Elements
of
Statistical Method

by

Albert E. Waugh
Professor of Economics, University of Connecticut

SECOND EDITION
SIXTH IMPRESSION

McGRAW-HILL BOOK COMPANY, INC.
NEW YORK AND LONDON
1943

ELEMENTS OF STATISTICAL METHOD

PRINTED IN THE UNITED STATES OF AMERICA

This book is affectionately

dedicated to

ALEXANDER E. CANCE

and

IRVING G. DAVIS

inspired teachers, tireless seekers after
new things, gentlemen

PREFACE TO THE SECOND EDITION

It is now five years since the first edition of this book appeared. I was gratified at the time by the reception accorded it by reviewers, and I have been pleased ever since to find a continuing demand which indicates that my colleagues who teach beginning courses in statistical method have found it useful. Several of these teachers have been kind enough to send me suggestions, many of which have been incorporated in this revision. To them I give my deepest thanks. I am especially indebted to Professor Roy W. Jastram of Stanford University, who has read the entire manuscript of this new edition. His knowledge of statistical methods and his insight into the problems of teaching have made his suggestions particularly valuable.

One of the most difficult problems in writing any elementary textbook is to keep it truly elementary—to write for beginning students rather than for one's colleagues. One is always tempted to include material merely because he finds it interesting without considering whether it should properly be treated in the first course or not. No beginning student anywhere is going to succeed in mastering more than a limited number of concepts in his first course, no matter how able his teacher is or how lucidly his textbook presents the material.

Yet quite naturally no two teachers approach the same subject in exactly the same way, and the elementary principles that one instructor decides to include in his introductory course will not be identical in all respects with those listed as essential by another. Much of the new material in this revised edition has been inserted at the request of other teachers who use the book, and in several such cases their arguments have been weighty enough to induce me to insert the material not only in the book but also in my own courses. To some extent these additions have been offset by omitting material that appeared in the first edition which seemed to be of minor importance.

As the book has grown somewhat larger, it has seemed wise to adopt the scheme of using numbered paragraph headings, so

that teachers wishing to assign parts of chapters or wishing to omit certain materials may refer more easily to the parts in question. At the same time, the tables and figures have been renumbered by chapters, so that Table 8.13, for example, is the 13th table of Chap. VIII. In several cases where instructors had found the chapters of the original edition too long for reasonable assignments, the original chapters have been broken up into two or three shorter chapters, although the numbering of the sections should in itself make the problem of assignments a good deal easier.

So much for the mechanical aspects of the alterations. The interested reader will easily discover the nature of the changes in content for himself. There are no changes in the first chapter save a few minor ones of purely verbal nature. The second chapter has been completely rewritten and somewhat enlarged. I have been pleased with the interest which both teachers and students have shown in the subject matter of this chapter, which is omitted entirely from many texts. Unfortunately the treatment in the first edition was not entirely precise, and I have attempted in this revision to put it in such form that it will not be open to misinterpretation. The third chapter, on frequency distributions, is also almost entirely rewritten and considerably enlarged.

The chapter on Measures of Central Tendency, which appeared as Chap. IV in the first edition, has here been broken down into two chapters; yet there has been relatively little expansion. In fact, the treatment of the mode has been condensed somewhat, since none of the measures ordinarily treated in elementary courses is reliable enough to warrant extended treatment. Somewhat more attention than before has been given to various sorts of weighted averages, and the attempt has been made to state in simple terms the advantages and disadvantages of the various sorts of averages and the relationships among them. The treatment of dispersion is changed but slightly, with the addition of a mere mention of the analysis of variance without actual treatment of the problem. Also both here and in the treatment of averages, simple checks on the accuracy of arithmetical computations have been explained.

Former Chap. VI now appears as Chaps. VII and VIII. The most important additions have been the inclusion of a short

section on the use of probability paper, which, for most beginning students, is a far easier method of testing for approximate normality of distribution than are the more mathematical methods; and the addition of material on the use of the chi-square test. I had formerly considered that the chi-square test was too advanced for inclusion in the beginning course, but on the suggestion of other teachers I have tried it with my own students for several years now and find not only that they have no trouble in absorbing the simpler aspects of the problem covered here, but also that it makes the purpose of fitting frequency curves clearer to them.

The new Chap. IX is little different from the old Chap. VII. Again there is the addition of mere mention of analysis of variance. Former Chap. VIII now appears as three chapters, numbered X, XI, and XII, with the addition of a good deal of new material in the first two of these new chapters, especially on the description of the simpler curvilinear trends and on the use of link relatives. The chapters on correlation reappear without significant change, as do the concluding chapters on Tabulation and Graphic Presentation and on the Collection and Analysis of Data.

The reader may at first wonder over the inclusion in an elementary book of material on analysis of variance, on multiple and joint and curvilinear correlation, etc. But it seems to the author that a beginning course should not only teach the student how to do a few elementary things, but should also put him on his guard against some of the commoner misuses of data and should prepare him for some of the concepts which he is likely to meet in more advanced work. Even a cursory examination of Chaps. X and XI, for example, will convince the reader that it is possible to tell a beginning student something about the nature of the problems of curvilinear and multiple correlation without actually taking up the mechanics with him. No student can know when it is wise and safe to use the methods of simple linear correlation unless he is at least aware of the fact that the sorts of situations described in these two chapters exist. Even if these two short chapters cannot be assigned *in toto*, the assignment of Secs. 14.1, 14.6, 15.1, 15.2, and 15.4, accompanied by a half-hour's lecture on the part of the instructor, should serve to protect the student and the general public against some of the commoner sorts of

errors in correlation work which have been perpetrated innocently by students who think of the methods of simple linear correlation as being *the* method which is always applicable when one wishes to investigate relationship between variables. And similarly, it has seemed too bad, when treating the significance of the difference between means, not to go at least far enough to let the student see that the concept can be broadened to cover those cases where several such means are being compared at once, as is done in the analysis of variance. In other words, the attempt has been made in this book not only to cover the elementary methods of statistics, but also to describe the nature of statistical concepts in general enough terms so that the student will be able to see the limitations as well as the usefulness of the methods discussed. The treatment of the more "advanced" topics does not go farther than this.

Finally, throughout the revision, the attempt has been made to retain the simple style of presentation best adapted to the beginner without sacrificing precision of statement or accuracy of treatment. The author will appreciate it greatly if users of the book will call his attention to any errors, since experience teaches that it is well-nigh impossible to eliminate them entirely.

ALBERT E. WAUGH.

UNIVERSITY OF CONNECTICUT,
STORRS, CONN.,
June, 1943.

PREFACE TO THE FIRST EDITION

This book is planned for the beginner in the field of statistics who has yet to learn "what it is all about." No attempt has been made to treat any aspect of the field exhaustively, and advanced students will find it necessary to consult other books and, particularly, to acquaint themselves with articles in the current technical statistical journals. The aim of this book is to introduce the student to statistical concepts and statistical nomenclature and to get him to think in statistical terms.

The book is not planned for the statistician of any particular field. It is not a book on business statistics or vital statistics or biometry. Its purpose is, rather, to present the statistical concepts on their own merits, and the illustrative materials are carefully chosen from diverse fields. Such a presentation helps the student to discover the wide range of usefulness of the tools with which he is working. It has been the author's purpose also to make clear from time to time the fact that one cannot safely apply statistical method in any field unless and until he has become a master of that field.

The few short problems which are given at the ends of the chapters in this book are not to be thought of as distinct from the text that precedes them. The attempt has been made to give the student new viewpoints on the subject matter of the chapters by the use of carefully selected questions which are not, in the main, mathematical in nature. Most of the questions can be answered by thoughtful reconsideration of the principles enunciated in the preceding chapters.

The student should not expect, of course, to confine himself to the short simple questions and problems included here. Additional problems can be formulated with ease by either student or teacher, and the morning's newspaper or current magazines will furnish abundant and interesting grist for the statistician's mill. It is decidedly important that the student should solve many problems in each section of the field that is studied. The use of some one of the admirable manuals of

statistical problems that are available is recommended.[1] But such problems cannot in any sense replace the exercises included here. These latter exercises serve not so much to give facility in numerical manipulation as to clear away mental cobwebs and make the statistical concepts stand out sharply.

In the writing of this book no attempt has been made either to omit all mathematical treatment or to include it all. The book is certainly not a catalogue in which the advanced student can find a compilation of mathematical proofs and the derivation of formulas. Nor is the book intended for grammar-school boys. It is assumed that the reader has a good command of college algebra. Many of the concepts can be more quickly understood by the student who has mastered the calculus, but such preliminary training is by no means essential for a satisfactory understanding of the subject matter of this elementary text. There is a feud of long standing between those statisticians who argue that no mathematical training is necessary for an understanding of actual statistical operations and those who argue that no formula can be used correctly except by one who understands how it was derived. In this dispute the author takes a middle ground. It is his belief that a thorough training in mathematics is desirable but is not a *sine qua non*. Fortunately there are several good texts available in which the qualified student can find the mathematical treatment developed *in extenso*. In the present text the author has included derivations and mathematical proofs only where they seemed necessary for a workable understanding of the elementary concepts. In many cases it has naturally been difficult to decide what to include and what to omit, and in cases of doubt the author has leaned toward omission.

No attempt has been made here to describe the development of any of the concepts treated. The names of great statisticians are notable by their absence. The author has been forced by lack of space to omit references to the men who have developed the science and techniques of statistics, just as the ordinary algebra text fails to tell who first demonstrated each rule and as

[1] The author's "Laboratory Manual and Problems for Elementary Statistics" McGraw-Hill Book Company, Inc., is specially planned to accompany this text, but any other similar compilation of problems can be used with proper planning.

the ordinary text in elementary economics fails to state in connection with each principle who originally formulated it.

This book does not present original statistical theory. Elementary texts seldom do (and almost never should) cover the "growing points" of the science that are still under dispute. As in any beginner's text, it has been necessary to omit many subjects altogether, but where any concept has been treated the aim has been to have the treatment accurate and up to date. In many cases where the limitations of space and of students' time made it impossible to cover a subject in detail (in advanced correlation analysis, for instance) a special effort has been made to cover the basic concepts in such a way that the student will find it easy to assimilate additional material himself and visualize the problems involved.

In the arrangement and approach to the subjects that have been covered, the author has been guided by his experience of over a decade in teaching courses in elementary statistics to students specializing in many fields. The aim has been to keep the presentation simple and readable, and it should be possible for any serious student to cover the subject here even without the aid of a teacher. The author owes much to his students—perhaps most of all to the more obtuse of them—who have forced him to tax his ingenuity in searching out new methods of attack on, and new illustrations of, old concepts.

References are not given at the end of each chapter, but the interested student will find a list of the more commonly available statistics texts in Appendix VII. No attempt has been made to make this bibliography complete; the author has included those books which he has found most helpful with his own students.

It is difficult to give proper credit for much of the material in this book. In many cases reference has been made, either in the body of the text or in footnotes, to specific sources from which I have drawn. A complete catalogue of such sources would, however, require more space for footnotes than has been given to text. For several years my classes have used the texts of Professors R. E. Chaddock and F. C. Mills. I have doubtless absorbed ideas of theirs until I do not know which ideas are my own and which originated with them. I have been greatly interested in and influenced by Ezekiel's splendid work on

correlation methods. Among more recent works I have found Professor Richardson's mathematical introduction especially stimulating. The dynamic and lucid exposition of statistical philosophy of my friend, Professor Henry Schultz, has colored my general attitude toward statistical problems, although the subject matter of this text is of such an elementary nature that I have had small occasion to draw upon it. Professor Frederick E. Croxton has kindly read almost the entire manuscript and has made suggestions and pointed out such troublesome errors as creep inevitably into a work of this kind. The book has been improved by his painstaking attention and frank suggestions of alternative methods of treatment.

It is too much to expect, I fear, that this book is free from errors. Teachers using it will be doing me a kindness if they will suggest improvements in the presentation and point out errors that may occur. Since I have read the proofs of the book myself, all errors must be laid at my door. But for the kind assistance of friends those errors would have been more numerous.

ALBERT E. WAUGH.

CONNECTICUT STATE COLLEGE,
STORRS, CONN.,
December, 1937

CONTENTS

CHAPTER IV

CHAPTER V

CHAPTER VI

CHAPTER VII

CHAPTER VIII

CHAPTER IX

CHAPTER X

CHAPTER XI

CHAPTER XII

CHAPTER XIII

CHAPTER XIV

CHAPTER XV

CHAPTER XVI

CHAPTER XVII

APPENDICES

ELEMENTS OF STATISTICAL METHOD

CHAPTER I

THE NATURE OF STATISTICS

1.1. Scientific Method.—Men have discovered facts in many ways. Some things have been learned by chance. Wisdom, so it is said, has been imparted to men in dreams and through miraculous revelation. There are those who assert that they are gifted with clairvoyance, by means of which they are able to discern things hidden from the ordinary mortal. And, of course, a large portion of all human knowledge has come down to us from unknown sources. The method by which it was originally discovered is not known.

The tremendous advances in human knowledge that have characterized the past century and a half have not come, in the main, from any of the sources above mentioned. Nowadays a great preponderance of the additions to our information are the result of plan and not the product of chance. We learn new things because we have gone about the learning process methodically, and not haphazardly. The methods which are used in acquiring knowledge are called *scientific methods*, and it should be remembered that they do involve the following of a definite plan.

1.2. Experimental Method.—The best known of the scientific methods, and the one which has been most fruitful, is called the *experimental method*. Galileo, we are told, was attracted by the swinging of lamps in the cathedral. He noted that the period of the swing varied, and he wished to determine what factors influenced the length of the period. He did not rely on dreams, nor did he, so far as we know, patronize the local fortuneteller.

He began to experiment. He went at the problem methodically, in an attempt to determine what forces were at work in the pendulum.

Now Galileo might have taken the first half-dozen pendulums that he encountered and studied them. If he had done so he would have found that the pendulums differed in many respects. In some the bob would be heavier than in others. In some the length would be greater than in others. In some cases there might be air currents which were not present in others. At the points where the pendulums were located there might be differences in air pressure, relative humidity, the attraction of gravity, etc. And under such circumstances when Galileo found that pendulum 1 oscillated more rapidly than pendulum 2 he would be unable to tell whether the difference was due to differences in length, weight, humidity, air pressure, or some combination of these forces. Hence Galileo was very careful to construct pendulums that differed in but one respect; that is, he would make several pendulums of the same length; he would protect them all from currents of air that might affect the rate of swinging; he would operate them all at the same time and place so that there would be no differences in barometric pressure and the like. In fact, these pendulums would be exactly the same and would be operated under identical conditions, except, let us say, that there would be variations in the weight of the bob. Under these circumstances if Galileo found that there were regular variations in the period of oscillation which corresponded to differences in the weight of the bob, he would know that the cause must be bob weight, since no other factors differed. If, on the other hand, these pendulums did not differ in period, he would know that bob weight did not affect the rate of swinging.

Having discovered the effect, if any, which was associated with the weight of the bob, Galileo would now construct pendulums which differed in nothing except length, and would ascertain the relationship of length to period. He would then, in turn, discover the effect of gravitational attraction, barometric pressure, etc. The important thing to note is that the method consists in keeping all forces save one constant, and in varying that force in order that the scientist may discover its effect, if any. This method of investigation is called the experimental

method, and where it is applicable scientists prefer it to all other methods.

1.3. Statistical Method.—But often men wish to discover facts in fields where the experimental method cannot be applied. Let us suppose, for example, that you wish to discover the forces that determine the price of milk in New York City. You would like to apply the experimental method. This means that you would have to try one thing at a time, keeping all other things constant, and note the effect of the changes. First you would make changes in the quantity of milk offered on the market, to determine whether or not the price varied as you varied the amount offered. But it would be necessary for you to see that there were no changes in the other factors. You would have to establish entire uniformity in people's wages, since presumably the amount that they will pay for milk depends partly on the amount of their incomes; you would have to make sure that the tastes of consumers remained constant, since changes in their desires would perhaps change the amount that they would pay; you would find that you were forced to fix the general price level so that there would be no changes from variations in the purchasing power of money; etc. But this is manifestly out of the question. Here the experimental method cannot be applied, because the many factors cannot be held constant as the scientist varies the one force in which he is interested for the moment. Thus in the social and biological sciences it is often impossible to make use of the experimental method.

We should be foolish, however, to neglect entirely those fields in which experimental method is out of the question. To be sure, it is very difficult to discover facts in such fields, but it is none the less desirable that facts be discovered, and the scientist must do what he can in the face of difficulties. He could, of course, fall back on chance or revelation. In such a case he would have left the field of science entirely, since he would be following no plan. As a matter of fact he is likely to fall back on another method, as a poor second choice. This other method (or body of methods) we call *statistical method.* When we apply statistical method to a problem we go at the problem systematically, as in the experimental method; but the system used is not the same. Being unable to hold forces constant, we perforce let them vary. But now we record the

variations in all the forces operating and attempt to determine the separate part which each plays in influencing the result. Under ordinary circumstances this method is much more difficult than the experimental method, and the results obtained are usually less accurate and less satisfying. But they are much better than no results at all.

The classification of scientific methods into those which are experimental and those which are statistical, like most classifications, is formal and arbitrary and not entirely realistic. When the scientist comes to work on a problem in practice, he almost always combines elements of both the statistical and the experimental approach. Many of the most important of the statistical methods were originated in the fields of physics and astronomy—fields that we usually think of as "exact" sciences. Even in these fields the scientist has to contend with errors of observation, and in addition he usually finds it impossible to record the values of all the variables which are involved.[1] Under such circumstances the "exact" scientist is forced to combine statistical methods with his experimental procedures. On the other hand, even the social scientist can and does use a certain amount of control in his investigations.

1.4. Statistics.—The word "statistics" is used in two senses which differ materially. It is sometimes used in the singular and sometimes in the plural. When used in the plural, it refers to numerical data. Thus if we say that there *are* statistics in the "World Almanac" we mean that there are numerical data there. When we use the word in the singular, we refer to a body of methods which are used in summarizing such numerical data. Statistics *is* a body of methods which are used when we wish to study masses of numerical data and to extract from them a few simple facts.

Originally statistics were gathered for public purposes. In fact, the word "statistics" and the word "state" come from the same root. Statistics were gathered for purposes of taxation and for military purposes. But nowadays there is almost no field in which statistics are not useful. Every science depends to some extent on the gathering of data and on the application to them of statistical methods. In some fields, as has been pointed out above, the statistical methods are almost the only

[1] See Chap. II for a discussion of these problems.

methods that can be used, while in other fields they are a minor supplement to other scientific methods.

1.5. Preliminary Admonitions.—It is important to remember that the purpose of statistical method is to simplify great bodies of numerical data. If you were shown a table containing 1000 figures, each figure representing the weight of a newborn baby, you would be confused by the very mass of material itself. But if these 1000 figures could be boiled down to one or two, you would comprehend them quickly. Thus if we discover that the average weight of girl babies at birth is 7.1 lb. and of boy babies 7.6 lb., we have derived two figures from the original 1000 and have made much simpler the problem of understanding the original data. To be sure, we must not be misled into believing that the original data were so simple as our conclusions. We must not come to the conclusion that each girl baby weighed 7.1 lb. and that each boy baby was just ½ lb. heavier. We have given up some of the detail of our original figures in order that we may get a simple, convenient, and easily understood general statement.

It is the purpose of statistical methods thus to simplify data. In too many cases students who have studied but a little statistics lose sight of this fact and come to believe that the purpose of statistics is to mystify the uninitiated. They try, by the use of uncommon terms and symbols, to impress the layman with their own erudition. Such an attitude shows complete lack of comprehension on the part of the student. Unless data are simpler and easier to understand after statistical methods have been applied than they were before, the time and trouble of applying the methods have been wasted. If statistical methods make data more complicated and harder to understand they are worse than useless. The student should try, in the case of each method discussed in this book, to see just how it makes masses of data simpler and more easily understood than they would otherwise be.

It is also wise to caution the student who is beginning the study of statistics that statistical methods cannot, in themselves, solve any problem. The original data must have been accurate, the methods must have been properly applied, and the results must have been interpreted by one who understands both the methods themselves and the field to which they have been

applied. Many a student feels that if he takes data—any data—and performs certain mystic necromancies, he will get results which, by some unknown power, are correct. He has them down on paper, in black and white, carried to seven decimal places, and hence they *must* be right. He feels, with Mephisto's pupil, *Denn was man schwarz auf weiss besitzt, kann man getrost nach Hause tragen.* It is important to realize the fact that no statistical method can, in itself, insure against mistakes, inaccuracy, or faulty reasoning and incorrect conclusion. These methods are to be thought of as tools which, when in proper hands and when applied to the materials for which they are designed, can turn out useful products, but which have no powers to work wonders by themselves.

