27.785
723	52 27 29	26.889	773	59 75 29	27.803
724	52 41 76	26.907	774	59 90 76	27.821
725	52 56 25	26.926	775	60 06 25	27.839
726	52 70 76	26.944	776	60 21 76	27.857
727	52 85 29	26.963	777	60 37 29	27.875
728	52 99 84	26 981	778	60 52 84	27.893
729	53 14 41	27.000	779	60 68 41	27.911
730	53 29 00	27.019	780	60 84 00	27.928
731	53 43 61	27.037	781	60 99 61	27.946
732	53 58 24	27.055	782	61 15 24	27.964
733	53 72 89	27.074	783	61 30 89	27.982
734	53 87 56	27.092	784	61 46 56	28.000
735	54 02 25	27.111	785	61 62 25	28.018
736	54 16 96	27.129	786	61 77 96	28.036
737	54 31 69	27.148	787	61 93 69	28.054
738	54 46 44	27.166	788	62 09 44	28.071
739	54 61 21	27.185	789	62 25 21	28.089
740	54 76 00	27.203	790	62 41 00	28.107
741	54 90 81	27.221	791	62 56 81	28.125
742	55 05 64	27.240	792	62 72 64	28.142
743	55 20 49	27.258	793	62 88 49	28.160
744	55 35 36	27.276	794	63 04 36	28.178
745	55 50 25	27.295	795	63 20 25	28.196
746	55 65 16	27.313	796	63 36 16	28.213
747	55 80 09	27.331	797	63 52 09	28.231
748	55 95 04	27.350	798	63 68 04	28.249
749	56 10 01	27.368	799	63 84 01	28.267
750	56 25 00	27.386	800	64 00 00	28.284

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
801	64 16 01	28.302	851	72 42 01	29.172
802	64 32 04	28 320	852	72 59 04	29.189
803	64 48 09	28.337	853	72 76 09	29.206
804	64 64 16	28.355	854	72 93 16	29.223
805	64 80 25	28.373	855	73 10 25	29.240
806	64 96 36	28 390	856	73 27 36	29.257
807	65 12 49	28 408	857	73 44 49	29.275
808	65 28 64	28 425	858	73 61 64	29 292
809	65 44 81	28.443	859	73 78 81	29.309
810	65 61 00	28.460	860	73 96 00	29 326
811	65 77 21	28 478	861	74 13 21	29.343
812	65 93 44	28.496	862	74 30 44	29.360
813	66 09 69	28 513	863	74 47 69	29 377
814	66 25 96	28 531	864	74 64 96	29.394
815	66 42 25	28.548	865	74 82 25	29.411
816	66 58 56	28 566	866	74 99 56	29.428
817	66 74 89	28 583	867	75 16 89	29.445
818	66 91 24	28 601	868	75 34 24	29.462
819	67 07 61	28.618	869	75 51 61	29 479
820	67 24 00	28.636	870	75 69 00	29.496
821	67 40 41	28.653	871	75 86 41	29.513
822	67 56 84	28.671	872	76 03 84	29.530
823	67 73 29	28.688	873	76 21 29	29.547
824	67 89 76	28.705	874	76 38 76	29.563
825	68 06 25	28.723	875	76 56 25	29.580
826	68 22 76	28.740	876	76 73 76	29 597
827	68 39 29	28.758	877	76 91 29	29.614
828	68 55 84	28.775	878	77 08 84	29.631
829	68 72 41	28.792	879	77 26 41	29 648
830	68 89 00	28.810	880	77 44 00	29.665
831	69 05 61	28.827	881	77 61 61	29.682
832	69 22 24	28.844	882	77 79 24	29.698
833	69 38 89	28 862	883	77 96 89	29.715
834	69 55 56	28.879	884	78 14 56	29.732
835	69 72 25	28.896	885	78 32 25	29.749
836	69 88 96	28.914	886	78 49 96	29.766
837	70 05 69	28 931	887	78 67 69	29.783
838	70 22 44	28 948	888	78 85 44	29.799
839	70 39 21	28.965	889	79 03 21	29.816
840	70 56 00	28.983	890	79 21 00	29.833
841	70 72 81	29.000	891	79 38 81	29.850
842	70 89 64	29.017	892	79 56 64	29.866
843	71 06 49	29.034	893	79 74 49	29.883
844	71 23 36	29.052	894	79 92 36	29.900
845	71 40 25	29.069	895	80 10 25	29.916
846	71 57 16	29.086	896	80 28 16	29.933
847	71 74 09	29.103	897	80 46 09	29.950
848	71 91 04	29.120	898	80 64 04	29.967
849	72 08 01	29.138	899	80 82 01	29.983
850	72 25 00	29.155	900	81 00 00	30.000