1.6. Suggestions for Further Reading.—Any textbook on logic will describe some of the characteristics of the various scientific methods. In addition, there are a number of interesting and instructive books on the subject of the scientific method itself. Karl Pearson's "Grammar of Science," now available in the handy Everyman's Library, is one of the best. For several interesting illustrations of pure chance discoveries, see Professor W. B. Cannon's article, The Role of Chance in Discovery, *Scientific Monthly*, Vol. 50, No. 3, March, 1940, pp. 204–209.

CHAPTER II

THE MEANING OF NUMBERS

It is impossible for us here to go into the philosophy of number theory, nor is it our intent to develop the history of the number concept. Before we work with numbers it is important, however, to understand just what we do and do not mean when we express facts in numerical form.

2.1. Accuracy of Measurement.—Most of the numbers that we use in scientific work represent measurements. These measurements are made with various kinds of instruments, varying from such relatively simple things as a foot rule to complicated apparatus such as that used to measure the speed of light. Yet no measuring instrument is completely accurate, nor is the operator of any measuring device completely dependable. Two men will read an instrument in slightly different ways, or the same man will read the same instrument in different ways at different times. The accuracy of a measurement will depend in part on the skill and the carefulness of the person making it. It will also depend in part on the instrument used. Some scales will give readings to the nearest pound, some to the nearest ounce, some to the nearest milligram. Yet even the finest, most delicate, most precise of scales has somewhere a limit of accuracy beyond which it cannot go. Similarly an instrument for measuring lengths may be as crude and inaccurate as the hand lead line used by mariners to ascertain the depth of water, which has markings at 2, 3, 5, 7, 10, 13, 15, 17, and 20 fathoms. Since a fathom is 6 ft., one might be able to discover with such a line that the depth of the water was between 60 and 78 ft. (10 and 13 fathoms), but for greater accuracy he would have to rely on his ability to estimate rather than on the instrument used. On the other hand, in manufacturing many kinds of machinery the tolerances are well below one hundredth of an inch, and various ingenious sorts of instruments have been developed which will distinguish and record lengths far smaller than this. Yet again, even the best

cross hair in the instrument, some observers tend always to record the transit just before it really takes place, some just after it really takes place, and some at approximately the right time. Some individuals thus have a biassed error in one direction and some in another, while others seem to have compensating errors. A popular magazine reported a few years ago[1] that when a large number of people were asked to estimate the length of a minute by ringing a gong at the beginning and end of their estimates their average guess was only 35 sec. If all or most of these people tended to err in the same direction, the error was a biassed one, as it very evidently was in this case. Bias may come from the unwillingness or the inability of people to give correct information, or from peculiar individual traits which lead an observer to read a scale incorrectly but always too high or too low. Biassed error can sometimes be discovered and eliminated. Random error can never be eliminated, although its effects may be reduced by getting large numbers of observations.

2.3. Significant Figures.—In abstract arithmetical work when one uses the number 15 he means just exactly 15—no more and no less. The number 15 and the number 15.0000 are assumed to mean the same thing. But we have just learned that in scientific work a number is seldom exact. When a scientist uses the number 15 he means "approximately 15." The convention which has been adopted in all the various fields of scientific work is that the scientist will write down as many digits as he knows, and then add zeros enough to locate the decimal point. Usually the last digit other than zero is an approximation which is correct to the nearest place. When an elementary physics book gives the speed of light as 186,000 miles per second, it is understood that the digits 1, 8, and 6 represent measurements, although the last digit, 6, is probably an approximation. The three zeros are not measurements at all. They are put there merely so that we shall know the position of the decimal point. In fact, Newcomb and Michelson's[2] determination of the speed of light is 186,324 miles per second. It is at once evident that the three zeros in the number 186,000 were not measured zeros. They merely indicated that the measurement was in thousands of miles

[1] *Collier's Magazine*, Vol. 108, No. 4, July 26, 1941, p. 8.

[2] "American Ephemeris and Nautical Almanac," 1940, p. xx.

rather than in miles. And even the more accurate figure 186,324 is probably not exactly correct. It means that the speed lies nearer to 186,324 than to 186,323 or 186,325. Thus we know only that the speed lies between 186,323.5 and 186,324.5 miles per second. The statement that the speed of light is 186,000 miles per second is taken by the scientist to mean, not that it is exactly 186,000 miles, but nearer to 186,000 than to 185,000 or 187,000. It is correct to the nearest thousand miles. It lies between 185,500 and 186,500 miles.

As we change the form of our statement, increasing its accuracy, we could give the speed of light successively as 190,000; 186,000; 186,300; 186,320; and 186,324 miles per second. Each time we get a little closer to the fact, but we never get the exact measurement save by chance. As we get more and more accurate in our statement, coming progressively closer to the exact figure without ever getting there, we say that we get more and more *significant figures.*

Webster's Dictionary defines significant figures as "figures that remain to a number or decimal after the ciphers at the right or left are canceled." Thus the number 1000 has one significant figure; the number 900 has one significant figure; the numbers 910 and 912 have, respectively, two and three significant figures. The following numbers have the number of significant figures indicated, in conformity with the rule given in Webster's definition:

Number	Significant Figures
200	1
20	1
2	1
0.2	1
0.02	1
0.002	1
210	2
217.352	6

Webster's definition, however, falls down in some cases. The number 321.4500 would have five significant figures if we canceled the zeros at the extreme right. But these zeros were not necessary to locate the decimal point. To be more accurate we could say:

1. Every digit except zeros is always significant.
2. Zeros are always significant unless
   a. They are at the extreme right of a number and to the left of the decimal point.
   b. They are at the extreme left of a number.

Thus, in the number 32,056, the zero is not at the extreme right or the extreme left, and it is therefore significant. The number has five significant figures. In the number 230.00, the zeros are at the extreme right, but they are not at the left of the decimal point. Hence they are significant, and the number has five significant figures. In the number 186,000, the zeros are at the extreme right *and* at the left of the decimal point, so they are not significant, and the number has three significant figures. In the number 0.003, the zeros are at the extreme left and are not significant. The number has one significant figure.

We can state the rule in another form by saying that all digits are significant figures except zeros which had to be included to show the location of the decimal point. In the examples in the preceding paragraph, it is clear that the zero in the number 32,056 was a measured zero, and was not put in to locate the decimal point.

Zeros at the extreme left of a number are always insignificant, since there is no other reason for using them than to locate the decimal point. But with the number 230.00, the final zeros did not have to be put in to locate the decimal point. They should not be put in at all unless they represent measurements. The number 230 and the number 230.00 do not mean the same thing in scientific work, even though in pure mathematics they are the same. In science the number 230 means that a measurement lies between 225 and 235; while the number 230.00 means that the measurement lies between 229.995 and 230.005. It is obvious that the figure 230.00 represents a far more accurate measurement than the figure 230. The convention in scientific circles is to write only as many digits as are known to be correct, adding enough zeros to locate the position of the decimal point if its position would not otherwise be evident. Digits so written (not including the zeros added merely to locate the decimal point) are called significant figures.

The student sometimes feels that the number 0.000324 must represent a very accurate measurement on account of the zeros which precede it. We have said that these zeros are not signifi-

cant. Suppose you have measured a distance and found it to be, as nearly as you can tell, 324 mm. You can express this measurement also by saying that it is 0.324 m. or that it is 0.000324 km. You have not increased the accuracy of the measurement by using larger units (kilometers instead of meters). You have expressed exactly the same measurement in three different ways, and each of the three numbers 324; 0.324; and 0.000324 has the same number of significant figures, namely, three.

2.4. Standard Notation.—Sometimes, to be sure, we may have zeros at the right-hand end of a number when they really are significant. Suppose I measure a distance and find that it is 100 ft. to the nearest foot. I know that it lies between 99.5 and 100.5 ft. Under these circumstances the two zeros really represent measurements, yet under our rules they would be called nonsignificant. Or, of course, one of the zeros might be significant while the other was not, if I had found the distance to be between 95 and 105 ft. to the nearest 10 ft. Thus we see that the number 100 may have one, two, or three significant figures, depending on the actual accuracy of the original measurement.

Often we can tell whether such zeros are significant or not by the context. Suppose, for example, that I am given a column of figures on cotton production for various years, and I find these figures:

Year	Production
1938	365,000
1939	371,000
1940	390,000
1941	396,000
1942	407,000
1943	417,000

I note that one of the numbers, that for 1940, appears to have but two significant figures according to our rules, yet it is fairly safe to assume that it, like the other numbers, was given to the nearest thousand and that one of its zeros is significant.

When one wishes to show which zeros are significant and which are not, it can be done easily by means of what is called *standard notation*. Our system of enumeration is based on the radix 10, and every number in our system can be stated in the form of some number multiplied by a power of 10. For example, the number 20 is 2×10; the number 200 is 2×100, or 2×10^2; the number 234 is 2.34×100, or 2.34×10^2. Here are several other numbers, each stated in the usual form and also in standard notation:

Usual Form	Standard Notation
15,325	1.5325×10^4
21	2.1×10^1, or 2.1×10
2.1	2.1×10^0, or 2.1×1
0.21	2.1×10^{-1}
0.021	2.1×10^{-2}
0.0021	2.1×10^{-3}

The student will see at once that when the exponent of 10 is positive it means, "move the decimal point so many places to the right." Thus 1.5326×10^3 means 1.5326 with the decimal point moved three places to the right, or 1532.6 When the exponent of 10 is negative, it means "move the decimal point so many places to the left." Thus 2.345×10^{-4} means 2.345 with the decimal point moved four places to the left, or 0.0002345.

Now let us go back to the problem that we raised at the beginning of this section. How can we indicate that one of the zeros is significant and one is not in the number 100? First we write it in standard form, when it could appear in any of the following ways:

$$1 \times 10^2$$
$$1.0 \times 10^2$$
$$1.00 \times 10^2$$

When one thinks of the numbers as pure numbers, these seem to be the same. But the first has one significant figure, the second has two, and the third has three. We note that the second number includes the expression 1.0. In this number the zero is significant under the rules as given on pages 11–12. In the number 1.00 both zeros are significant under our rules. Therefore we see that when we write 1×10^2 we mean that only the figure 1 is significant; if we write 1.0×10^2 we mean that the 1 and one of the zeros are significant; and if we write 1.00×10^2 we mean that all three digits are significant. A number is in standard form when we have written the first digit followed by a decimal point, and then such other digits as are significant, with the entire number multiplied by whatever power of 10 is necessary to put the decimal point in the proper place.

2.5. Computations with Approximate Numbers.—The rules that we all learn for the simple arithmetical procedures of addition, subtraction, multiplication, and division are based on the

assumption that we are using "pure numbers"—numbers that are exactly accurate and mean exactly what they say. But when we use numbers that are merely approximations, as we almost always do in science, these rules are likely to give us misleading conclusions. For example, suppose you are to find the average of the numbers 7, 4, and 6. By common arithmetical methods you would add the three numbers and divide by 3 to get

$$17/3 = 5.666666666666 \ldots,$$

carrying out the computation to as many 6's as your patience would permit. But when we remember that these numbers are not accurate, we realize that the number 7 means "somewhere between 6.5 and 7.5," the number 4 means "somewhere between 3.5 and 4.5," and the number 6 means "somewhere between 5.5 and 6.5." The sum of our three numbers could be as low as $6.5 + 3.5 + 5.5 = 15.5$; or it might be as large as

$$7.5 + 4.5 + 6.5 = 18.5.$$

In the first case the average would be $15.5/3 = 5.17$ approximately. In the latter case the average would be $18.5/3 = 6.17$ approximately. What is the sense in carrying out the answer as 5.666666666 . . . when we are not sure of even the first digit? The long line of 6's does not represent actual measurement, and the numbers should not be there. They are not significant. They are not digits which are really known at all. We see at once that when our original figures are more or less inaccurate there is no reason to carry out computations to large numbers of decimal places which contain only a pretended accuracy. Therefore we shall need to use new rules for the simple arithmetical processes to adapt them for use with approximate numbers.

2.6. Multiplication and Division of Approximate Numbers.—When multiplying or dividing two or more approximate numbers the following rules should be used:

1. Round off the numbers with the largest number of significant figures until they have but one more significant figure than does that one of the numbers which has the smallest number of significant figures.
2. Multiply or divide the rounded numbers in the usual way.
3. Round off the answer (product or quotient) until it has no more significant figures than has that one of the original figures which contained the smallest number of significant figures.

These rules can best be understood by illustrations.

What is the area of a table top that measures 72×36 in.? Both numbers have two significant figures. No preliminary rounding off is necessary. We multiply $72 \times 36 = 2592$. We round off the answer to two significant figures to get 2600 sq. in., which we give as the answer. The answer 2592 sq. in. should not be given, since it contains a pretended accuracy. When the original measurements are given as 72 in. and 36 in., the area may be anywhere between 2646.25 sq. in. (as it would be if both original figures had the greatest possible values) and 2538.25 sq. in. (as it would be if both original figures had the smallest possible value). To pretend that we know the area to the nearest square inch when we do not even know it for certain within 100 sq. in. is likely to mislead ourselves as well as others. The student will notice that even our rules give slightly more significant figures in the answer than we are sure of, since they give an answer of 2600 sq. in. while we are not even sure of the second digit. If the rules for computation given here seem to the student to be rough, approximate, and inaccurate, he may rest assured that these rules do not discard any accuracy which really existed, but only imaginary accuracy.

One more example. We measure the circumference of a large cylindrical water tower, using a foot rule. Because of the crudeness of our apparatus we are not sure of the result to thousandths of inches, but in our best judgment the circumference is 530 in. the first two numbers only being significant. What is the diameter of the tower? We look in our textbooks and find that the diameter can be found by dividing the circumference by 3.14159265358979323846 . . . , or we consult our memories and recall that we can divide by 3.1416; 3.14; or $3\frac{1}{7}$. Since our measurement of the circumference was inaccurate (as far as we can be sure) at the start, it would be foolish to use extremely accurate values of π in our computation. Our first rule says to round off the number with the most significant figures. Our number with the least significant figures, 530, contains two; so we round off the value of π to three significant figures—one more than we have in the least accurate figure. This gives us a value of 3.14 for π. We now divide to find

$$\frac{530}{3.14} = 168 \text{ in.}$$

We know at once that we do not need to carry the computation further, because we want the result to two significant figures only. We carry it one more place to discover whether it is nearer 160 or 170, but we now round off the result and give our final answer as 170 in. for the diameter. If we want to know the diameter with more accuracy, it will not help us to use a more accurate value of π or to carry our result to more decimal places. The only way we can learn the diameter more accurately is to measure the circumference more accurately at the start.

2.7. Addition and Subtraction of Approximate Numbers.—The rules for addition and subtraction of approximate numbers are based, not on the significant figures, but on the decade, or number of the column counting from the decimal point. Suppose we are told that the distance from New York to Chicago is 910 miles and 10 miles the other side of Chicago we come to a fork in the road. We are to take the right fork and proceed 150 ft., where we find a house. We walk up the front walk to the door, a distance of 38 ft. There is a table 7 ft. 2 in. from the front door. On the table, 0.32 in. from the edge, is a box. How far is the box from New York City? We see that it would be foolish to add together 910 miles, 150 ft., 38 ft., 86 in., and 0.32 in., even though each of the measurements contains two significant figures. If we know the distance from New York to Chicago only to the nearest 10 miles, we cannot start adding inches and make sense. Therefore our rules for addition and subtraction of approximate numbers are

1. Arrange the figures in a column, all in the same units (all in feet or miles or inches, for example) with the decimal points over one another.
2. Find the column containing one or more nonsignificant figures which is farthest to the left.
3. Round all the other figures off so that their last significant figure is in this column.
4. Add or subtract in the usual way, using the rounded figures.
5. Round off the answer (sum or difference) so that it has its last significant figure as far to the left as that one of the original numbers whose last significant figure was farthest to the left.

This sounds a good deal more complicated than it is. Again we illustrate.

It is estimated that the number of immigrants coming to the United States between the close of the Revolutionary War and 1820 was 250,000. From 1820 to 1900 the number coming was

19,123,606, and from 1901 to 1940 the number was 19,166,837. What was the total number of immigrants from the Revolutionary War to 1940? Familiar methods would lead one to add the three numbers and find a total of 38,540,443. A moment's thought will show us, however, that we are not justified in doing this. Our first number is an estimate. Possibly the number of immigrants in this first period was really 250,001, or maybe it was 251,395. The four zeros in the number 250,000 are not supposed to represent exact measurement. They represent some unknown numbers, which would take their places if we knew the actual facts. So if we set up our numbers for addition in the usual manner, we might substitute question marks for these zeros to show that we do not really know the values in those columns. This would give us the following problem in addition:

$$\begin{array}{r} 25?,??? \\ 19,123,606 \\ \underline{19,166,837} \end{array}$$

This is like being told to add 6 and 7 to some unknown number. The answer will also be unknown. Therefore we follow the rules just given. We round off the two latter numbers until they have their last significant digit in the "thousands column," since the nonsignificant digit farthest to the left (the first zero in 250,000) is in this column. This gives us our problem in this form:

$$\begin{array}{r} 250,000 \\ 19,124,000 \\ \underline{19,167,000} \\ 38,541,000 \end{array}$$

This gives us an answer of 38,541,000 instead of an answer of 38,540,443. The latter had an unwarranted and pretended accuracy in its final digits. Even our rule carries us one column further than we are sure of, since in the thousands column we added 7 and 4 to an unknown number. Hence we now round off the answer until its last significant figure is in the "10-thousand column," since in our original numbers we had one case where the last significant figure was in this column. This gives us a final answer of 38,540,000 immigrants. If we state our rule loosely (and the meaning should now be understandable), we can say that we first round off to one more column than is really

known, and then add or subtract. We then round off our answer to the last column that is really known.

It is quite possible to lose many, or even most, of our significant figures in the process of subtraction. For example, let us ask how many more immigrants came to the United States from 1901 to 1940 than came from 1820 to 1900. The original figures are given in the preceding paragraph, and we subtract as follows:

$$\begin{array}{r} 19{,}166{,}837 \\ -19{,}123{,}606 \\ \hline 43{,}231 \end{array}$$

We started with two numbers, each containing eight significant figures. Our difference contained but five significant figures.

2.8. A Horrible Example.—Suppose you are asked to find, as accurately as you can, the weight of the earth. You look up the necessary original data and find the following:

a. The volume of a sphere is $4.1888r^3$ where r is the radius.

b. Estimates of the polar radius of the earth vary from 6,356,079 to 6,356,992 m.

c. Estimates of the equatorial radius of the earth vary from 6,377,397 to 6,378,388 m.

d. A meter is 3.28 ft.

e. The density of the earth is 5.5 times that of water.

f. Water weighs 62.5 lb. per cubic foot.

All these figures are approximations, and from them you wish to ascertain the weight of the earth. You decide that you will use as the radius the average of the four figures given in *b* and *c*, which gives you a value of 6,367,214 m. If you follow the rules of arithmetic, forgetting the rules that we have just learned, your computations will be as follows:

1. Cube of radius is 258,135,859,576,097,196,344 cu. m.
2. Volume is found, by multiplying (1) by 4.1888:

   1,081,279,488,591,955,925,045.7472 cu. m.

3. Each cubic meter contains 3.28^3 cu. ft., or

   35.287552 cu. ft.

4. Volume of earth in cubic feet is product of (2) and (3):

   38,155,706,190,222,051,486,759.9066988544 cu. ft.

5. From *f* and *g* above, a cubic foot weighs

   343.75 lb.

6. Weight of earth is product of (4) and (5):

   13,116,024,002,888,830,198,573,717.927731200000 lb.

If, on the other hand, we use the rules that have been given, we note first that the density of the earth is given to but two significant figures, as 5.5 times that of water. Therefore we round off all our other figures to three significant figures and state them in standard notation, thus:

a. Volume of a sphere is $4.19r^3$.
b. Radius of earth is 6.37×10^6 m.
c. A meter is 3.28 ft.
d. Density of earth is 5.5 times that of water.
e. Water weighs 62.5 lb. per cubic foot.