TABLE OF SQUARES AND SQUARE ROOTS — *Continued*

Number	Square	Square Root	Number	Square	Square Root
901	81 18 01	30 017	951	90 44 01	30.838
902	81 36 04	30 033	952	90 63 04	30.854
903	81 54 09	30.050	953	90 82 09	30.871
904	81 72 16	30.067	954	91 01 16	30.887
905	81 90 25	30.083	955	91 20 25	30.903
906	82 08 36	30.100	956	91 39 36	30.919
907	82 26 49	30.116	957	91 58 49	30.935
908	82 44 64	30 133	958	91 77 64	3c.952
909	82 62 81	30 150	959	91 96 81	30 968
910	82 81 00	30.166	960	92 16 00	30.984
911	82 99 21	30 183	961	92 35 21	31.000
912	83 17 44	30 199	962	92 54 44	31.016
913	83 35 69	30.216	963	92 73 69	31.032
914	83 53 96	30 232	964	92 92 96	31.048
915	83 72 25	30.249	965	93 12 25	31.064
916	83 90 56	30.265	966	93 31 56	31.081
917	84 08 89	30 282	967	93 50 89	31.097
918	84 27 24	30.299	968	93 70 24	31.113
919	84 45 61	30.315	969	93 89 61	31 129
920	84 64 00	30 332	970	94 09 00	31.145
921	84 82 41	30 348	971	94 28 41	31.161
922	85 00 84	30.364	972	94 47 84	31 177
923	85 19 29	30 381	973	94 67 29	31.193
924	85 37 76	30 397	974	94 86 76	31.209
925	85 56 25	30.414	975	95 06 25	31 225
926	85 74 76	30.430	976	95 25 76	31.241
927	85 93 29	30 447	977	95 45 29	31.257
928	86 11 84	30 463	978	95 64 84	31.273
929	86 30 41	30 480	979	95 84 41	31.289
930	86 49 00	30.496	980	96 04 00	31.305
931	86 67 61	30.512	981	96 23 61	31.321
932	86 86 24	30 529	982	96 43 24	31.337
933	87 04 89	30 545	983	96 62 89	31.353
934	87 23 56	30 561	984	96 82 56	31.369
935	87 42 25	30.578	985	97 02 25	31.385
936	87 60 96	30 594	986	97 21 96	31.401
937	87 79 69	30 610	987	97 41 69	31.417
938	87 98 44	30 627	988	97 61 44	31.432
939	88 17 21	30.643	989	97 81 21	31.448
940	88 36 00	30.659	990	98 01 00	31.464
941	88 54 81	30 676	991	98 20 81	31.480
942	88 73 64	30 692	992	98 40 64	31.496
943	88 92 49	30.708	993	98 60 49	31.512
944	89 11 36	30.725	994	98 80 36	31.528
945	89 30 25	30 741	995	99 00 25	31.544
946	89 49 16	30.757	996	99 20 16	31.559
947	89 68 09	30.773	997	99 40 09	31.575
948	89 87 04	30 790	998	99 60 04	31.591
949	90 06 01	30 806	999	99 80 01	31.607
950	90 25 00	30 822	1000	100 00 00	31 623

TABLE III

MINIMUM VALUES OF SIGNIFICANCE RATIO REQUIRED FOR SIGNIFICANCE AT VARIOUS LEVELS[1]