We then carry out our computations, remembering that when we multiply two numbers we add the exponents of the figure 10, and when we raise to the nth power we multiply the exponent by n. This gives us the following steps:

1. Cube of radius is 258×10^{18}, or 2.58×10^{20}
2. Volume of earth is (1) times 4.19, or

$$10.8 \times 10^{20}, \text{ or } 1.08 \times 10^{21}$$

3. A cubic meter contains 3.28^3 cu. ft., or

$$3.53 \times 10 \text{ cu. ft.}$$

4. Volume of earth is product of (2) and (3), or

$$3.81 \times 10^{22} \text{ cu. ft.}$$

5. From *d* and *e* above, a cubic foot weighs

$$3.44 \times 10^2 \text{ lb.}$$

6. Weight of the earth is product of (4) and (5):

$$13.1 \times 10^{24}, \text{ or } 1.31 \times 10^{25} \text{ lb.}$$

The answer which we get this way, namely, 1.31×10^{25} lb., is as accurate as we can possibly get with data as inaccurate as those with which we started. We have saved ourselves a great deal of arithmetic and have not misled ourselves with fictitious accuracy. The long, tedious processes of ordinary arithmetic gave an answer to the 12th decimal place, which means that we pretend to know the weight of the earth down to the last tiny grain of sand; while as a matter of fact we started with a density of the earth about which we knew only that the figure lay between 5.45 and 5.55. The correct method of computation can be carried out on a slide rule with accuracy as great as is warranted by the original figures. It actually took 4½ min. to get this answer with the slide rule. The student might try timing himself on even the single process of cubing 6,367,214, which we took in both cases to be the radius of the earth, to see how much time is saved by following the correct procedure. But it should be emphasized that *the rules given in this chapter are not intended to save time.* Timesaving is just a pleasant by-product. The rules are used to keep us from giving a purely fictitious accuracy to our results. The rules are intended to keep us from saying that we know something when we really merely conjecture it. A

very large part of the data obtained with ordinary measuring instruments is inaccurate enough at the start so that we introduce no new inaccuracy if we carry out our computations with a slide rule.

2.9. Rounding Off Numbers.—In rounding off numbers the accepted practice is to leave the last digit retained just as it is if the quantity rounded off amounts to less than half of one of the units retained, but to increase by one unit the last digit retained if the quantity rounded off exceeds half of one of the units retained. If the quantity rounded off amounts to exactly one-half a unit, the convention is to leave unchanged the last digit retained if it is even, but to increase it by unity if it is odd. This means that the last unit of the rounded number is left even if it was even or is made even if it was odd—both times, of course, if the quantity rounded off is exactly half a unit. The purpose of these rules is to increase the number kept half the time and leave it unaltered half the time, on the theory that such a procedure will tend to balance the positive and the negative errors on the average over any large number of cases. We can illustrate the rules as follows:

Round off each of the following numbers to three significant figures:

236,941 is rounded off to 237,000 since the part rounded off (941) is greater than one-half (500). The last unit kept is in the thousands, and 941 exceeds half a thousand.

236,241 is rounded off to 236,000 since the part rounded off (241) is less than half a thousand, and the last number kept is in the thousands.

2,365,001 is rounded off to 2,370,000 since the last significant figure is in the 10 thousands and the part rounded off (5,001) exceeds half of 10,000.

236,500 is rounded off to 236,000. The part rounded off (500) is just half a unit, so we leave it even.

235,500 is also rounded off to 236,000. The part rounded off is exactly half a unit, so we make the uneven number even by increasing it one unit.

209,700 would be rounded off to 3.00×10^5. If we do not use standard notation we get 300,000 when we round it off, yet obviously some of the zeros are significant. We were rounding to three significant figures, so we gave the 3 and the two zeros, writing in standard form.

These rules are not followed invariably in statistical computation, especially when the use of computing machines makes it as easy to carry a result to 10 as to 2 places. Also, for reasons that will become evident later, it is sometimes necessary to carry out intermediate computations to several decimal places in order to ensure the desired accuracy of final conclusions. The student

should always be on his guard, however, against "accuracy" that is purely fictitious, and the rules given above should be kept constantly in mind as guides.[1]

2.10. Suggestions for Further Reading.—The student will probably learn more about this subject by working out many examples than he will by further reading. C. H. Richardson, "An Introduction to Statistical Analysis," Harcourt, Brace and Company, New York, 1935, gives a good and simple treatment in his introductory chapter. A brief but faulty statement appears in the 13th edition of the Encyclopaedia Britannica (also in the 11th and 12th editions) in the article on arithmetic, section VII on approximation, subsection 82 on degree of accuracy. It would be good practice for the student to discover for himself the error in the definition of significant figures there given. The last section of Chap. V in E. F. Lindquist's "A First Course in Statistics: Their Use and Interpretation in Education and Psychology," Houghton Mifflin Company, Boston, 1938, gives a good brief discussion. David Brunt, "The Combination of Observations," Chap. I, Cambridge University Press, London, 1931, discusses accidental and constant errors, with comments on errors caused by measuring instruments and those caused by the observer. He also points out (with proof) that the arithmetic mean of a number of observations is more accurate than the observations themselves. In general, perhaps the best treatment of the subject is found in books in the field of physics or astronomy which deal with the problem of measurement. See, for example, William Chauvenet, "A Manual of Spherical and Practical Astronomy," Vol. II, Appendix, J. B. Lippincott Company, Philadelphia; or Dascom Greene, "An Introduction to Spherical and Practical Astronomy," Appendix, Ginn and Company, Boston. A splendid treatment appears in Willford I. King's "The Elements of Statistical Method," Chap. VIII, The Macmillan Company, New York, 1922.

[1] It can be shown with little difficulty that one can expect, on the average, somewhat more accuracy in the arithmetic average of a set of numbers than there is in the original numbers themselves. This can be shown either empirically by finding the average of a set of numbers, rounding off some of the known digits, and comparing the average of the rounded numbers with the average of the original numbers, or on a priori grounds. We should, on the average, be able to carry one more significant figure in our average than there is in the original figures if we take the average of 10 numbers, two more significant figures if we take the average of 100 numbers, three more significant figures if we take the average of 1000 numbers, etc. If we have the weights of 100 people, each weight to the nearest pound, we can theoretically carry the average weight to the nearest hundredth of a pound. In practice, if we are dealing with large numbers of cases, it is probably safe to carry the average to one more significant figure than the original figures. For an explanation of this fact, see Raymond Pearl, "Medical Biometry and Statistics," 2d ed., pp. 362*ff.*, W. B. Saunders Company, Philadelphia, 1930.

EXERCISES

1. How many significant figures are there in each of the following numbers: 2200; 134.6; 0.00054; 19,000.00; 1.300×10^3?

2. Multiply 345.982 by 13.6, assuming that both are approximate numbers.

3. The value of π has been computed to several hundred decimal places. How would you decide, in the case of any actual problem, how many places to use?

4. What answer would you give to the critic who says that rounding off of original figures and of answers makes conclusions inaccurate?

5. Round off each of the following numbers until it has two significant figures: 3456.7; 0.0009460; 1821; 1871; 18,501; 19,500; 18,500; 19,999.

6. A distance has been measured as 540,000 ft. correct to the nearest 10 ft.; that is, it is known that the true distance is between 540,005 and 539,995 ft. Write the measurement in such a way as to make it clear just how accurate the number is.

7. When dealing with pure numbers you have been taught the "table of nines" as $9 \times 1 = 9$, $9 \times 2 = 18$, $9 \times 3 = 27$, etc. Write the "table of nines" as it would appear if the numbers involved were approximate measurements.

8. Suppose a man is asked how many people attended a boxing bout last night. He reasons as follows: "The parking lot must be about 2½ acres in size. It was about three-quarters full. I suppose you can get about 500 or 600 cars to the acre—call it 500. And suppose there were two people to the car. That would make 1875 people in attendance." Comment on his answer. Assuming that his original figures are reasonable, what would you give as the answer?

9. I own a rectangular building lot. The frontage on the street has been surveyed and found to be 97.53 ft. The depth of the lot has not been accurately measured, but I paced it as 30 paces. My pace is approximately 5 ft. What is the area of the lot?

10. In an old graveyard we find a tombstone "Sacred to the Memory of Garland Waggoner, died August 14, 1731, aged 89 years." In another plot we find another stone "Sacred to the Memory of Howard D. Newton, who departed this life August 14, 1731, aged 74 years, 8 months, and 7 days." How much older was Waggoner than Newton?

CHAPTER III

THE FREQUENCY DISTRIBUTION

3.1. The Frequency Table.—The statistician usually works with large numbers of data. Originally, of course, these data are in the form of individual measurements. For example, the following figures are the marks received by 90 students on an examination in elementary economics, the highest possible mark being 208:

104	57	85	203	128
121	81	105	107	100
166	109	138	75	114
75	118	109	101	81
65	143	102	107	157
149	94	165	151	181
49	158	95	206	55
81	191	142	85	82
114	79	81	136	133
122	76	103	158	43
159	150	88	176	133
153	89	89	156	112
136	92	106	112	90
119	156	82	84	163
147	179	123	104	85
131	73	107	164	158
168	93	154	102	112
69	139	142	113	147

Even here, where we have but 90 figures, the impression received by inspecting the data is not sharp and clear-cut. Moreover, this method of listing the data takes much room. Hence statisticians usually condense results of this kind into more usable form. For example, they might make a table showing the number of times each mark occurred. This would appear like Table 3.1.

Here we have the advantage that the figures have been arranged in order of magnitude, but we still have too many

TABLE 3.1.—MARKS RECEIVED BY 90 STUDENTS ON AN EXAMINATION IN ELEMENTARY ECONOMICS. HIGHEST POSSIBLE MARK, 208

Mark	Number	Mark	Number	Mark	Number	Mark	Number
206	1	151	1	113	1	84	1
203	1	150	1	112	3	82	2
191	1	149	1	109	2	81	4
181	1	147	2	107	3	79	1
179	1	143	1	106	1	76	1
176	1	142	2	105	1	75	2
168	1	139	1	104	2	73	1
166	1	138	1	103	1	69	1
165	1	136	2	102	2	65	1
164	1	133	2	101	2	57	1
163	1	131	1	95	1	55	1
159	1	128	1	94	1	49	1
158	3	123	1	93	1	43	1
157	1	122	1	92	1		
156	1	121	1	90	1	Total...	90
154	1	119	1	89	2		
153	1	118	1	88	1		
152	1	114	2	85	3		

entries for easy comprehension. These data would usually be condensed even more into the following form:

TABLE 3.2.—MARKS RECEIVED BY 90 STUDENTS ON AN EXAMINATION IN ELEMENTARY ECONOMICS. HIGHEST POSSIBLE MARK, 208

Mark	Number of Cases	Mark	Number of Cases
200–209	2	110–119	8
190–199	1	100–109	14
180–189	1	90– 99	5
170–179	2	80– 89	13
160–169	5	70– 79	5
150–159	11	60– 69	2
140–149	6	50– 59	2
130–139	7	40– 49	2
120–129	4		
		Total.....	90

Of course, it is not necessary that we group the marks in classes of 10. We might choose to group them in classes of 50, in which case we should have had the following:

TABLE 3.3.—MARKS RECEIVED BY 90 STUDENTS ON AN EXAMINATION IN ELEMENTARY ECONOMICS. HIGHEST POSSIBLE MARK, 208

Mark	Number of Cases
200–249	2
150–199	20
100–149	39
50– 99	27
0– 49	2
Total.....	90

It will be noticed immediately that as we group the data in larger and larger classes we gain simplicity but lose detail. In neither Table 3.2 nor Table 3.3 do we know a single mark that was assigned. We cannot tell whether or not anyone received a mark of 102. To be sure, we know that 39 students received marks from 100 to 149, but whether any one or all of them received the mark of 102 is not stated. We have condensed our data and made it easier to get an idea of the distribution of marks received by this class. But we have done it at the cost of exactitude.

Let us note, first, several things about the form of these tables. When data are arranged as these are, so that we are told the number of times that each of various values occurs, we say that we have a *frequency table*. Frequency tables may give each value that occurs and tell the number of times that it occurs, as does Table 3.1, page 25. More commonly they divide the data into classes and show the number of cases that fall within the limits of each class. This is the form in which Tables 3.2 and 3.3 appear.

3.2. Class Limits.—Let us turn our attention now to the numbers used to denote the classes. In Table 3.3, we find the classes described as follows:

200–249
150–199
100–149
etc.

Each class is bounded by two figures, which are called the *class limits*. The class limits of the first class listed in Table 3.3, for

example, are the numbers 200 and 249. The larger of these numbers (249) is called the *upper limit*, and the smaller (200) is the *lower limit*.

Class limits are not always exactly what they seem on casual inspection of a frequency table. Suppose, for example, that we have been testing samples of rubber bands produced in a certain factory to find out how heavy a load each band will carry before it breaks; and we find 123 samples to be distributed as in Table 3.4.

TABLE 3.4.—HYPOTHETICAL EXAMPLE OF BREAKING POINTS OF 123 RUBBER BANDS

Breaking Point (pounds)	Number of Cases
4– 6	5
7– 9	23
10–12	68
13–15	21
16–18	6

At first sight it would seem that the 68 rubber bands which are listed together all broke at weights of 10 or 12 lb. or somewhere in between. We should be likely to conclude that the class limits are 10 and 12 lb. It would be surprising that we should have bands breaking between these points, but none breaking at weights between 9 and 10 lb. If the upper limit of one class is 9 lb. and the lower limit of the next class is 10 lb., we have no place to classify weights of 9.3 lb., for example.

At this point we need to recall from the preceding chapter just what these numbers mean. We remember that the number 9 means "between 8.5 and 9.5," so when the upper limit of a class is given as the number 9, the actual limit is 9.5. The lower limit of the next class is given as the number 10, but this means "between 9.5 and 10.5," so the actual lower limit is 9.5. Thus we see that there is really not a "no-man's land" between the classes. The *stated limits* are 10 and 12, but the *actual limits* are 9.5 and 12.5.

Yet the actual limits are not always halfway between the stated limits. If we had a frequency table showing the numbers

of families with various numbers of children, the first two classes might have the following stated limits:

1–3
4–6

Here it would be incorrect to say that the upper limit of the first class was really 3.5 children. There are no families with 3.5 children. We know from the nature of the data that the upper limit and the stated limit in this case are the same. We have to decide what the stated limits mean by our knowledge of the characteristics of the data with which we are working, and it is always dangerous for a statistician, no matter how competent he may be in technical statistical theory, to work with data that he does not understand.

There is, however, still another sort of case that needs explanation. Purely for convenience of tabulation, statisticians have agreed that when the stated upper limits of all classes end with the digit 9 the upper limit is to be considered as extending clear up to the lower limit of the next class as stated. For example, we might restate our hypothetical example of the breaking points of rubber bands by transforming the class limits of Table 3.4 into the form given in Table 3.5.

TABLE 3.5.—HYPOTHETICAL EXAMPLE OF BREAKING POINTS OF 123 RUBBER BANDS

Breaking Point (pounds)	Number of Cases
4– 6.9	5
7– 9.9	23
10–12.9	68
13–15.9	21
16–18.9	6

In Table 3.5 each upper limit ends with the digit 9. Hence the upper limit of the lowest class, for example, is taken to be, not 6.9 as stated and not 6.95, halfway between the stated upper limit and the stated lower limit of the next class above, but as 6.999. . . . In other words, any value as large as 4 but not so large as 7 would be put in this class, even if it fell short of 7 by an exceedingly small amount. This method of evaluating class

limits is an exception to the general rule for interpreting the meaning of numbers—an exception that is made merely because the statistician, in looking over his original data, can save a good deal of time if he knows that every value which begins with 4, 5, or 6 goes in this class, regardless of the decimal which may follow it.

In order to make the meaning of class limits even more evident, some authors state them thus:

4 and under 7
7 and under 10
10 and under 13
etc.

These limits obviously mean exactly the same as those of Table 3.5. Other writers, in an effort to save time, write these same class intervals thus:

4–
7–
10–

In these cases the upper limits are not stated, and it is understood that the entries mean "from 4 up to but not including 7," "from 7 up to but not including 10," etc. The lower limits are stated, and the class is supposed to run up to the lower limit of the class that follows. This method of statement, however, is likely to be difficult for the novice to interpret. While it may be a useful timesaver for the statistician's own private work, it is not so good for publication as the other methods which have been mentioned.

The student should develop the habit of inspecting every frequency table that he comes across to see if he can determine the actual class limits as distinct from those stated in the table. This sort of practice will do more than anything else to show the advantages of some statements and the disadvantages of others. When class limits are properly given, there is no room for doubt as to where any particular value should be classified. As one further example, suppose we are classifying men according to their weights, and two adjacent classes have the following stated class intervals:

173–182
183–192

There is no question in this case where we should put a man who

weights 182.3 lb. The actual upper limit of the lower class is 182.5, and the item 182.3 should be included there. To be sure, one could not be certain where he should put a case of exactly 182.5 lb. Some authorities would favor dividing such a case between the classes, giving each of them one-half a case. Even easier, if our measurements have been made as accurately as to the tenth of a pound, would be to state our class limits thus:

172.5–182.4
182.5–192.4

Now it is obvious where the case reported as exactly 182.5 goes. By stating our class limits to a decimal accuracy as great as the measurements that are to be classified, we are able to eliminate all doubt on classification.

3.3. Overlapping Class Limits.—One sometimes sees tables published with the upper limit of one class coinciding with the lower limit of the next class, thus:

25–27
27–29
29–31
etc.

In such cases it is impossible to tell where to classify an item that is exactly 27 or exactly 29. It seems to belong in two classes. It is confusing to the reader, and bad practice generally, to use such overlapping class limits.

3.4. Open-end Classes.—Another bad practice often followed in making frequency tables is to set up a first class, or a last one, or both, in such a way that it is impossible to tell what the class limits are. We can illustrate this with Table 3.6, which is in many ways an example of how not to make a frequency table.

TABLE 3.6.—AGES OF HORSES ON UTAH FARMS. HYPOTHETICAL DATA

Age (years)	Number of Horses
0– 2	35
2– 5	78
5–10	220
10–20	715
Over 20..	31

In this table there are not only overlapping class intervals, and class intervals of unequal length, but also it is impossible to tell whether the 31 oldest horses were all very close to 20 years old, or whether they ran all the way up to 35, 40, or 50. Such a class is called an *open-end class,* and the inclusion of such classes in a frequency table materially reduces the value of the table to the statistician. If it seems necessary for any reason (such, for example, as those stated in Sec. 3.18) to leave an open-end class at either end of a table, it would add greatly to the value of the table if some further facts were given about the items included in the open-end class. In Table 3.6, for example, it would help materially if an asterisk were placed beside the figure 31 and a statement were then made below the table saying, "The average age of these 31 horses was 22.4 years," or words to that effect. When an open-end class is used, one should give either the total or the average of the items in the class.

3.5. Class Intervals.—The difference between the actual lower limit of any class and the actual lower limit of the next larger class is called the *class interval.* The class interval can also be defined as the distance between class marks (see Sec. 3.6). In Table 3.5 the class interval is 3 lb., since the lower limit of each class falls short by 3 lb. of the lower limit of the next larger class. Similarly the class interval of Table 3.4 is 3 lb., although the class limits are stated in different form. In Table 3.3 the class interval is 50.

There are decided advantages in setting up the classes of a frequency table in such a way that all classes have the same class interval. In Table 3.6 no two classes have the same class interval. This would make many statistical computations unnecessarily difficult, and should be avoided if reasonably possible. At a later point in this chapter (see Sec. 3.18), reference is made to some kinds of cases where it seems wise to make exceptions to the general rule and to use unequal class intervals, but unless there is some good reason to the contrary, the rule that class intervals should be equal throughout any given frequency table is a good rule to follow.

3.6. Class Marks.—For many of the statistical computations that we shall describe in the following chapters, it is necessary to know the *class mark* or the *class mid-point* of each class in a frequency table. This is the value midway between the actual

upper and lower limits of the class. In Table 3.5, for example, the class marks are 5.5, 8.5, 11.5, 14.5, and 17.5 lb. The actual limits of the smallest class are 4 and 6.99999 (approaching 7 as a limit), and the point halfway between them is found by adding them and dividing by 2. In Table 3.4 the class marks are 5, 8, 11 lb., etc. In this table the actual limits of the smallest class are 3.5 and 6.5, and the point halfway between is 5.

Sometimes, frequency tables are given with the classes defined by their mid-points rather than by the class limits. For example, Table 3.4 could be recast in the form of Table 3.7, and the two tables would be understood to mean exactly the same thing.

TABLE 3.7.—HYPOTHETICAL EXAMPLE OF BREAKING POINTS OF 123 RUBBER BANDS

Breaking Point (pounds)	Number of Cases
5	5
8	23
11	68
14	21
17	6

In this table the reader would understand that all 68 of the rubber bands in the central class did not break at exactly 11.0 lb. with no others breaking until exactly 3 lb. more had been added. He would decide that the values given in the left-hand column are class marks. If he needed the class limits he would realize that, just as the class marks are halfway between the actual limits, so the actual class limits are exactly halfway between the class marks if the latter are equally spaced.

It will be seen immediately from Table 3.7 that the class interval can be determined from the class marks just as easily as from the class limits if the class interval is constant throughout the table. Where there are open-end classes, however, or where the class intervals are unequal, the problem is not so simple. Yet for most purposes, the class marks are the important things, and we can struggle along with unequal class intervals if the class marks are known.

3.7. Cumulative Frequency Tables.—Instead of describing the numbers of rubber bands that broke within certain ranges of weight, we might equally well have listed the numbers that broke at or below given weights, or those which broke at or above given weights. If we go back to Table 3.5 we see at once that only 5 bands broke at weights below 7 lb. Twenty-eight bands broke at weights below 10 lb. (since we would have to include the 5 that broke below 7 lb. and the 23 that broke between 7 and 9.9 lb.). Likewise 96 broke at weights below 13 lb., 117 at weights below 16 lb., and 123 at weights below 19 lb. These figures can be derived directly from those of Table 3.5 (as the student should verify for himself), and we could state them in the form of Table 3.8.

TABLE 3.8.—HYPOTHETICAL EXAMPLE SHOWING NUMBERS OF RUBBER BANDS WITH BREAKING POINTS BELOW STATED AMOUNTS

Breaking Point (pounds)	Number of Bands Which Broke at Weights below Those Stated at Left
4	0
7	5
10	28
13	96
16	117
19	123

Such a table is called a *cumulative frequency table.* Another form of cumulative frequency table could be made up showing the number of rubber bands with breaking points more than the stated amounts. For example, we could transform the data of Table 3.5 or 3.8 into the form shown in Table 3.9.

Sometimes, as in Table 3.8, our cumulative frequency table lists the numbers of cases smaller than given amounts. In such cases the table starts at zero and the numbers get larger and larger until they equal the total number of items studied. Thus Table 3.8 starts at zero and rises to 123, since there were 123 rubber bands in the hypothetical example. At other times, as in Table 3.9, the cumulative-frequency table lists the numbers of items larger than given amounts. In such cases the table

starts with the total number of items studied and the numbers get smaller and smaller until they reach zero. The former (as in Table 3.8) are called *less-than frequencies,* while the latter (as in Table 3.9) are called *more-than frequencies.*

TABLE 3.9.—HYPOTHETICAL EXAMPLE SHOWING NUMBERS OF RUBBER BANDS WITH BREAKING POINTS ABOVE STATED AMOUNTS

Breaking Point (pounds)	Number of Bands with Breaking Point Equal to or above That Stated at the Left
4	123
7	118
10	95
13	27
16	6
19	0

3.8. Graphic Presentation: the Histogram.—As was pointed out in Sec. 3.1, the impression received by inspecting large numbers of individual figures is not sharp and clear-cut. In order to get a quick impression of the approximate sizes of the items, the statistician usually classifies them in a frequency table. The figures in Table 3.2 or 3.3 give one a very much quicker and more accurate idea of the marks that were received by these students than can be obtained from all the distracting detail of the original figures which are given on page 24. But, as will be pointed out in more general terms in Chap. XVI, perhaps the fastest way of all to get a general impression of the characteristics of a mass of statistical material is to present them in pictorial form, by means of graphs.

When dealing with frequency distributions, one of the simplest of the graphical methods of presentation is the *histogram.* This is made by laying out a horizontal scale, representing the sizes of the items (that is, the students' marks, or the breaking points of the rubber bands, etc.), and erecting thereon bars of various length, the lengths of the bars showing the numbers of cases. The data of Table 3.3, for example, are shown in a histogram in Fig. 3.1. It will be noted that the frequencies of the five classes of the table are now represented by five bars.

The base line is marked off to represent the marks received, and since the class limits in the table run from 0 to 250 our scale runs through the same values. The chart enables us to see at a glance that the commonest marks were those between 100 and 150, that there were very few marks below 50 or above 200, etc.

If the class intervals in our original table had been unequal in size, it would have been much more difficult to make our

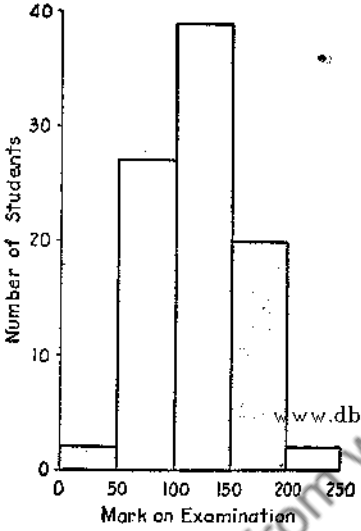


FIG. 3.1.—Frequency histogram of data of Table 3.3.

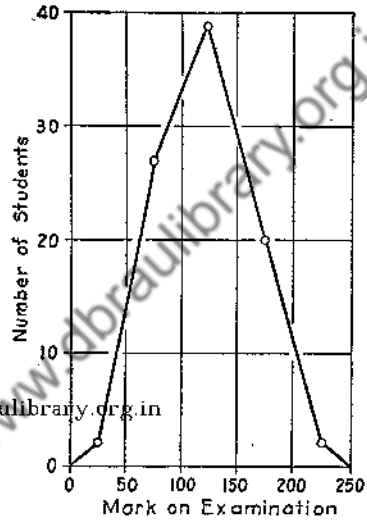


FIG. 3.2.—Frequency polygon of data of Table 3.3.

histogram. Here again we note the value of equality in class intervals. When it is necessary to depict data from a table in which there are unequal intervals, however, adjustments can and should be made as explained in Sec. 3.19.

3.9. Graphic Presentation: the Frequency Polygon.—As an alternative to the histogram, the data of Table 3.3 could be represented by a line connecting the mid-points of the tops of the bars of Fig. 3.1. In such a case we would locate the class marks on the horizontal scale at the base, and over each class mark we would locate a point corresponding with the frequency within the class. These points would then be connected by straight lines, as in Fig. 3.2.

It will be noted in Fig. 3.2 that our frequency distribution is represented by a line that starts on the base line at a point

one half class interval below the mid-point of the smallest class. If then passes through a series of points, each one vertically above one of the class marks on the scale at the base. Each of these points lies at a distance above the base scale proportional to the frequency in the class in question. The line finally falls back to the base line one half class interval above the largest class mark.

In the case of the *frequency polygon*, as such a figure is called, as in the case of the histogram, the problem becomes more complex when class intervals are unequal, and in such cases it is necessary to make adjustments similar to those described in Sec. 3.19.

The frequency polygon is perhaps more likely to be misleading than is the histogram, since the uninitiated is apt to attempt to read frequencies from the line at points between class marks. It must be remembered that both the histogram and the frequency polygon are based on a frequency table that gave the numbers of cases within various classes, but showed us nothing about how they were distributed within those classes. The bars of the histogram obviously show the facts for classes as a whole, but the unwary reader is likely to select some particular point on the base line of the frequency polygon and try to read the corresponding frequency from the line. For example, in Fig. 3.2 he is likely to interpret the diagram to mean that 14 people received marks of 50, since the line seems to have a height of about 14 above the point that represents a mark of 50 on the scale at the base. Yet a glance at Table 3.1 will show that really not a single student received a mark of 50. The frequency polygon must be interpreted as showing the numbers of cases within classes, and not the numbers of cases at particular points.

3.10. Graphic Presentation: the Frequency Curve.—If we could make our class intervals smaller and smaller, the bars in Fig. 3.1 would become narrower and narrower. Likewise the numbers of cases in the classes would become smaller and smaller (see Fig. 3.6). But if we could study larger and larger numbers of cases—not 90 students, but 900, or a million, or a limitless number—we could still make the bars narrower and narrower without making them disappear altogether. In such a case the line connecting the tops of the bars in Fig. 3.2 would probably come closer and closer to a smooth curve. The scientist assumes

that the values in a frequency distribution are not just chance affairs, but that they are distributed according to some law. The smooth curve which we should get if we could study enough cases is called a *frequency curve*. We often try by one means or another to estimate what these frequency curves must be like. A large part of Chaps. VII and VIII is devoted to a study of certain types of frequency curves and the methods of describing them. We can always make a frequency polygon or a histogram from a frequency table, showing how the items were actually

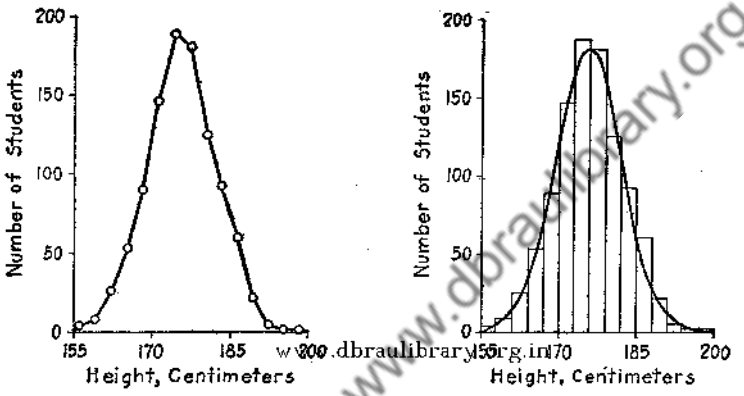


FIG. 3.3.—Frequency polygon, frequency histogram, and frequency curve, all based on the data of Table 5.1, page 82.

distributed; but we can never do more than estimate the nature of the frequency curve that underlies the data. When we do estimate such a curve by any of the means described hereafter, we usually draw it as a smooth curve on our diagram, thus distinguishing it from a frequency polygon, which is drawn with straight lines with breaks at the class marks. Figure 3.3 shows at the left a frequency polygon of the data of Table 5.1, page 82, while the right half of the figure shows a histogram of the same data. Superimposed on the histogram is a frequency curve computed by methods described in Chap. VII. While the histogram shows the way that the heights were actually distributed in the cases studied, the frequency curve represents, on the basis of certain assumptions, the underlying law of the distribution of men's heights.

3.11. Graphic Presentation: the Ogive.—Just as the frequency polygon represents an ordinary frequency table, so we could draw

a chart that showed the data of a cumulative frequency table such as Table 3.8 or 3.9. At the left of Fig. 3.4, the data of Table 3.8 have been so depicted, while at the right of the same figure are shown the data of Table 3.9. Charts of this kind, representing cumulative frequency distributions, are called *ogives*. Often the vertical scale is drawn to represent percentages of the total number of cases, running from 0 to 100 per cent. Such an arrangement makes it easier to compare two ogives based on different numbers of cases.

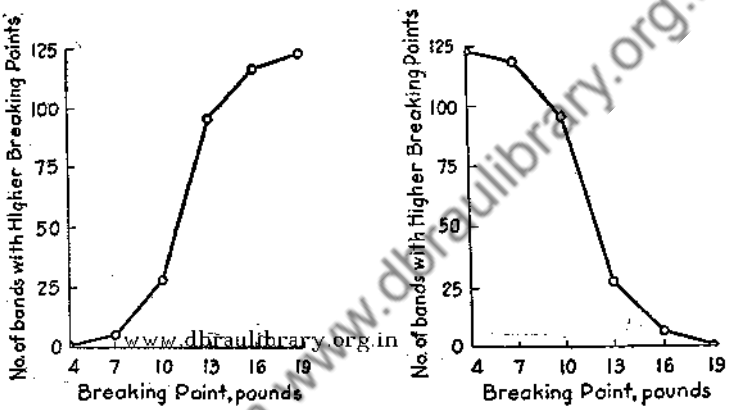


FIG. 3.4.—Two ogives. The left-hand section is a "less-than" ogive, and the right-hand section is a "more-than" ogive.

3.12. What to Look For in a Frequency Table.—As was pointed out in Sec. 3.10, the scientist assumes that every frequency distribution tends to follow some design or pattern. Different plants or animals or observations of physical phenomena are not all exactly alike, but neither do they differ planlessly. Until one has learned to think in terms of frequency distributions, he has not really become a scientist. Frequency tables, or their graphic counterparts, are basic to an understanding of scientific work in general, and particularly to that aspect of it which we study in statistics.

The trained statistician gets from a frequency table a good summary picture of the distribution on which it is based. He notes the approximate maximum and minimum sizes of items which are included and the points, if any, of heaviest concentration. If the curve rises toward a high point somewhere toward

the center, he notices this fact and the approximate position of the peak of the curve.

As an illustration we can look back again at Table 3.3, or at either Fig. 3.1 or 3.2, both of which are based on that table. The statistician looking at this diagram would note at once that the marks tend to pile up in the center, somewhere near the score of 125, and that marks above 200 or below 50 are very unusual. If the class interval had been smaller and the number of classes correspondingly larger, it is likely that even more information could be derived at a glance.

3.13. Common Shapes of Frequency Curves.—While a histogram or frequency polygon might assume almost any shape, long experience with varied kinds of data has shown that most distributions tend to fall into one or another of a relatively small number of classes. It is consequently assumed that most frequency curves assume a reasonably small number of shapes.

By far the largest proportion of frequency distributions seem to be *mound-shaped* or *humpbacked*, with small numbers of cases near the extremes and larger numbers of cases near the center. With certain kinds of data, there seems to be good reason for anticipating that the values would be arranged in some such pattern, as we shall see in Chap. VII; but even where there is no a priori ground for expecting it, we find over and over again that distributions of radically different kinds of data from distant branches of science assume this mound-shaped form.

Sometimes the mound-shaped distribution is symmetrical, with the right-hand side of the curve presenting a mirror image of the left-hand side. In other cases, even though there is a high point in the curve somewhere between the two extremes, the curve is asymmetrical, or *skewed*. A symmetrical frequency curve is shown at the left in Fig. 3.5 and an asymmetrical curve at the right. We shall have occasion to study symmetry and lack of symmetry at a later stage (see Chap. VIII).

But it would be a mistake to assume that frequency distributions are always mound-shaped. Sometimes a frequency distribution starts with a high point at the left end and falls lower and lower as one moves toward the right. Possibly a curve might start at a low point on the left and run higher and higher until its highest point was at the extreme right. Such a dis-

tribution is called a *J-shaped distribution* as distinct from the mound-shaped distributions which are much more common.

Suppose you were to investigate all the women in the United States between the ages of 15 and 20, and you were to find how many had never been married, how many had been married once, how many twice, how many three times, etc. Your data would form a frequency distribution, and the chances are pretty good that it would be J-shaped, starting very high on the left and falling lower and lower. Or imagine that you could classify

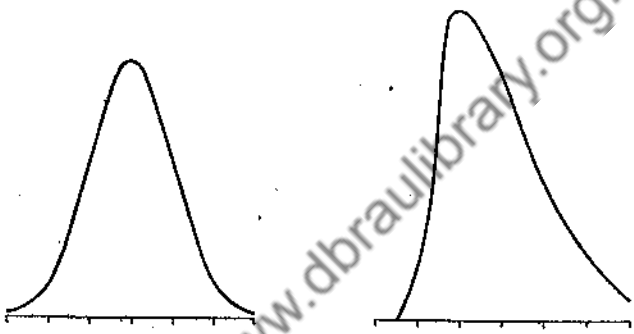


FIG. 3.5.—Symmetrical frequency curve at the left, and skewed curve at the right.

the men of the United States according to the numbers of warts on their noses. Presumably you would again find a J-shaped distribution, with the largest class being those with no nasal warts, the next-largest class being those with one such disfigurement, and with the number of men falling as the number of warts rose. These illustrations should help the student to understand that there is nothing unnatural about J-shaped distributions—that mound-shaped distributions are not "correct" or "proper." With certain kinds of data one seems in practice to find mound-shaped distributions, but with other sorts of data it is just as natural to find other patterns.

Once in a long time one finds a distribution that yields a curve with a low spot in the middle and high spots at both ends. Such a distribution is called a *U-shaped distribution*. It has been shown that the percentage of cloudiness at certain weather stations seems to follow the U-shaped distribution; that is, there are many days when the sky is completely obscured by clouds, and many days when there are no clouds at all. But as one comes closer and closer to the point where half the sky *is*

clouded and half clear, he finds fewer and fewer days to use as illustrations. It has also been suggested that marks in certain difficult and advanced college courses tend to run in U-shaped distributions, on the theory that the only people who elect the courses are those who are either very good at the subject or too ignorant even to understand their own lack of ability. U-shaped distributions are so uncommon in practice, however, that statisticians look anxiously for them to use as illustrations, and the ordinary student does not need to expect to encounter them.

3.14. Common Shapes of Ogives.—Since an ogive is a graph of a cumulative frequency curve, it is evident that there will be a definite relationship between the shape of the frequency curve and the shape of the ogive based on the same data. An ogive, as we have seen, either starts at the bottom and works to a maximum or starts at the maximum and works down to zero (see Sec. 3.11). But in order for an ogive to be a straight line, it would be necessary for the frequencies of all classes in the frequency table to be equal, since in that case each time we added a new class we would add the same frequency, and our line would rise always at the same rate. This uncommon sort of frequency distribution is called a *rectangular frequency distribution*, and it would be represented by a frequency table in which each class had the same frequency; or by a histogram in which all bars were the same length; or by a frequency polygon which was a straight, horizontal line; or by an ogive which was a straight line, rising or falling according to whether we have more-than or less-than frequencies (see Sec. 3.7).

But the commonest kind of frequency curve, as we saw in the preceding section, is the mound-shaped curve. In such a distribution, the first few classes are small, getting larger and larger for a time, reaching a maximum ultimately, after which they get smaller and smaller. If we are to add these classes to form an ogive, it is evident that our line will start out low (if we have less-than frequencies), and at first we shall add only small increments to it. But as we pass to the larger classes, the frequencies become greater, so each time we add a little more than the time before. For this reason, our ogive rises more and more steeply until we reach the point corresponding to the peak of the frequency polygon or the largest frequency in the frequency table. Thereafter we keep on adding classes, and consequently

our ogive continues to rise—but each added class is smaller than the one before it, and therefore we add less and less each time. For this reason our ogive rises more and more gradually, finally tending to flatten out and approach the horizontal as it nears the top. We thus see that the ogive which corresponds to a mound-shaped curve is (if we use less than frequencies) a rising line in an S-shape (see Fig. 5.2, page 96). If we use more-than frequencies, a similar line of reasoning will show that we get a falling line with a reverse S-shape. In fact, it is because mound-shaped distributions are most common, and because their ogives have this characteristic S-shape, that these cumulative-frequency curves are called *ogives*. The student will recall that the so-called *ogee curve* of architecture or in furniture is an S-shaped curve, and the ogive gets its name from its common S shape.

3.15. Making a Frequency Table: How Many Classes?—It is now time to leave our general discussion of the nature of frequency distributions and pay some attention to the practical problems encountered in the actual making of frequency tables. If you were faced with the problem of making a table from a large number of original figures such as those listed on page 24, your first problem would be to determine how many classes to make. Should you divide the 90 marks into 17 different groups, as in Table 3.2, or into 5 different groups, as in Table 3.3, or should you decide on some other number?

It is evident at once that the number of classes in a frequency table depends on the size of the class interval. In Table 3.2, where the class interval is 10, there are many more classes than there are in Table 3.3, where the class interval is 50. In fact, the number of classes and the size of the class interval will be roughly, though not exactly, in inverse proportion.[1]

[1] The student with a mathematical turn of mind will be interested in proving for himself that there is one case where one can tell in advance something about the relationship between the number of classes and the size of the class interval. This is the case where we have made a table with some given class interval, and we make a new table with a smaller class interval $1/n$th as large as the old one, where n is an integer. In such a case, if the n new classes are contained wholly within one of the old classes, not overlapping at the limits, it should be easy for the student to demonstrate that the new table with smaller classes may have as many as n times the former number of classes (in which case the class interval and the number of classes have varied in exact inverse proportion), or the new number of

It may seem at first that the number of classes is immaterial. But some idea of the importance of a correct choice in the matter may be obtained from Fig. 3.6, which shows the data of Table 3.1, page 25, in four different histograms, with class intervals of 100, 50, 25, and 10. It will be seen at once that when the class interval gets too small the diagram loses that simple regularity which characterises the underlying law of the distribution.