Degrees of Freedom ($N-1$)	Levels of Significance				
	20%	5%	2%	1%	0 1%
1	3.078	12.706	31.821	63.657	636.619
2	1.886	4.303	6.965	9.925	31.598
3	1.638	3.182	4.541	5.841	12.941
4	1.533	2.776	3.747	4.604	8.610
5	1.476	2.571	3.365	4.032	6.859
6	1.440	2.447	3.143	3.707	5.959
7	1.415	2.365	2.998	3.499	5.405
8	1.397	2.306	2 896	3.355	5.041
9	1.383	2.262	2.821	3.250	4.781
10	1.372	2.228	2.764	3.169	4.587
11	1.363	2.201	2.718	3.106	4.437
12	1.356	2.179	2.681	3.055	4.318
13	1.350	2.160	2.650	3.012	4.221
14	1.345	2.145	2.624	2.977	4.140
15	1.341	2.131	2.602	2.947	4.073
16	1.337	2.120	2.583	2.921	4.015
17	1.333	2.110	2.567	2.898	3.965
18	1.330	2.101	2.552	2.878	3.922
19	1.328	2.093	2.539	2 861	3.883
20	1.325	2.086	2.528	2.845	3.850
21	1.323	2.080	2.518	2.831	3.819
22	1.321	2.074	2.508	2.819	3 792
23	1.319	2.069	2.500	2.807	3.767
24	1.318	2.064	2.492	2.797	3.745
25	1.316	2.060	2.485	2.787	3.725
26	1.315	2.056	2.479	2.779	3.707
27	1.314	2.052	2.473	2.771	3.690
28	1.313	2.048	2.467	2.763	3.674
29	1.311	2.045	2 462	2.756	3.659
30	1.310	2.042	2.457	2.750	3.646
40	1.303	2.021	2.423	2.704	3.551
60	1.296	2.000	2 390	2.660	3.460
120	1.289	1.980	2 358	2.617	3.373
∞	1.282	1.960	2.326	2.576	3.291

[1] This table is taken by consent from *Statistical Tables for Biological, Agricultural and Medical Research* by Professor R. A. Fisher and F. Yates, published at 13/6 by Oliver & Boyd Ltd., Edinburgh.

INDEX

Arithmetic mean, *see* Mean
Average deviation, *see* Mean deviation
Averages, 51–68; *see* Mean, Median, Mode

Bias in sampling, 141

Central tendency, measures of, 51–58; *see* Mean, Median, Mode
Class interval, size of, 16–17, 19–21; limits of, 17, 20–21, 24–27; midpoint of, 27
Coefficient of correlation, meaning of, 160–166, 198–204; computation of, 167–174; as measure of regression, 175–182; in regression equations, 164–167; as measure of reliability of prediction, 186–190; assumption of rectilinearity, 190–191; reliability of, 191–195; influence of range of talent on, 195–198; as measure of test validity, 213–215; as measure of test reliability, 216–218
Coefficient of reliability, 216–218; ways of estimating, 218–220; significance of in test evaluation, 223–227
Column diagram, *see* Histogram
Comparable measures, *see* Percentile rank *and* Standard scores
Composite measures, 150–152
Confidence, levels of, 104–106
Confidence interval, for true mean, 106–108, 118–121, 134–136; for true proportion, 123–127
Continuous series, 24
Correlation, meaning of, 153–160; linear, 156; non-linear, 156; *see also* Coefficient of correlation
Crude mode, 61
Cumulative frequency, curve of, 42–48
Curves, types of, 48–50; normal probability, 81–101
Curvilinear relationship, 156, 190

Data, continuous and discrete, 24
Deciles, 32
Deviation, *see* Variability
Differences, reliability of, 129–136, 138–139; *see* Standard error
Discrete data, 24
Distribution, frequency, *see* Frequency distribution

Errors, in rounding numbers, 61–66; importance of in statistical work, 67; of sampling, 102–143; of estimate, 186–190; of measurement, 205–213, 220–223; *see also* Standard error

Fitting, method of, normal curve to observed distributions, 99–101
Frequencies, reliability of percentage, 125–129
Frequency distribution, need for, 11–12; construction of, 13–28; graphic representation of, 39–50; cumulative, 42–46; types of, 48–49

Graphs, *see* Histogram, Polygon, Cumulative frequency curve
Grouping, in frequency distributions, 13 ff.; natural, 21