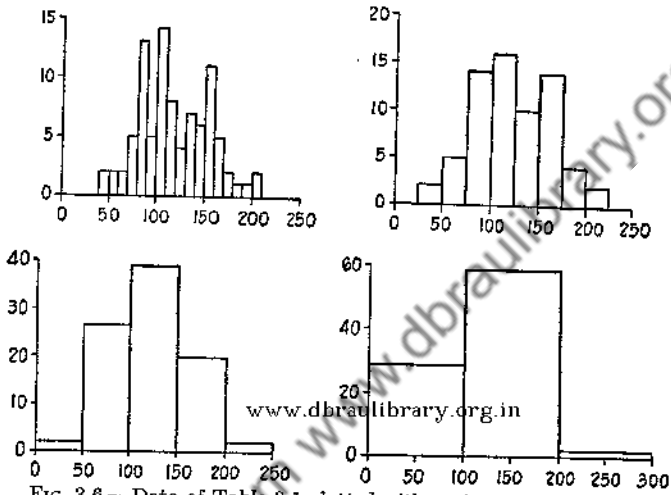


FIG. 3.6.—Data of Table 3.1 plotted with various class intervals.

We begin to get all sorts of erratic variations in the lengths of the bars. This is the result, at least in part, of the fact that the number of cases in each class has become very small and, therefore, particularly unreliable and subject to chance fluctuation. Just as you would probably hesitate to estimate the average weight of newborn giraffes after having seen but two or three of them, so you would not expect to get much accuracy from a class that contained but three or four cases.

Inspection of Fig. 3.6 shows, however, that as the class interval grows larger, and the number of classes grows smaller, we get enough cases in each class so that the erratic variations tend to disappear, and the underlying pattern becomes much plainer. To be sure, we can go too far in this direction, making

classes may be as much as $2(n - 1)$ smaller than n times as large as the first grouping. If the first coarse grouping had m classes, the new grouping may yield as many as mn or as few as $mn - 2(n - 1)$ classes.

so few classes that no pattern at all is evident. If we were to take a class interval of 300 in this illustration, all our cases would fall in the first class from 0-299, and we would see nothing of the nature of the distribution at all. Hence we can say, first, that we want the number of classes to be both small enough and large enough to show the nature of the distribution.

In addition, we are to see at later points in this book that the statistician often treats all the cases in a given class as though they were equal in size, all being equal to the class mark. This is a very useful assumption, and will not give us any great error if the classes are reasonably narrow. But if the classes get too wide, it will patently be unwise to assume that all the cases in a given class are even approximately the same size.

Also if we make our class intervals too small, we lose one of the major advantages that we seek from classification in frequency tables. Suppose, in the case of the student marks, that we set our class interval as low as one unit. Then we get Table 3.1, page 25. Here we have almost as many classes as the original number of cases. It is the purpose of frequency table to condense and compress our data, to rid them of their minor peculiarities, and to present them in summary form so that we can grasp them quickly. We could even use a class interval smaller than one, say one-tenth. Then our classes would look something like this

81.05–81.14
81.15–81.24
81.25–81.34
etc.

Then, since our original figures were all whole numbers, 9 out of every 10 classes would be vacant.

We shall mention in Sec. 6.9 another suggestion as to the number of classes in a frequency table, but for the time being we can summarize by saying that the number of classes should be large enough (and the class interval small enough) so that all items in a class may reasonably be treated as equal without too much error, and so that the general pattern of the distribution is not obscured by lumping together too large a proportion of the items in a very small number of classes. On the other hand, the number of classes should be small enough (and the class interval large enough) so that our data are compressed into a

reasonably small number of classes, so that there will be no vacant classes unless near the extremes of the data, so that the pattern of the distribution is not obscured by erratic fluctuations from class to class, and so that there is a reasonably large number of items within each class except possibly near the extremes of the data.

3.16. Making a Frequency Table: Rules of Thumb.—Instead of a general discussion of the issues involved, many authors have contented themselves with stating arbitrary rules as to the numbers of classes to be used in a frequency table. Perhaps commonest are statements that the number of classes should vary between 12–25 or between 15–20, or some other arbitrary limits.

The student will see at once that no general statement can be made which will cover all frequency distributions. For example, if we have but 25 cases we evidently cannot use even as many as 10 classes and get reasonable numbers of cases within most of them. On the contrary, if we have 10,000 cases we may well be able to spread them over 50 classes and still get a good, smooth curve which shows well the general nature of the underlying pattern. The number of classes and the number of cases is directly related.

At least one author has made an effort to set up a definite rule by means of which the student can determine the number of classes for his frequency table if the number of cases is known. This rule, which we shall call *Sturges' rule*, after its author,[1] states that the number of classes is determined by the formula

$$m = 1 + 3.3 \log N$$

where m is the number of classes and N is the number of cases. For example, if we have 842 cases and wish to make a frequency distribution, Sturges' rule would tell us to find the number of classes as follows:

$$m = 1 + 3.3(2.92) = 1 + 9.6 = 10.6$$

In other words, we should use about 10 or 11 classes for our distribution.[2]

[1] H. A. STURGES, The Choice of a Class Interval, *Journal of the American Statistical Association*, Vol. 21, 1926, pp. 65–66.

[2] The student will notice that the numbers in this computation have been

Sturges' rule is based on assumptions that we have not yet studied. In particular, the formula is derived from a consideration of the expansion of the binomial, which, as we shall see in Chap. VII, gives a good approximation to many of the more common frequency distributions. Solving Sturges' rule for various sizes of classes and numbers of cases gives us the directions which are summarized in Table 3.10.

TABLE 3.10.—NUMBERS OF CLASSES TO USE IN FREQUENCY TABLE WITH VARIOUS NUMBERS OF CASES; ACCORDING TO STURGES' RULE

If the Number of Cases Lies between	Use This Number of Classes
1	1
2	2
3–5	3
6–11	4
12–22	5
23–45	6
46–90	7
91–181	8
182–362	9
363–724	10
725–1,448	11
1,449–2,896	12
2,897–5,792	13
5,793–11,585	14
11,586–23,171	15
23,172–46,341	16
46,342–92,681	17
92,682–185,363	18
185,364–370,727	19
370,728–741,455	20
741,456–1,482,910	21

rounded off greatly, since we want an answer with but one or two significant figures. The formula is sometimes given as

$$m = 1 + 3.321920091(\log N)$$

but our study in Chap. II should have demonstrated the foolishness of such pretended accuracy.

For example, our earlier computations showed us that with a distribution of 842 cases we should use 10.6 classes. Table 3.10 shows us immediately without computation that we should use 11 classes.

Sturges' rule is easy to apply and has the advantage of definiteness; but for ordinary frequency distributions, that is about as far as it goes. Most statisticians seem to be agreed that the Sturges formula gives too many classes when the number of cases is small and too few classes when the number of cases is large. No statistician, for example, would think of making a frequency distribution of 4 classes if he had only 8 cases, nor would he feel that a distribution of 1000 cases need be confined to 11 classes. Actually, the choice of the number of classes to use will have to depend mainly on the nature of the data studied, and on the units in which they are stated, far more than on any arbitrary rule laid down in advance. Perhaps we can say again that we want to get the class interval small enough so that all items in a class can be treated as roughly the same size, but that, subject to this restriction, the fewer classes we can make and still show the underlying pattern of the distribution, the better.

3.17. Making a Frequency Table: Choosing the Class Interval. When, by means of Sturges' rule or by some other means, we have decided on the approximate number of classes for our table, the next problem is that of selecting a class interval which will yield that number of classes. Let us go back, for example, to the data with which we opened this chapter, giving marks received by 90 students on an examination (see page 24). Sturges' rule would tell us that we should have 7 or 8 classes. Suppose, for purposes of illustration, that we accept these figures as correct. What class interval should we use to get 7 or 8 classes? The answer is easy to determine. Inspection shows that the highest mark received by anyone was 206 and the lowest mark was 43. The range, or difference between the highest and lowest values in the distribution, was $206 - 43 = 163$. If we want to divide these 163 units into 7 classes we get $16\frac{3}{7} = 23+$ as our class interval. If we want to get 8 classes we should use $16\frac{3}{8} = 20+$ as our class interval.

It would be very foolish for a statistician to follow any rule so slavishly that he would set his class interval at exactly $16\frac{3}{7}$ or $21\frac{2}{7}$ just because his arithmetic yields that quotient. He will

save time and effort and get results just as good if he uses a class interval that is reasonably convenient. In the two cases just illustrated, for example, the statistician would be likely to choose an interval of 25 where the rule gives 23+, and an interval of 20 where it gives 20+. Class intervals that are in tens or multiples of tens, or in units or exact decimal values of units, are by far the easiest to use. Class intervals of 1, 2, 3, 5, 10, 20, 25, etc., are the most common. We can therefore state our rule for finding the class interval as follows: (1) Find the range (difference between the largest and smallest value). (2) Divide the range by the number of classes that you have set up as being approximately right. (3) Use the quotient as the approximate class interval, but round it off to a whole number, and if possible to some number easy to work with in classifying the items.

Where one is dealing with very large numbers of cases, it is not even necessary to determine the exact range. A rather hurried inspection will usually show approximately the largest and smallest items, and from them an approximate range can be computed which is just as good as the exact range, since our answer is to give but an approximation to the class interval at any rate.

3.18. When to Use Unequal Class Intervals.—It has been pointed out over and over again in this chapter that there is advantage in using uniform class intervals throughout a frequency table *if it can be done reasonably.* But now we must pay some attention to those cases where there is good reason to use unequal class intervals. First let us see what sorts of cases there are in which unequal intervals may be desirable. The principal reasons for equal class intervals are that the frequencies are then directly comparable from class to class and that statistical computations are greatly facilitated. But even these advantages do not in all cases outweigh the advantages of unequal intervals.

In the first place, we have cases of badly skewed distributions, where one end of the curve runs far, far away from the peak. For example, a frequency distribution of the incomes received by people in the United States would show that most of the incomes are bunched rather closely around $500 to $2000. Relatively few people receive incomes below $500 and relatively few over $2000 per year (see Fig. 5.3). If we want to show how these incomes are distributed, we cannot take a class interval

of $5000, or even of $2000, or we shall lump all these cases together in one class and obscure the shape of the distribution entirely. Yet suppose we were to decide on a class interval of $1000 (which would still be far too large for actual use). The largest incomes were several million dollars a year. If we are to include enough classes at $1000 per class to reach the highest incomes, it would be necessary to make thousands of classes. This is out of the question. Hence we make small classes where the cases are numerous and larger classes where the cases are sparse. If we were to make uniform class intervals, and have a reasonably small number of classes, it is clear that we would get a J-shaped distribution, since our first class would have to contain incomes from zero to perhaps $100,000 or more. When a statistician finds a J-shaped distribution, he always tests the frequencies at the more populous end by trying smaller class intervals to see whether it is actually J-shaped or contains a hidden mound near the end.

A second reason for using nonuniform class intervals may be to get similar cases together. Vital statistics are often classified in the following age groups, for example:

Under 1 year	10–14.9
1–1.9	15–19.9
2–2.9	20–29.9
3–3.9	30–39.9
4–4.9	40–49.9
5–9.9	50–59.9
	etc.

In such a classification, the very young children are put in small groups largely because it is thought that the problems of children under the age of 1 year differ enough from the problems of children between one and two years so that it will be advantageous to classify them separately, while the problems of people of the age of 50 and those of people aged 59 may be roughly similar from the standpoint of the vital statistician. The scientist would be foolish to lump together cases which should be treated separately merely in order to retain uniformity of class intervals.

A third reason for using nonuniform class intervals in some tables is to keep data confidential. Uniform class intervals are

likely to bring small frequencies in classes near the extremes. Where there are only one or two cases in an extreme class, it may be easy for informed people to figure out who is who, discovering what income this firm gets, or what are the costs of that firm, etc. Many figures, especially those collected by the government, are obtained on the basis of promises that they will be held in confidence, and uniformity of class intervals may make this retention of confidence impossible.

For any of these reasons, or, perhaps, for others, it may be decided that the frequency table should be made up with unequal class intervals even though there are many disadvantages of such a course. In such a case, certain precautions should be taken to make sure that the results are not misleading.

TABLE 3.11.—ILLUSTRATING THE USE OF UNEQUAL CLASS INTERVALS

Class Limits	Number of Cases
0– 9	5
10– 19	22
20– 29	35
30– 39	39
40– 49	41
50– 59	39
60– 69	35
70– 89	48
90–109	28
110–129	16
130–169	12

3.19. How to Use Unequal Class Intervals.—Let us turn our attention to Table 3.11, in which there are unequal class intervals. The dangers inherent in the use of such grouping are immediately apparent. As we glance through the table, we get the impression that the heaviest concentration of cases falls in the class 70–89. It appears that the frequencies get larger and larger until we reach a peak at 40–49, after which we have a slight fall, rising again to an even higher peak in the class 70–89, after which the frequencies fall again.

Yet when we look at the table carefully, we notice that the class interval is twice as great in the 70–89 class as it is in any of the preceding classes. If the frequencies were concentrated

just as heavily in this class as in the preceding one, there should be twice as many cases, since the class interval is twice as great. Yet there are not twice as many cases. The preceding class has 35 cases as compared with 48 cases here.

If we are to make the cases comparable, it should be evident that we must divide each frequency by its class interval. This would give us a new table such as Table 3.12.

TABLE 3.12.—ILLUSTRATING ADJUSTED UNEQUAL CLASS INTERVALS

Class Limits	Frequency per Unit of Class Interval
0– 9	0.5
10– 19	2.2
20– 29	3.5
30– 39	3.9
40– 49	4.1
50– 59	3.9
60– 69	3.5
70– 89	2.4
90–109	1.4
110–129	0.8
130–169	0.3

Now we see that the frequencies build up smoothly from each end toward the middle and that there is really one high point rather than two. Perhaps this can be visualized even better from Fig. 3.7. In the upper part of this figure, we see a histogram of the data of Table 3.11 made without any adjustment for inequality of class intervals, and therefore giving the incorrect impression. The lower part of the figure shows the data correctly plotted from Table 3.12, and in this case one gets the correct impression at once. In making this correct histogram, each bar covers a width on the base line corresponding to its class interval, and the area of the bar (width times height) is proportional to the frequency actually found in the class. In order to get this proportionality, the heights of the bars are not proportional to the original frequencies, but proportional to the adjusted frequencies of Table 3.12.

Since our class intervals in Tables 3.11 and 3.12 are 10, 20, and 40, it is immaterial whether we make our adjustments by

dividing by 10, 20, and 40 or by dividing by 1, 2, and 4. Either will put our results in the same proportions. It is perhaps easier to state the rule for adjusting frequencies where there are unequal class intervals by saying that one divides each frequency by the corresponding class interval, but as long as the frequencies are

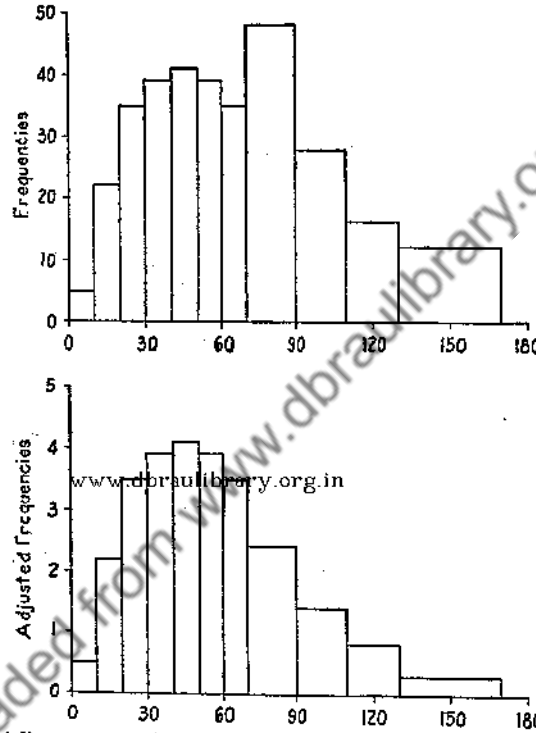


FIG. 3.7.—Adjustment of histogram for unequal class intervals. Same data with and without adjustment.

divided by numbers proportional to the class intervals the proper results will be obtained.

3.20. Logarithmic Frequency Classes.—When a frequency histogram looks like the lower part of Fig. 3.7, some statisticians suggest that there may be real advantage on technical and theoretical grounds in using unequal class intervals of a particular kind. These are class intervals so arranged that the successive lower class limits will be in constant proportion, rather than differing by constant amounts. Suppose that we have a frequency distribution in which the items run from a low value of 15 to a high value of 200, and we want to divide them into 10 classes of this kind. We might fall back on the familiar formula for geometric progressions, which tells us

that the last term of such a progression can be found from the first term by the formula

$$L = F(1 + r)^n$$

If we let L represent the largest value in the distribution, F the smallest value, and n the number of classes desired, we can solve for the value of $1 + r$ (using logarithms) and discover the required value by which each lower limit is to be multiplied to find the lower limit of the next class. In our problem we have assumed a lowest value of 15 and a highest value of 200, with 10 classes desired. Hence our formula becomes

$$200 = 15(1 + r)^{10}$$

Solving, we discover that $1 + r = 1.295$. This tells us that we should multiply each lower limit by 1.295 to find the next lower limit. If we start with the lower limit of the smallest class at 15 and multiply repeatedly by 1.295, we get the following lower limits:

15, 19.42, 25.15, 32.57, 42.18, 54.62, 70.73,
91.60, 118.62, 153.61, 198.92

Had we not dropped decimals, this last figure would have been exactly 200. We would now set up our class limits as follows:

15.0– 19.4	54.7– 70.7
19.5– 25.1	70.8– 91.6
25.2– 32.5	91.7–118.6
32.6– 42.1	118.7–153.6
42.2– 54.6	153.7–198.9

It will be noticed that these class intervals are unequal, with an interval of about 4.5 at the beginning and an interval of about 45 at the end. We could now go back to our original figures and distribute them among these 10 classes to get our frequency distribution with unequal intervals. Some statisticians have suggested that if this type of distribution yields a frequency polygon which is more symmetrical than that obtained from the same data with equal class intervals, one should use the logarithmic class intervals and should use the geometric mean instead of the arithmetic mean (see Chap. V). Logarithmic frequencies of another kind can be fitted by a method described by George R. Davies in the *Journal of the American Statistical Association*.[1] The beginning student, however, will do well to confine himself to equal class intervals or to the simple adjustments suggested in Sec. 3.19.

3.21. Making a Frequency Table: Locating the Class Marks.—Having decided upon our class interval in accordance with the directions of Sec. 3.17, we still have to decide where to locate the class marks, or, what amounts to the same thing, where to locate the class limits. Suppose that we have decided to use a class

[1] Vol. 20, p. 467.

interval of 10, and the smallest value in our distribution is 68. Shall we set up our first two classes thus:

60–69
70–79
etc.

or shall we use the following:

62–71
72–81
etc.

or might we decide on such a peculiar arrangement as

64.38–74.37
74.38–84.37
etc.

In each of these cases, the class interval is 10, and it is obvious that there are limitless other combinations of class limits that could be used, still retaining the class interval at this size. The decision as to the size of the class interval has not completed our problem of setting up our frequency table, since we still must locate the class limits.

Unless there is some good reason to the contrary, we usually take class limits that are whole numbers, such as those given in the first two of the three illustrations of the preceding paragraph. And if the class interval is 5, 10, 25, 50, or 100, or some such number, we commonly make each lower limit an exact multiple of the class interval, as in the first of the three illustrations in the preceding paragraph. Some writers have suggested that the class marks, rather than the lower limits, be made whole numbers and, if possible, multiples of 10. Their argument is that such an arrangement will save time in later computation when, as we shall see, the assumption is made that all values in a class are equal in size to the class mark. But when the table is such that computations can be made by the short method rather than by the long method (as explained in Chap. V), there is no advantage in having integral class marks, and classification of items is speeded up usually by having integral class limits.

Sometimes the data being tabulated run down to zero and then stop, negative values being impossible. For example, we

might consider a frequency distribution showing the number of corporations hiring various numbers of employees. No concern would hire a negative number of men, so that our values may not run below zero. In such a case, if the values actually do run down to or close to zero, it is evident that we cannot maintain uniform class intervals at the lower end of the table with some locations of class marks, while we can with others. Suppose again that we are using a class interval of 10. If we have classes of 33–42 and 23–32, etc., our low classes will have to be 3–12 and 0–2. This make the lowest class interval smaller than the others. Sometimes, then, the location of the class marks would be determined by our wish to keep even our smallest class uniform with the others.

A third consideration in the locating of class marks becomes prominent in those distributions where certain values are common and other values either do not appear at all or appear uncommonly. For example, it may be that tickets to a ball game are sold at 25 cents, 50 cents, and $1, but that no other values occur. No ticket will be purchased at 38 cents or at any intermediate value. Yet again, we might be listing the numbers of rooms in houses, in which case we would get 5-room houses or 6-room houses, but there would be no houses with 5.4 rooms. Distributions of this character, where only certain values are possible, are called *discrete distributions*. We can contrast them with *continuous distributions*, in which any intermediate value can occur. For example, men's heights do not necessarily fall at 70 or 71 in. or any other particular value. It is quite possible for a man's height to be 70.342 in., or any other conceivable value within the whole range from the shortest to the tallest person. Most distributions with which the statistician deals are either continuous, or the breaks are so small compared with the range of the data that they can be safely treated as continuous. As an example of the latter, if we were classifying incomes received by people in the United States one might argue that the distribution is discrete, since one can receive $1,043.21 or $1,043.22, but not between. Yet 1-cent breaks are so small when compared with the vast range of incomes that there is little error in saying that the distribution is continuous.

Some people have described distributions as *homograde* where we have called them discrete and as *heterograde* where we have

called them continuous. But unfortunately there is not uniformity of usage for these terms, since other authorities would say that a homograde series is one in which there are only two possibilities—a characteristic is either present or it is absent. Thus a division of people into those who are vaccinated and those who are unvaccinated would, by this definition, constitute a homograde series, while a heterograde series would be one that showed variation in magnitude. Most statisticians would distinguish these last two cases in other terms by saying that when we study those things which are either present or absent we are studying *attributes*, while when we study things in which the magnitude can assume many different values we are studying *variables*. There is no confusion if we speak of continuous and discrete data, nor if we speak of attributes and variables. There is, however, difference in the usage of the words homograde and heterograde.

With discrete data, it is natural that there should be bunching of values at particular points, since no intermediate values can occur. But sometimes we get a similar bunching even in cases where intermediate values could occur. For example, when people are asked their ages they usually give whole numbers of years, leaving off intermediate fractions; and census data also show that there is a very real tendency for people to give their ages in multiples of 5 or 10, stating that they are 40 or 45 years old even though they may really be 41 or 44. Estimated values are particularly likely to show this bunching. If we ask people to estimate distances or ages or weights, they are apt to do it in whole numbers and in multiples of 5 or 10. A farmer may say that he grows 30 acres of wheat or 35 acres, but he is very unlikely to state that he grows 31.398 acres even though that may be the fact. Since many statistical computations are based on numbers which were originally estimates it is worth while to keep this fact in mind.

Regardless of the reason for bunched values, whether because the data are discrete, or because they are estimated, or because they are given only approximately, or for any other reason, data which are characterized by points of marked concentration should be tabulated with the points of bunching at the class marks. This is because we shall later assume that all items in the class are at the class mark, and the error will be minimized if the bunched cases are actually located there.

3.22. Summary: Directions for Making a Frequency Table.—We can now summarize the actual steps involved in making a frequency table as follows:

1. Decide whether or not to use equal class intervals. Use equal intervals if reasonably possible. See Sec. 3.18 for suggestions concerning the use of unequal intervals. The remaining steps summarized below assume the use of equal class intervals.

2. Decide how many classes to use. For suggestions see Secs. 3.15 and 3.16.

3. Find the range or the approximate range between the largest and the smallest values in the distribution.

4. Divide the range found in step 3 by the number of classes found in step 2. Use the quotient as a first approximation to the class interval.

5. Select a class interval that is convenient—usually a whole number and possibly a multiple of 5 or 10—using the approximation of step 4 as a basis.

6. Decide where to locate the class limits. For suggestions see Sec. 3.21.

a. Lower class limits should usually be whole numbers, often multiples of 5 or 10.

b. Make sure that the lowest class can be included without altering the class interval.

c. If there is bunching for any reason (as in discrete series or where values are based on estimates) put the popular values at class marks.

7. Having laid out the class limits, distribute the original values among the classes, noting how many items fall in each class.

3.23. Suggestions for Further Reading.—It is impossible to cover adequately a good deal of important material on frequency distributions in one chapter. In Chaps. VII and VIII we shall discuss again certain particular forms of frequency distributions which are of especial statistical importance. Karl Pearson suggested methods of treating some of the commoner frequency curves. His original memoirs on the subject may be found in *Philosophical Transactions*, A, at the following three points: Vol. 186, pp. 343*ff.*; Vol. 197, pp. 443*ff.*; and Vol. 216, pp. 429*ff.* Perhaps even better for the general student than these scattered references would be the more compact treatment found in the first six chapters of W. Palin Elderton, "Frequency Curves and Correlation," C. & E. Layton, London, 1927. An even more condensed summary, with directions and illustrative examples but little discussion of underlying theory, may be found in C. B. Davenport and M. P. Ekas, "Statistical Methods in Biology, Medicine, and Psychology," John Wiley & Sons, Inc., New York, 1936. Chapter 7 of the "Handbook of Mathematical Statistics," edited by H. L. Rietz, Houghton Mifflin Company, Boston, 1924, gives a valuable discussion of frequency curves including both Pearson's forms and others.

EXERCISES

1. Suppose that you want to divide data into classes with uniform class intervals of five units. You wish to have the value 5 and its multiples at the

class marks. List some of the classes as they would appear in a frequency table.

2. Give several examples of data which are discrete and several examples of continuous data.

3. Classify the data given on page 24 into a frequency table. Make the class interval 25, and have the value 25 and its multiples at the midpoints of the classes.

4. Go to the library and record the number of pages in each of the first 100 books that you find. Classify the results in a frequency table, following the rules of Sec. 3.22.

5. Select from this or some other book three or four pages of solid reading matter, not broken up by illustrations, tables, or formulas. Count the number of words in each line, and make a frequency table showing the numbers of times that various numbers of words appear. It will be best to select only complete lines, omitting those which begin and end paragraphs if they are shorter than ordinary lines. Make your own rules on the treatment of abbreviations, hyphenated words at the ends of lines, etc. Continue until you have counted 150 to 200 lines.

6. The "World Almanac" has, for several years, published a table giving facts about "Noted Americans of the Past." The years of birth and of death are given for each such noted person. If we find the approximate age at which each of these people died by subtracting the year of birth from the year of death, we get the following figures (taken from the 1941 edition, page 660, and being data on the first 204 persons listed in alphabetical order):

33	77	76	87	79	80	80	81	75	75	66	89
56	83	74	71	65	52	76	76	50	69	73	49
94	87	70	86	69	71	85	78	83	72	71	46
53	65	66	81	90	91	78	58	81	91	84	48
81	70	68	74	88	75	48	74	66	73	77	77
76	65	72	63	54	85	74	65	48	63	47	69
59	73	46	67	77	58	60	99	59	72	73	65
84	76	41	77	50	84	75	81	72	68	45	62
84	95	59	84	86	66	62	65	60	85	66	68
62	75	60	75	59	73	72	73	59	56	62	92
65	39	30	64	76	50	78	83	82	68	37	78
82	73	92	67	81	52	71	41	94	79	78	76
56	71	70	81	48	78	93	25	71	67	34	78
62	77	67	76	89	84	55	65	92	86	79	86
83	71	69	71	63	73	45	71	83	36	77	59
44	55	37	38	57	77	80	81	40	50	64	88
64	74	85	58	71	65	77	81	84	71	70	51

Make a frequency table of these figures, following the rules of Sec. 3.22.

7. In any table the class marks always fall exactly halfway between the actual class limits. Under what circumstances will the actual class limits fail to fall exactly halfway between the class marks?

8. The families of 898 working-class men in Bolton, England, were classified in 1924 according to the number of rooms occupied by the family with the following results:[1]

Number of rooms	2	3	4	5	6	7	8	Total
Number of families	15	477	227	169	7	2	1	898

It will be noted that this table is a frequency table shown horizontally as contrasted with our usual vertical arrangement. Show the data in a histogram.

9. Show the data of the preceding exercise in a frequency polygon.

10. Make a "less-than" ogive of the data of Exercise 8.

11. Make a cumulative frequency table from the data of Exercise 8.

12. Solve Sturges' rule for the data of Exercise 6.

13. Divide the data of Exercise 6 into four classes with logarithmic frequency classes (see Sec. 3.20).

14. Suppose we are given the data of Table 3.13. Note that the class intervals are not equal. Make a histogram of the data, making proper adjustment for the inequality of class intervals (see Sec. 3.19).

TABLE 3.13.—FREQUENCY TABLE WITH UNEQUAL CLASS INTERVALS

Size of Items	Number of Cases
150–159	15
160–169	60
170–179	85
180–189	98
190–199	105
200–209	104
210–219	97
220–229	83
230–239	62
240–259	88
260–279	56
280–309	45
310–339	15

15. Make a "more-than" ogive of the data in Table 3.13. Decide in advance whether or not it is necessary to make any correction for the inequality in class intervals.

[1] Figures quoted in R. G. D. Allen, "Mathematical Analysis for Economists," p. 411, The Macmillan Company, New York, 1939.

CHAPTER IV

MEASURES OF CENTRAL TENDENCY

4.1. Averages.—We have seen that the statistician commonly groups masses of data together into frequency tables so that they will be easier to comprehend. But often he wishes to go even further, to compute some one number which will in some definite way represent all the numbers of the group. Any number that, in this way, is used to represent a whole series of values is called an *average* of those values. To be sure, the word "average" is used in common speech to mean one particular kind of representative figure—a representative figure computed in a particular way. But technically there are many kinds of averages, and sometimes the statistician uses one and sometimes another. These various representative values or type values or averages are computed in various ways, and they represent the group in various ways. It is the purpose of this chapter to investigate some of the more commonly used averages and to ascertain their characteristics.

Although there is no limit to the number of ways in which one could select a value as representative of the group, there are in practice only a few ways in which statisticians find it worth while to attack the problem. We shall confine our study to the methods that are in most common use among statisticians. In this chapter, we shall consider the ways of finding representative values when each of our original figures is given individually; in the following chapter, we shall study the same problem as it is handled when the data have been grouped together in frequency tables. And at the end of the next chapter, we shall study the use and interpretation of the results.

4.2. The Arithmetic Mean: Ungrouped Data.—The arithmetic mean[1] is the measure most people have in mind when they use the word "average." The concept is familiar to every student and needs no discussion here. The arithmetic mean of a series of

[1] This measure is called indiscriminately the "arithmetic mean," the "arithmetic average," or merely the "mean."

values is the quotient obtained by dividing the sum of the values by the number of values. We can symbolize this computation as follows:

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

where each of the original figures is represented by X.

N = the number of cases.
Σ means "the sum of."
$\bar{X}$ represents the mean of the X's.

Thus the formula should be read, "The mean of the X's is the sum of the X's divided by the number of cases."

Let us illustrate. We have five numbers ($N = 5$), as follows:

7; 4; 6; 3; 10

If we add them (ΣX) we get 30. Thus, since $\Sigma X = 30$ and $N = 5$, our formula becomes

$$\bar{X} = \frac{\Sigma X}{N} = \frac{30}{5} = 6$$

It is well to become accustomed to statistical symbols in a case such as this, where the student knows in advance what is expected of him. Every student knows how to find the average of these five numbers without taking a course in statistics and without having a Greek-letter formula to guide him. But this is a good opportunity for him to discover that statistical formulas are but shorthand directions for computations. If one understands the symbols, one knows that $\Sigma X/N$ says, "Add up the X's and divide the sum by the number of cases." And since the symbols always mean the same thing, when they have once been mastered it is easy to follow their directions. It would pay to learn them as they come. So far we have these:

X always refers to the figures with which you start. (If you start with two series of figures, one may be called X and one Y, or one may be called X_1 and the other X_2, etc.)

Σ (the Greek capital letter sigma) means "the sum of the things which follow." It is the sign for addition.

N always means "the number of cases."

If we were to find the average of the 90 examination marks given on page 24, we should use the same formula; that is, we should divide the sum of the marks by 90, the number of marks.

4.3. Weighted Arithmetic Mean.—Sometimes we wish to find the average of several numbers which are not of equal importance. In such a case it is necessary for us to add one slight complication to our method. The addition can best be explained in terms of an illustration. Suppose that there are, in a given high school, 100 freshmen, 80 sophomores, 70 juniors, and 50 seniors. On a given day 15 per cent of the freshmen are absent, 5 per cent of the sophomores, 10 per cent of the juniors, and 2 per cent of the seniors. What percentage is absent for the school as a whole? The student is likely to attempt to find the answer by adding the four percentages and dividing by 4. This would give him the following incorrect answer:

$$\frac{15 + 5 + 10 + 2}{4} = \frac{32}{4} = 8$$

We can quickly find, however, that 8 per cent is not the correct answer. There must have been 15 freshmen absent (15 per cent of 100), 4 sophomores absent (5 per cent of 80), 7 juniors absent (10 per cent of 70), and 1 senior absent (2 per cent of 50). This makes 27 students absent altogether out of a student body of 300. Our correct answer, then, is 9 per cent rather than 8 per cent.

In such a case, we commonly find the correct average by a process known as *weighting*. We determine how important each of our original numbers is and assign it a weight proportionate to its importance. We then multiply each number by its weight and add the products. The sum of the products is then divided by the sum of the weights. If we add one new symbol to those already listed in Sec. 4.2, we can represent the weight assigned to any number by the letter W. Then our formula for a weighted arithmetic mean would be

$$\bar{X} = \frac{\Sigma(XW)}{\Sigma W}$$

We see at once that this formula gives the following directions:

1. Multiply each original value (X) by the corresponding weight (W).
2. Add the products thus obtained.
3. Divide this sum by the sum of the weights.

To set up our hypothetical example in formal manner, so that we may see how the formula works, we list our original percentages (15, 5, 10, and 2 per cent) as the values of X. We assign to each a weight (W) equal to its importance. In this case the weights are the numbers of students on which each percentage was based, on the grounds that a percentage based on 100 cases deserves more weight than one based on only 10 or 15 cases. Our problem appears as follows:

X	W	XW
15	100	1500
5	80	400
10	70	700
2	50	100
Totals.....	300	2700

$$\bar{X} = \frac{\Sigma(XW)}{\Sigma W} = \frac{2700}{300} = 9$$

This time we get the correct answer, 9 per cent, at once.

The student should note that we do not weight an average merely because we are fond of statistical computation, nor because we wish to impress the layman, but because weighting gives the right answer.

Let us take one further illustration. In 1940, the populations per square mile of the New England states were approximately as follows:

State	Population per Square Mile
Maine.....	28.3
New Hampshire.....	54.7
Vermont.....	39.5
Massachusetts.....	539.6
Rhode Island.....	648.2
Connecticut.....	356.0

What was the average density of population in New England? If we add the six numbers and divide their sum by 6, we shall get an incorrect answer, 277.7 persons per square mile. Again

it is necessary, if we want the right answer, to weight the average. When we stop to think about it, we realize that the large figure given for Rhode Island is based on very few square miles, while the small figure for Maine is based on many more square miles. We should not give the original figures equal importance, but should weight them according to the number of square miles in each state. We can then work out the correct population density for New England, as follows:

State	Population per square mile (X)	Area (thousands of square miles) (W)	XW
Maine	28.3	29.9	846.17
New Hampshire	54.7	9.0	492.30
Vermont	39.5	9.1	359.45
Massachusetts	539.6	8.0	4,316.80
Rhode Island	648.2	1.1	713.02
Connecticut	356.0	4.8	1,708.80
Totals		61.9	8,436.54

The average population per square mile can now be found by the formula for the weighted arithmetic mean, as follows:

$$\bar{X} = \frac{\Sigma(XW)}{\Sigma W} = \frac{8{,}436.54}{61.9} = 136.3$$

The average density of population per square mile in New England was 136.3.

Let us note why this weighted answer is the correct one. If we multiply the population per square mile of any state by the number of square miles in the state, we shall get the total population of the state. If we then add up these products for each of the states, we shall get the total population of the district. And if we divide this total population by the total area, we shall get the population per square mile in the total area. This is exactly what we did in the example above with the single exception that we carried out our computations in thousands of square miles instead of single square miles in order to save time. We could well have rounded off our computations even further in accordance with the rules of Chap. II.

Strictly speaking, every arithmetic average is weighted. If we add several numbers and divide by the number of them, we

have merely weighted them all equally—given each a weight of 1. We can see that if each value is given a weight of 1 our formula for the weighted mean is reduced to the ordinary formula for the "unweighted" arithmetic mean.

Whenever we are taking the average of several percentages, averages, ratios (note that population per square mile is a ratio), or any other numbers which for any reason differ in their importance, we must weight the average if we are to get the right answer. Sometimes the weights are entirely arbitrary, as when a teacher decides to weight the final examination in a course twice as heavily as a regular examination given during term time or to weight laboratory work half again as heavily as recitations. One observation of a solar eclipse may be weighted more heavily than another because of better visibility, more accurate instruments, more experienced observers, or for any other reason. We shall note in the next chapter one other common type of case in which weighting is necessary.

4.4. The Median: Ungrouped Data.—The median is the value so chosen that there are just as many cases larger in value than the median as there are cases smaller in value than the median. In other words, if we arrange all the values in order of size, with the smallest item on one end and the largest item on the other, and if then we select a value in such a way that there will be the same number of items on each side of it, the value so selected is the median.

It is easier to illustrate this concept than to describe it. If we take the five values which we used in illustrating the arithmetic mean, we recall that they were

$$7;\quad 4;\quad 6;\quad 3;\quad 10$$

First we arrange them in order of magnitude. When we have a series of values arranged in order of size, we say that we have an *array*. If we arrange these five items in an array we have

$$3;\quad 4;\quad 6;\quad 7;\quad 10$$

Now let us select such a number that there will be just as many values above as below. If we select the value 4, we find but one value smaller and three values larger: this will not meet our requirement. If we select the value 5, we find two values smaller and three larger: again we have not met the requirement. Obvi-

ously the only value that will suit us is 6: there are two values below 6 and two values above it.

Let us illustrate again with the data on page 24. First we must arrange them in an array. This gives us the following:

206; 203; 191; 181; 179; 176; 168; 166; 165; 164;
163; 159; 158; 158; 158; 157; 156; 154; 153; 152;
151; 150; 149; 147; 147; 143; 142; 142; 139; 138;
136; 136; 133; 133; 131; 128; 123; 122; 121; 119;
118; 114; 114; 113; 112; 112; 112; 109; 109; 107;
107; 107; 106; 105; 104; 104; 103; 102; 102; 101;
101; 95; 94; 93; 92; 90; 89; 89; 88; 85; 85; 85;
84; 82; 82; 81; 81; 81; 81; 79; 76; 75; 75; 73;
69; 65; 57; 55; 49; 43.

Now that the items are arranged in an array, we must select a value that will divide the distribution into two parts with the same number of items in each part. We might start out at random, taking items that looked likely and seeing how many were larger and how many smaller. We might, for example, start with the value 110 and count the items which exceed 110 and those which are smaller. Trial will show that there are 47 items which are larger than 110 and 43 items which are smaller: it is obvious that we must select a value somewhat larger than 110. We could continue to try items in this manner until we discovered a value which met the requirement. Such a method, however, would be very wasteful of time. It is obvious, to begin with that we want the item that lies at the center of the distribution. Suppose we arrange three items and want one that will divide the values evenly: we must obviously choose the second item. If we have four items we must select a point between the second and third. If we have five, as in the first illustrative example above, we know that we must choose the third. Experiment will show that, if we are to select the item that will divide the distribution into two equal parts, we must select the item that is $(N + 1)/2$. Thus, if there are 5 items we must select item number $(5 + 1)/2$, which equals 3. In this case we have 90 items, and $(90 + 1)/2 = 45.5$; that is, we must select a value which is halfway between the values of the 45th and the 46th items. If we count in our array we discover that the 45th item from the bottom had a value of 112 and the 46th also has a

value of 112. The median will be halfway between them, or, since they are identical in size, will be $(112 + 112)/2 = 112$.