Histogram, 39–41, 45–46; uses of, 50
Hypotheses, testing, 110–115, 130–136, 138–139

Index of reliability, 223 (footnote)
Interval, *see* Class interval

Levels of confidence, 106–108

Mean, definition of, 52; computation of, 52–60; number of significant figures in, 61–66; uses of, 68; reliability of, 106–123, 136–138
Mean deviation, definition of, 70; computation of, 71–74; characteristics of, 77–80
Measurement, nature of in education and psychology, 205–213
Median, definition of, 60–61; computation of, 27; uses of, 68; reliability of, 124
Median deviation, definition of, 70; characteristics of, 77–80
Mode, 61; uses of, 68

Normal probability curve, definition of, 81; properties of, 81–85; ordinates under, 82–84; area relationships, 85–87; uses of in type problems, 88–92; significance of, 93–99
Normality, law of, 95–99
Null hypothesis, 120–136, 138–139

Ogive, 42–46

Ordinates, relations between, of normal curve, 82–84

Percentages, reliability of, 125–129
Percentile curve, 42–46
Percentile ranks, 32–33; computation of, 33–35; use and interpretation of, 37–38
Percentiles, 32–33; computation of, 35–37; use and interpretation of, 37–38
Polygon, frequency, 41–42, 45–46; smoothing, 46–48
Prediction, uses of *r* in, 158–160; use of regression equations in, 184–192; reliability of, 186–190
Probable error, 123, 125; *see* Standard error
Product-moment correlation coefficient, *see* Coefficient of correlation
Proportion, reliability of, 125–129

Quartile deviation, *see* Semi-interquartile range
Quartiles, 32

Range, of a frequency distribution, 17; as a measure of variability, 69
Range of talent, effect on *r*, 197–200
Ranks, 32
Rectilinearity, 156; assumption of, 190
Regression, phenomenon of, 175–182; equations in z-score form, 182; in raw score form, 185; uses of in prediction, 182–190
Reliability, meaning of, in sampling, 100–101; of the mean, 100–121, 134–136; of the median, 124; of the quartile deviation, 124; of the standard deviation, 124; of differences, uncorrelated, 129–133; of correlated differences, 134–136, 138–139; of per cents and proportions, 125–129; of test scores, 215–227; coefficient of, 216–220; significance of in test evaluation, 223–227

Sampling distribution of mean, 104–105
Sampling, investigation by, 102 ff.; random, 103, 140; errors in, 102 ff.; biased, 141; controlled, 142; methods of, 139–143; in test construction, 208–211
Scatter-diagram, 154–155
Scores, standard, 145
Semi-interquartile range, 70; reliability of, 124
Series, continuous and discrete, 24
"Short" method, of computing mean, 55–60; of computing mean deviation, 71–74; standard deviation, 75–77; coefficient of correlation, 167–174
Significance, statistical, 130–133
Significance ratio, 132
Significant, differences, 130–133; correlation, 191–195
Significant digits, 61–66
Skewness, 48–49
Small sample theory, 136–139
Smoothing frequency distributions, 46–48
Spearman, product-moment correlation coefficient, *see* Coefficient of correlation
Spearman-Brown prophecy formula, 218–220
Standard deviation, definition of, 71; computation of, 75–77; characteristics of, 77–80; reliability of, 124
Standard error, of mean, 106, 115–124; of median, 124; of quartile deviation, 124; of standard deviation, 124; of percentages and proportions, 125–129; of differences, 129–136; of correlation coefficient, 191–195; of estimates based on regression equations, 186–190; of measurement (of test scores), 220–223
Standard or z-scores, 145–146; computation of, 147–150; used in securing composites, 150–152
Statistics, purposes of, 1–3; major aspects of instruction in, 3–6; organization of instructional materials, 6–9; how to study, 9–10

Tabulation, of measures in frequency distributions, 11 ff.
Test scales, characteristics of, 29–31, 205–213
Tests, uses of correlation in evaluating, 205–227
"True" scores, 220
T-scales, 149

Validity, meaning of, 213; measurement of, 214–215; relation of, to reliability, 223–227
Variability, measures of, 69–80; *see also* Quartile deviation, Mean deviation, Median deviation, Standard deviation