When the number of items in the array is even, the median is taken as the arithmetic average of the two central items. When the number of items is odd, the median is the value of the item which is item number $(N + 1)/2$ from either end. Note that we do not find the median by evaluating the expression $(N + 1)/2$. This formula merely tells us the *position* of the median. If there are 571 items, and we wish to find the median, we arrange them in array. The median is not

$$\frac{(N+1)}{2} = \frac{(571+1)}{2} = 286$$

but the median is the value of 286th item in the array. In other words, the median is found by first arranging the items and then counting them, finally taking the value of the item which is central, or, if there is no single central item, the average of the two central items.

Note, then, that, if the median mark given on an examination was 112, we mean that as many students received more than 112 as received less than 112. If we say that the median height of 100 men was 5.5 ft., we mean that as many men were taller than 5.5 ft. as were shorter than 5.5 ft. This might not be at all true of the arithmetic mean, as you will observe from the following example. The mean of the items

$$4; \quad 5; \quad 6; \quad 7; \quad 203$$

is $225\frac{4}{5} = 45$. But 45 is exceeded by only one of the items; four items are smaller. The median of these five items would be the value of the $(N + 1)/2$ item, or the value of the $(5 + 1)/2$ item, that is, the third item. This item has a value of 6: there are two items above it and two items below. Here it will be noted that the median and the arithmetic mean do not necessarily have the same value.

4.5. The Mode: Ungrouped Data.—The mode is the value that occurs most frequently. The modal income of wage-earners in the United States is the most common income, the income which is received by more people than receive any other income. If we say that the modal size of farm in a given community is

78 acres, we mean that there are more farms of this size than of any other size.

In any statistical problem with continuous data and with fine enough measurements, it is probable that no two values will exactly coincide. Hence there will be no one value occurring more often than any other value. It may be that no two men in the United States are of exactly the same height if we could measure them with enough exactitude. How, then, could there be a modal height? In such a case we should group the data and compute the estimated mode from the groups of a frequency table. Or it may be that the crudity of our measurements will be such that the data are already grouped. Thus if we can measure heights only to the nearest inch, so that all men between 5 ft. 8.5 in. and 5 ft. 9.5 in. are recorded as being 5 ft. 9 in. tall, then we have patently grouped together many men whose heights are really slightly different. In this way we may find that many men seem to have the same height, and we get the same results that we obtain by conscious grouping of the cases.

TABLE 4.1.—FREQUENCY OF APPEARANCE OF VARIOUS NUMBERS OF BLACK CARDS IN 102 DEALS OF 10 PLAYING CARDS

Number of Black Cards	Frequency
0	0
1	1
2	5
3	12
4	18
5	34
6	22
7	7
8	2
9	0
10	1

When we have discrete data, on the other hand, the mode may be easier to ascertain. Let us illustrate such a case. Table 4.1 shows the number of times that various numbers of black cards appeared in 102 deals of 10 playing cards. Here the data are patently discrete, since we may get 4 black cards or 5 black cards, but never 4.26 black cards. And here the mode is also plainly marked. The commonest number of black cards—the number

which appeared more often than any other single number—was 5. One may often find distributions in which there is no mode, that is, in which no single value appears more often than others. In yet other cases there may be two or more modes or points of concentration. In these respects the mode differs from the median and the arithmetic mean, since there is always one median or arithmetic mean and never more than one.

4.6. The Geometric Mean: Ungrouped Data.—The arithmetic mean of a group of values was found by adding them and dividing the sum by their number. The geometric mean is computed by multiplying the values together and taking the nth root. Thus the geometric mean of the numbers 7, 9, and 11 is

$$\sqrt[3]{7 \times 9 \times 11} = \sqrt[3]{693} = 8.849$$

This method of computation is useful if we are to average but two or three numbers, but if we are asked to average 12 or 50 or 200 numbers we discover that the process involves the extraction of the 12th or 50th or 200th root. This is out of the question. We can arrive at the same result, however, by another method. The student will recall that adding the logarithms of numbers is equivalent to multiplying the numbers themselves together, and that dividing a logarithm by n is equivalent to extracting the nth root of the number. We can, therefore, work our problem by adding the logarithms of the n numbers, dividing the result by n, and taking the antilogarithm of the quotient. For example, suppose we are required to find the geometric mean of the numbers 12, 17, 33, 21, and 162. The long process would involve the multiplication of the five numbers and the extraction of the fifth root. The short method involves the addition of the logarithms of the five numbers, the division of the sum by 5, and the taking of the antilogarithm. The process follows:

X	$\log X$
12	1.07918
17	1.23045
33	1.51851
21	1.32222
162	2.20952
	$\Sigma(\log X)$ 7.35988

$$\frac{\Sigma(\log X)}{N} = \frac{7.35988}{5} = 1.47198$$

Geometric mean = antilog 1.47198 = 29.65

This process of discovering the geometric mean can be symbolized by the formula

$$\log M_g = \frac{\Sigma(\log X)}{N}$$

where M_g represents the geometric mean and the other symbols have the meanings already attached to them.

In both illustrative problems of this section, we have found geometric means which are smaller than the arithmetic means of the same numbers. Experiment will show that unless all the numbers being averaged are identical in size the geometric mean of a group of numbers is always smaller than their arithmetic mean. And, of course, if a single one of the original numbers is zero, their geometric mean is also zero.

4.7. The Harmonic Mean: Ungrouped Data.—The harmonic mean of a group of numbers is the reciprocal of the arithmetic mean of their reciprocals. Thus, if we wish to find the harmonic mean of seven numbers, we first take their reciprocals. We then find the arithmetic mean of these reciprocals and take the reciprocal of the result. (The reciprocal of any quantity is the quotient that results when unity is divided by that quantity.) In the preceding section we found the geometric mean of the numbers 7, 9, and 11. Their harmonic mean would be found as follows:

$$\frac{1}{\frac{\frac{1}{7} + \frac{1}{9} + \frac{1}{11}}{3}} = \frac{1}{\frac{0.142857 + 0.111111 + 0.090909}{3}}$$

$$= \frac{1}{\frac{0.344877}{3}} = \frac{1}{0.114959} = 8.7$$

The values of the reciprocals are, of course, discovered from tables of reciprocals. It is not necessary to compute them each time.

We note that the harmonic mean (which we can symbolize as M_h) of the numbers 7, 9, and 11 is 8.7; the geometric mean we have found to be 8.849, and the arithmetic mean is 9. If we take the second example which we used with the geometric mean, and compute the harmonic mean of the numbers 12, 17, 33, 21, and 162, we find the following:

$$\frac{1}{M_h} = \frac{1/12 + 1/17 + 1/33 + 1/21 + 1/162}{5}$$

$$= \frac{0.2262516}{5} = 0.0452503$$

$$M_h = \frac{1}{0.0452503} = 22.1$$

If we again compare the results obtained by the three methods, we find

Arithmetic mean = 49.0
Geometric mean = 29.65
Harmonic mean = 22.1

Experiment will show that whenever we average a group of values the arithmetic mean will be larger than the geometric mean, and the latter will be larger than the harmonic mean (unless all the values averaged are of the same size, in which case the three averages will be identical).[1]

It is somewhat easier to compute the harmonic mean by a method other than that so far used. We have seen that the harmonic mean is based on the arithmetic mean of the reciprocals of numbers, and it was to show this that we used the method heretofore presented. But note that

$$\frac{1}{\frac{\Sigma(1/X)}{N}} = \frac{N}{\Sigma(1/X)}$$

so that in practice we divide the number of items by the sum of their reciprocals. To compute the harmonic mean of our last example again by the shorter method, we have

$$M_h = \frac{N}{\Sigma(1/X)} = \frac{5}{0.2262516} = 22.1$$

It is impossible to compute the harmonic mean of any set of numbers if one or more of these numbers is zero, since division by zero is not allowed in mathematics.

Not only is it true that the value of the geometric mean of any set of numbers always lies between their arithmetic and their harmonic means,

[1] For proof of the fact that these inequalities will persist except when the items averaged are identical in size, see Davis and Nelson, "Elements of Statistics," pp. 96*ff.*, Principia Press, Bloomington, Indiana, 1935.

but in the special case where we are dealing with two numbers we can show that the geometric mean of the two numbers is also the geometric mean of their arithmetic and harmonic means. Suppose we let the two numbers be represented by x and y. Then their arithmetic mean is $(x + y)/2$, their geometric mean is $\sqrt{xy}$, and their harmonic mean is

$$\frac{2}{\frac{1}{x}+\frac{1}{y}} = \frac{2}{\frac{x+y}{xy}} = \frac{2xy}{x+y}$$

Using these formulas, we notice that the geometric mean of the arithmetic and the harmonic means is

$$\sqrt{\left(\frac{x+y}{2}\right)\left(\frac{2xy}{x+y}\right)} = \sqrt{xy}$$

But we have just seen that this is the geometric mean of the two original numbers; so it is evident that for this particular case (where there are but

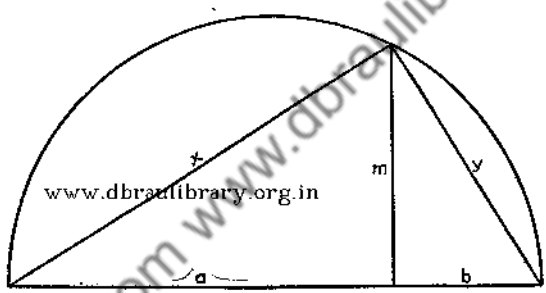


FIG. 4.1.—Relationship of the arithmetic and geometric means.

two numbers involved) the geometric mean of the two numbers is identical with the geometric mean of their arithmetic and harmonic means.

Perhaps the relationship between the sizes of the arithmetic mean and the geometric mean of two numbers can be most easily visualized by means of the diagram in Fig. 4.1. Here we have a semicircle, with its diameter cut into two sections a and b by the perpendicular m. From the point where the perpendicular cuts the arc, lines x and y are drawn to the ends of the diameter. Since the lines x and y form an angle which is inscribed in a semicircle, the angle between them is a right angle.

The arithmetic mean of the lengths a and b is $(a + b)/2$. This is the radius of the circle. Obviously, then, the arithmetic mean of a and b is constant no matter where the perpendicular m is erected. But we note in the diagram that

$$x^2 + y^2 = (a + b)^2 = a^2 + 2ab + b^2$$

But

$$x^2 = a^2 + m^2$$
$$y^2 = b^2 + m^2$$

Therefore

$$a^2 + m^2 + b^2 + m^2 = a^2 + 2ab + b^2$$
$$2m^2 = 2ab$$
$$m = \sqrt{ab}$$

But $\sqrt{ab}$ is the geometric mean of the lengths of a and b. We see that the perpendicular is the geometric mean of the two segments of the diameter, while their arithmetic mean is the radius of the circle. When the perpendicular is raised at the center of the diameter, $a = b = m$; and we have the limiting case in which the two original values are equal and the arithmetic mean equals the geometric mean. But whenever the perpendicular is erected at any point other than at the center of the diameter, the perpendicular will be shorter than the radius, and the geometric mean will be smaller than the arithmetic mean.

4.8. The Quadratic Mean: Ungrouped Data.—The quadratic mean of a group of numbers is found by squaring the numbers, finding the arithmetic average of the squares, and taking the square root of the result. We can illustrate again with the three numbers 7, 9, and 11. The squares of the numbers are 49, 81, and 121. The sum of these squares is 251, and their arithmetic average is 83.66. The square root of 83.66 is 9.15. If we take the other set of numbers with which we have illustrated our earlier averages, 12, 17, 33, 21, and 162, we proceed to find their quadratic mean as follows:

X	X^2
12	144
17	289
33	1,089
21	441
162	26,244
	$\Sigma X^2 = 28{,}207$

$$\frac{\Sigma X^2}{N} = \frac{28{,}207}{5} = 5{,}641.6$$
$$M_q = \sqrt{5{,}641.6} = 75.1$$

If we represent the quadratic mean by the symbol M_q, we can describe these calculations by the following formula:

$$M_q = \sqrt{\frac{\Sigma X^2}{N}}$$

If we bring together the four averages which we have so far

computed, all based on the numbers 12, 17, 33, 21, and 162, we find the following values:

$$M_q = 75.1$$
$$X = 49.0$$
$$M_g = 29.6$$
$$M_h = 22.1$$

It may seem to the student that only one of these can be correct, and that the other three must be in error. When, moreover, it is pointed out that one always[1] finds this same general sort of thing —the quadratic mean largest, followed by the arithmetic mean, the geometric mean, and the harmonic mean the smallest—the question arises why one ever computes such peculiar averages. Suffice it to say at this point that sometimes one of these averages is "correct" and sometimes another, depending on what the figures represent and in what way we wish to typify them. Just as we saw (Sec. 4.3) that one sometimes weights an arithmetic mean because such a procedure does actually give him the right answer, so we shall see that sometimes one uses the harmonic, the geometric, or the quadratic mean because it gives the right answer. The discussion as to which kind of average to use under which circumstances is found toward the close of the following chapter (see Sec. 5.22).

4.9. Quartiles, Deciles, and Percentiles: Ungrouped Data.— The median is sometimes called an "average of position"; that is, it is defined as the value of an item which holds a certain position in the array. It is, we have seen, the item which is so located that it divides the array into two parts, there being the same number of items in each part. We could, of course, find the two points which divide the array into three parts or the seven points which divide the array into eight parts. In fact we do often wish to find the points which divide the array into 4, 10, or 100 parts.

The three points which divide the array into four parts in such a way that each part contains the same number of items are called the *quartiles*. Just as we found that the median item could be found by counting $(N + 1)/2$ items from either end, so the quartiles can be found by counting $(N + 1)/4$ items from

[1] Except in the limiting case where all the original values are equal. In such a case, all four averages will be equal also.

each end. If we revert to the case we used in illustrating the median—the examination marks which were listed on page 24 and arranged in an array on page 66—we discover that there are 90 marks. Hence the position of the first quartile (the first quartile is always the smallest of the quartiles and the third quartile the largest, with the second quartile between) will be $(N + 1)/4$ or $(90 + 1)/4$ or $9\frac{1}{4}$ or 22.75 items from the bottom. If we count up 22 items, we arrive at the value of 88. The 23d item is 89. Hence a point $\frac{3}{4}$ of the way between them will be 88.75, and we say that the first quartile (symbolized by Q_1) is 88.75.

Now let us count down from the top 22 items. This brings us to a value of 150. The 23d item has a value of 149. If we locate the value which is $\frac{3}{4}$ of the way from 150 to 149, we get 149.25. Hence we say that the third quartile (Q_3) is 149.25. The second quartile must obviously be at the center of the array; that is, it is identical with the median. Hence we never speak of the second quartile, but say that the two quartiles and the median divide the array into four parts in such a way that each part contains the same number of items as each other. To summarize our results for the array of examination marks, we could say

$$Q_1 = 88.75$$
$$\text{Med.} = 112 \text{ (see page 67)}$$
$$Q_3 = 149.25$$

These three values are so chosen that they divide the array as required.

We can give the formulas for the positions of the quartiles, then, as follows:

$$Q_1 = \frac{(N + 1)}{4}$$
$$Q_2 = \frac{2(N + 1)}{4} = \frac{(N + 1)}{2} = \text{Med.}$$
$$Q_3 = \frac{3(N + 1)}{4}$$

This latter value for Q_3 shows how far up we would have to count from the bottom in order to reach the third quartile. In our illustration we counted down $(N + 1)/4$ items from the top. Experiment will show that either method gives the same result.

The *deciles* are the nine points which so divide the array that each part contains the same number of cases as each other part. In this case, as the name implies, the array is divided into 10 groups. The formulas for the positions of the deciles, starting with the first (smallest) decile (D_1) follow:

$$D_1 = \frac{(N+1)}{10}$$

$$D_2 = \frac{2(N+1)}{10}$$

$$D_3 = \frac{3(N+1)}{10}$$

etc.

If we compute the first two deciles from the data on examination marks used before (page 24), we find

$$D_1 = \frac{(N+1)}{10} = \frac{91}{10} = 9.1$$

9th item = 75
10th item = 76
1/10 of the way from 9th to 10th = 75.1

Hence

$D_1 = 75.1$

$$D_2 = \frac{2(N+1)}{10} = \frac{182}{10} = 18.2$$

18th item = 84
19th item = 85
2/10 of the way from 18th to 19th = 84.2

And the last (9th) decile would be

$$D_9 = \frac{9(N+1)}{10} = \frac{819}{10} = 81.9$$

81st item = 164
82d item = 165
9/10 of the way from 81st to 82d = 164.9

Note that in each case here the items have been found to be one unit apart. Suppose, in the last illustration, that the 81st item had been 164 and the 82d item had been 168. The point 9/10 of the way between the two would be at 167.6. This point is discovered as follows: The entire distance between 164 and 168 is 4.

Nine-tenths of 4 is 3.6. Since we are going from the value 164 toward the value 168, we add the 3.6 to the 164, getting 167.6, which would be the ninth decile under such circumstances.

The 99 points which divide the array into 100 parts in such a way that the parts contain equal numbers of items are called the *percentiles*. As the student would anticipate from what has gone before, the formulas for the position of the percentiles (where P_1 is the first percentile, P_2 the second percentile, etc.) are

$$P_1 = \frac{(N+1)}{100}$$

$$P_2 = \frac{2(N+1)}{100}$$

$$P_3 = \frac{3(N+1)}{100}$$

and so on, until we reach

$$P_{99} = \frac{99(N+1)}{100}$$

If we take but one example in this case, using the examination marks again for purposes of illustration and computing the value of the 19th percentile, we find

$$P_{19} = \frac{19(N+1)}{100} = \frac{19(91)}{100} = \frac{1729}{100} = 17.29$$

17th item is 82
18th item is 84
29/100 of the way from 82 to 84 is 82.58

Thus

19th percentile is 82.58

4.10. Use of Quartiles, Deciles, Etc.—In the preceding chapter (see Sec. 3.5) it was pointed out that there are usually advantages in using uniform class intervals throughout a frequency table, although we saw in Sec. 3.18 that there are cases where it is worth while to make exceptions. Now we notice that, while a frequency table usually keeps the class interval constant and has varying frequencies in the classes, a set of quartiles or deciles or percentiles amounts to the same thing as keeping the frequencies of the classes constant and varying the class interval. For example, if we illustrate again with the 90 examination marks which

appear in array on page 66, we find the following nine deciles:

75.1; 84.2; 93.3; 104.4; 112; 126;
142; 153.8; 164.9

These points are not equally spaced. The first amounts to an open-end class including all cases below 75.1. The next is a class running from 75.1 to 84.2, with a class interval of 9.1. The next class, running from 84.2 to 93.3, also has a class interval of 9.1. The next class runs from 93.3 to 104.4, with a class interval of 11.1. The other class intervals are 7.6, 14, 16, 11.8, and 11.1. Then there is the final class running 164.9 and over. While these classes have unequal class intervals (and in a mound-shaped distribution the class intervals will ordinarily be smaller toward the center of the distribution; the student should make sure that he sees why this is true) they contain equal numbers of cases. In our illustrative problem, each class contains 9 marks. In fact, that is just how we drew them up. We defined the percentiles, for example, as the 99 points which divided the distribution in 100 parts in such a way that there were equal numbers of cases in the various parts. We see, then, that here is another case where, in practice, one occasionally wishes to get away from the equal class intervals which are so useful when we are expecting to carry on further computation.

4.11. Summary of Averages with Ungrouped Data.—If each item is stated separately, rather than being grouped with others in a frequency table, we compute the various measures of central tendency (also called averages, or types) as follows:

1. The arithmetic mean ($\bar{X}$), also called the arithmetic average or the mean:
   Add the numbers given.
   Divide the sum by the number of cases.
   Formula:

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

2. The weighted arithmetic mean.
   Assign a weight to each number.
   Multiply each number by its weight.
   Add the products just obtained.
   Divide the sum of the products by the sum of the weights.
   Formula:

$$\bar{X} = \frac{\Sigma(XW)}{\Sigma W}$$

3. The median (Med.):
   Arrange the data in array.
   Count $(N + 1)/2$ items from either end.
   The value of this item is the median.
4. The mode (Mo.):
   Count the number of times that each value occurs.
   The value occurring most frequently (if any) is the mode.
5. The geometric mean (M_g):
   Find the logarithms of the values.
   Add these logarithms.
   Divide the sum by the number of cases.
   Take the antilogarithm of the quotient.
   Formula:

$$\log (M_g) = \frac{\Sigma(\log X)}{N}$$

6. The harmonic mean (M_h):
   Find the reciprocals of the numbers.
   Add the reciprocals.
   Divide the number of cases by the sum of the reciprocals.
   Formula:

$$M_h = \frac{N}{\sum\left(\frac{1}{X}\right)}$$

7. The quadratic mean (M_q):
   Find the square of each of the original numbers.
   Add the squares.
   Divide this sum by the number of cases.
   Take the square root of the quotient.
   Formula:

$$M_q = \sqrt{\frac{\Sigma X^2}{N}}$$

8. The quartiles (Q_1, Q_3):
   Arrange the data in an array.
   Count $(N + 1)/4$ items from the lower end.
   The value of the item located here is Q_1.
   Count $3(N + 1)/4$ items from the lower end.
   The value of this item is Q_3.
   Q_2 is the median.
9. The deciles and percentiles (D_1, D_2, etc.; P_1, P_2, etc.):
   Arrange the data in an array.
   For the deciles find the value of the items which are multiples of $(N + 1)/10$ from the end.
   For the percentiles find the values of the items which are multiples of $(N + 1)/100$ from the end.

Having defined our terms, and having seen how these averages are computed when each of our original figures is given, we shall

turn our attention in the following chapter to the methods used for finding these averages when the data are grouped together in frequency tables.

EXERCISES

1. Find the deciles of the data given in Exercise 6 at the end of Chap. III (see page 58).

2. Find the 89th percentile of the data of Exercise 6, page 58.

3. Company A buys electricity at 3 cents per kilowatt-hour, Company B at 2 cents, and Company C at 5 cents. Company A uses 10,000 kw.-hr., Company B uses 8,000, and Company C uses 20,000. What was the average cost per kilowatt-hour? Use a weighted average. Explain why you use the particular weights you do, instead, for example, of using the capitalizations of the companies or the numbers of their employees as weights.

4. Company A pays its employees an average wage of \$28 per week. Company B pays an average of \$35 per week. What figures would you need for weights before you could find the average weekly wages of the employees of both companies combined?

5. If you were given the wheat yield (bushels per acre) in each of the 48 states of the United States, and you wanted to compute the yield (bushels per acre) for the United States, why would you have to weight the average of the 48 yields, and what figures would you use to weight them with?

6. Find the quadratic mean, the arithmetic mean, the geometric mean, and the harmonic mean of the numbers 40 and 10.

7. Show in the preceding example that the geometric mean of the two numbers is also the geometric mean of their harmonic and arithmetic means.

8. If you knew the batting average of each member of a baseball team, and wanted the team's batting average, what additional information would you need? Under what circumstances would you get the correct answer if you took the simple arithmetic average of the figures for the various members of the club?

CHAPTER V

MEASURES OF CENTRAL TENDENCY (*Continued*)

5.1. Averages from Grouped Data.—We discovered in Chap. III that the statistician seldom retains his figures in their original form, since there are too many of them to be handled easily, and since the large number of figures tends to be confusing rather than enlightening. In order to compress the data within reasonable limits, and to make it possible to get an idea of the general nature of the distribution at a glance, he ordinarily classifies the data in a frequency table, showing merely the numbers of cases which fall in various classes.

It would at first seem that when we have the data so arranged it would be impossible to subject them to further statistical manipulation. How can we find the averages of the data? How can we compute the value of the arithmetic mean, the median, the mode, or any of the other summary figures that we studied in the preceding chapter?

It would evidently be a foolish waste of time for the statistician to classify his data in frequency tables if thereafter he could carry on no further computations. In this chapter, we shall see how it is possible to compute the various averages of figures even if they have been grouped or classified in a frequency table. We shall then see how and when each average should be used, and how it should be interpreted.

As for the computations themselves, we can understand the problem most easily in connection with an illustrative example. Castle gives figures (see Table 5.1)[1] showing the heights of 1000 Harvard students between the ages of 18 and 25, measured at the Harvard gymnasium in the years 1914–1916.

We have already seen that in such cases we do not know the value of a single item. We do not know the exact height of a

[1] W. E. Castle, "Genetics and Eugenics," Harvard University Press, Cambridge, Mass., 1916. By permission of the president and fellows of Harvard College. (Data are adapted from data on p. 61 of this book.)

single student out of this group of 1000. To be sure, we can tell something about the distribution of heights. We know that no student was shorter than 154.5 cm. and that none was taller than 199.5 cm. We know that most of them were between 170 and 180 cm. tall. But whether one student, or 15, or none, had a height of 168.3 cm. we cannot tell. How, then, can we tell anything about the average height, since we do not know the heights of any of the individuals? How, when our items have

TABLE 5.1.—HEIGHTS OF 1000 HARVARD STUDENTS, AGES 18 TO 25

Height (centimeters)	Number of Students
155–157	4
158–160	8
161–163	26
164–166	53
167–169	89
170–172	146
173–175	188
176–178	181
179–181	125
182–184	92
185–187	60
188–190	22
191–193	4
194–196	1
197–199	1
Total.....	1000

lost their identity in a frequency table, can we add them or multiply them together or arrange them in order of magnitude? How can we perform any of the operations which we have needed to perform in order that we may compute the various measures of central tendency? As a matter of fact, we can do these things only if we make certain assumptions. We must now discuss these assumptions and see how they enable us to compute averages from grouped data. For the statistician computes averages from such data quite as often as from ungrouped data, and as a matter of fact he usually prefers to do so, because grouped data save him time and cause little inaccuracy.

Since we do not know the exact height of a single one of the 53 students whose heights are recorded as falling between 164 and 166 cm., we must assume something about the heights. We might assume that the heights were evenly distributed over the 3-cm. range from 163.5 to 166.5, no two of the men being the same height and the differences in their heights being equal. We might assume that these 53 students were all of exactly the same height, in which case we should be likely to assume that they were all located at the middle of the class interval, or at 165 cm. Or we might make other assumptions that seemed reasonable. No matter what assumption we make we shall be likely to be somewhat in error, but we can surely choose a value for these heights that will not be in error for any one of the 53 men by more than 1.5 cm.; that is, if we choose to assume that the students are all the same height (165 cm.), the error will not be large in any individual case.

5.2. The Arithmetic Mean: Grouped Data.—When we compute the arithmetic mean of data which are grouped in frequency tables, we usually assume that the total of the values in any class is just what it would be if all the items were located at the mid-point of the class. This is the same as assuming that the items in the class are evenly distributed throughout the class: either assumption would give the same total. In the case that we have been using as an illustration, it is assumed that the total height of the 53 men in the group whose heights vary from 164 to 166 cm. is the same as the total height of 53 men who are each 165 cm. tall. In other words, their total height will be $53 \times 165 = 8745$ cm. Of course, if these 53 men were evenly distributed over the range from 163.5 to 166.5 cm., no two men being of the same height, the average height would still be 165 cm., and the total height would still be 8745 cm., as we discovered on the other assumption. Hence it makes no difference in this case whether we assume that the data are concentrated at the mid-points of classes or are evenly distributed throughout the classes. In computing the mean we shall make the former assumption, as involving less arithmetic.

If we assume that the 1000 students are located at several points, these being the class mid-points (or *class marks*, as they are sometimes called), then we can easily determine the average height. We say that the four shortest students have each a

height of 156 cm., or a total height of 624 cm. The next eight students, being concentrated at the height of 159 cm., have a total height of 1272 cm. If we continue thus throughout the table, multiplying in each case the class mark by the number in the class, we shall determine the total height of the students in each class. If, then, we add these products, we shall get the total height of the 1000 students. And we have already discovered that the average height is the total height divided by the number of cases. Thus we obtain our average easily once it is assumed that the heights are concentrated at the class marks.

Table 5.2 illustrates the process of finding the arithmetic mean from frequency data. In the first column the class limits are

TABLE 5.2.—COMPUTATION OF ARITHMETIC MEAN FROM FREQUENCY DISTRIBUTION (LONG METHOD)

Height (centimeters)	Class Mark (X)	Number of Students (f)	Total Height (fX)
155–157	156	4	624
158–160	159	8	1,272
161–163	162	26	4,212
164–166	165	53	8,745
167–169	168	89	14,952
170–172	171	146	24,966
173–175	174	188	32,712
176–178	177	181	32,037
179–181	180	125	22,500
182–184	183	92	16,836
185–187	186	60	11,160
188–190	189	22	4,158
191–193	192	4	768
194–196	195	1	195
197–199	198	1	198
Totals.....		1000	175,335

given as they appear in the original. But we assume that all the items within any class are located at the class mark, or mid-point, which is shown in the second column. We assume that any height between 154.5 and 155.5 was recorded as 155. Hence our first class presumably includes heights starting at 154.5; likewise, at its upper end, it presumably includes heights up to 157.5. This is a range of 3 cm., and we would find the class mid-point by

adding half this range (1.5 cm.) to the lowest limit of the range (154.5 cm.). Thus the class mark would be 154.5 + 1.5 = 156 cm. And since we assume that all four students in the class were concentrated at this value, it must be that our X's for these four students (that is, the original values with which we start our problem) are 156. Similarly the table shows 8 students with heights of 159 cm., 26 students with heights of 162 cm., etc. The number of students in each class of the frequency table we indicate by the letter f, which always represents the frequency with which items occur in a class of a frequency table. It is easy to remember that f stands for frequency.

Finally, in the last column, we have the total height of the people in each group. If each of four people measures 156 cm. in height, their total height is $4 \times 156 = 624$ cm. Similarly throughout the table we have multiplied each class mark by the class frequency (the number of items in the class) to get the total height of the people in the class. Hence we label the final column fX, since it is found by multiplying the X values by the f values.

If we summate the last column we find the total height of the 1000 students to be 175,335 cm. If 1000 students have a total height of 175,335 cm., then the average height is 175,335/1000 = 175.335 cm. But since our original figures are given to the nearest even centimeter only, we should not give the average to three decimals (even though we should probably not be far in error by doing so).[1] Therefore we round off our result to even centimeters, making it 175 cm.[2]

We can summarize the directions for computing the mean from frequency distributions in a formula, as follows:

$$\bar{X} = \frac{\Sigma(fX)}{\Sigma f} = \frac{\Sigma(fX)}{N}$$

This formula says, "Multiply each X by the corresponding f, and add the products. Divide the sum of the products by the sum of

[1] See RAYMOND PEARL, "Medical Biometry and Statistics," 2d ed., pp. 362*ff*., W. B. Saunders Company, Philadelphia, 1930.

[2] In Castle, *op. cit.*, from which these data are extracted, the average height is given as 174.4 cm. This is contrasted with our 175.3 cm. If one were to assume that the class intervals as given mean 155–157.9, 158–160.9, etc., the average would, of course, be even higher. I have not discovered the cause of the discrepancy.

the frequencies (which is, of course, the total number of cases, or N)."

If we compare this formula for the arithmetic mean of items in a frequency table with our formula for the weighted arithmetic mean (see page 62), we discover that they are similar save for one substitution. It we write the two formulas side by side this will be immediately apparent.

$$\frac{\Sigma(XW)}{\Sigma W} \qquad \frac{\Sigma(fX)}{\Sigma f}$$

We see that the formula for use in frequency tables is a duplicate of the formula for the weighted arithmetic mean except that we have substituted the symbol f for the symbol W. Obviously, then, when we find the arithmetic average of numbers classified in a frequency table we have really computed a weighted arithmetic mean, using the frequencies as weights.

5.3. Arithmetic Mean: Short Method.—The method we have just used for determining the average of data grouped in a frequency table is not the shortest possible method. In fact, it is not a method which would be used in practice. We have presented it merely to show that it is possible to compute the mean from grouped data if we make the proper assumptions. Any statistician who wished to compute such a mean would always use what is called "the short method." With this method we start by guessing at the mean and then adjusting the guess to meet the facts. This method is easiest to understand in connection with an illustration, and for this we shall use the data on student heights which appear in Table 5.1.

In Table 5.3 the class limits are listed in the first column. Since the first class contains those items which vary in size from 154.5 to 157.5 cm., we take the mid-point of the class as 156 cm. and list it in the second column. In the same way the other class marks are determined. Then comes the first step that is new. We look over the data and guess at the mean, choosing one of the class marks as the guessed mean. In this case we chose 174 as the provisional or guessed mean. We then set down the number of steps by which each class differs from the mean, and these deviations we list under the heading "class deviations," and we symbolize them by the letter d. Thus the first class (155–157) is six classes smaller than the guessed mean, so we label it -6.

The class labeled "194–196" is seven classes larger than the guessed mean, so we label it +7. In this way we state each class in terms of the difference from the mean, measuring our differences in units of the class interval.

The fourth column is one with which we are already familiar. Here appear the frequencies as before. The last column is the product of the class deviations and the corresponding frequencies (fd).

TABLE 5.3.—COMPUTATION OF ARITHMETIC MEAN FROM FREQUENCY DISTRIBUTION (SHORT METHOD)

Height (centimeters)	Class Mark (X)	Class Deviation (d)	Frequency (f)	(fd)
155–157	156	−6	4	− 24
158–160	159	−5	8	− 40
161–163	162	−4	26	−104
164–166	165	−3	53	−159
167–169	168	−2	89	−178
170–172	171	−1	146	−146
173–175	174	0	188	0
176–178	177	+1	181	181
179–181	180	+2	125	250
182–184	183	+3	92	276
185–187	186	+4	60	240
188–190	189	+5	22	110
191–193	192	+6	4	24
194–196	195	+7	1	7
197–199	198	+8	1	8
Totals.....			1000	+445

Now if we total the last column we find $\Sigma(fd) = +445$, and this divided by Σf (or N) = $^{445}/_{1000} = 0.445$. This tells us that the true average of these data is 0.445 class intervals above the guessed mean. (Had the sign of Σfd been minus, the true average would have been below the guessed average.) Now the class interval is 3 cm., and 0.445 class intervals make 1.335 cm. The guessed average was 174 cm., and, if we add thereto the correction of 1.335 cm. which we have just found, we get 175.335 cm. as the average. This is exactly the same as the average which we

found when we carried on the computations by the long method, as will be seen by reference to page 85.

What we have actually done by this short method is to assume a mean and to find on the average how far the items fall from this assumed mean. Our unit of measure is the class interval, which in this case is 3 cm. We know that if the mean is correctly chosen the sum of the positive deviations will exactly offset the sum of the negative deviations, so that, when, as in this case, the positive deviations are larger than the negative, the assumed mean is too small and must be raised enough so that positive and negative deviations will balance. If the negative deviations exceed the positive we must, on the other hand, lower the average. Our process consists in finding out by how many units (class intervals) we must adjust the assumed mean to make it coincide with the true mean.[1]

Any point may be chosen as the assumed mean, although the work involved is much less if the assumed mean is at the mid-point of one of the classes. Also it helps somewhat if the class chosen is near the middle of the distribution, so that the numbers used are as small as possible.

The so-called "short process" of computing the mean seems like a long process when described in such detail. If the student will compute a mean from a frequency distribution, using first the long and then the short method, and timing the process, he will discover that the short method is correctly named. We may summarize the steps of the short method as follows:

1. List the class marks.
2. Locate a guessed mean near the middle of the distribution and at a class mark.

[1] That the algebraic sum of the deviations from the arithmetic mean must equal zero is shown by the following:

Each deviation from the mean may be defined as follows:

$$x = X - \bar{X}$$

The sum of the deviations is, then,

$$\Sigma x = \Sigma(X - \bar{X}) = \Sigma X - N(\bar{X})$$

But since

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

$$\Sigma X = N(\bar{X})$$

It is therefore obvious that $\Sigma X - N(\bar{X}) = 0$ and that $\Sigma x = 0$.

3. State the other classes in terms of class deviations from the guessed mean. The deviations are plus and minus, and the signs are important. Values lower than the guessed mean are minus; others are plus. The class containing the guessed mean is marked "0."

4. List the frequencies of the classes.

5. Multiply each frequency by its class deviation, keeping the plus and minus signs.

6. Add the products obtained in the preceding step.

7. Divide the sum just obtained by the sum of the frequencies (that is, by N) and multiply the quotient thus obtained by the class interval.

8. Add the result obtained in the preceding step to the guessed mean of step 2.

If we are to boil these directions down into a convenient formula which gives directions for computing the mean from grouped data, we shall need some new symbols. Let d represent the distance (measured in units of the class interval) from the assumed mean. Let $\bar{X}'$ represent the assumed mean (as distinct from the real mean, which is represented by $\bar{X}$). Let Ci represent the class interval, which in the illustration was 3 cm. Then our formula is as follows:[1]

$$\bar{X} = \bar{X}' + Ci\left(\frac{\Sigma fd}{\Sigma f}\right) = \bar{X}' + Ci\,\frac{\Sigma(fd)}{N}$$

In our case this becomes

$$\bar{X} = 174 + 3\left(\frac{+445}{1000}\right) = 175.335$$

[1] On p. 85 we defined the mean of a frequency distribution thus:

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

But each actual value of X is equal to the guessed mean plus (or minus) an amount equal to the number of class deviations times the class interval. That is,

$$X = \bar{X}' + (Ci)(d)$$

Substituting this in the formula above gives

$$\bar{X} = \frac{\Sigma\{f[\bar{X}' + (Ci)(d)]\}}{N}$$

$$= \frac{\Sigma[f(Ci)(d)]}{N} + \frac{\Sigma[f(\bar{X}')]}{N}$$

$$= Ci\,\frac{\Sigma(fd)}{N} + \frac{\bar{X}'(\Sigma f)}{N}$$

Since, however, $\Sigma f = N$, this becomes

$$\bar{X} = \bar{X}' + Ci\,\frac{\Sigma(fd)}{N}$$

As before, we should round off the answer to 175 cm. to correspond with the accuracy of the original figures.

5.4. Checking Accuracy of Computations.—Whenever a good statistician gets an answer to any problem, he immediately checks it to see if it is reasonable. Suppose, for example, that we had found in our last example an average height of 487 cm. We know by looking at our original table that this is out of the question, since no student had a height greater than 199.5 cm. Yet it amazes teachers of statistics year after year to have students turn in answers on examinations which are as obviously wrong as this one. It is a good plan to study your data before you start your computations, estimating roughly what answer should be expected. Then if the computations give an answer which differs widely from that expected, one should question the accuracy of his work.

TABLE 5.4.—COMPUTATION OF ARITHMETIC MEAN FROM FREQUENCY DISTRIBUTION (SHORT METHOD) WITH CHARLIER CHECK

Height (centimeters)	Class Mark (X)	Class Deviation (d)	Frequency (f)	(fd)	$f(d+1)$
155–157	156	−6	4	− 24	− 20
158–160	159	−5	8	− 40	− 32
161–163	162	−4	26	−104	− 78
164–166	165	−3	53	−159	− 106
167–169	168	−2	89	−178	− 89
170–172	171	−1	146	−146	0
173–175	174	0	188	0	188
176–178	177	+1	181	181	362
179–181	180	+2	125	250	375
182–184	183	+3	92	276	368
185–187	186	+4	60	240	300
188–190	189	+5	22	110	132
191–193	192	+6	4	24	28
194–196	195	+7	1	7	8
197–199	198	+8	1	8	9
Totals.....			1000	+445	+1445

Fortunately with some statistical computations it is possible to check the accuracy of the work as one proceeds, so that at the end of any step in the process he can tell whether or not errors have been made. When such checks are possible, it is wise for the student to get in the habit of using them, since they take little time and effort at worst, and at best may save hours of rechecking and recalculating. In the case of the arithmetic mean computed from a frequency table, it is possible to apply what is known as the "Charlier check" to prove the accuracy (or demonstrate the inaccuracy) of our arithmetic.

To compute the Charlier check for the arithmetic mean, we merely add one column to our table by showing the values of $f(d + 1)$. This gives us Table 5.4, which the student should compare with Table 5.3. It will be seen at once how the figures in the last column are derived.

The first figure in the new last column is -20. This is found by adding 1 to the value of d (which gives us $-6 + 1 = -5$) and multiplying by the value of f (which is 4). Each figure in the last column is found similarly, by adding 1 to the value of d, and multiplying by the corresponding value of f. To take one more case, the fourth from the last item in the column is the number 132. It was found by adding 1 to the value of d (which was 5) to get the value of $d + 1$, or 6. This value was multiplied by the corresponding value of f (22) to get the value 132 in the last column.

When this check is applied we find that, if our arithmetic has been correct, the sum of the last column, $\Sigma [f(d + 1)]$, will always be equal to the sum of the totals of the two preceding columns. In the case of Table 5.4, we note that the last column yields a total of 1445, which is the sum of the two preceding totals, 1000 and 445. Sometimes the sums of one or two of the columns are negative, and the student must be careful to keep track of signs. For example, it might be that the sum of the frequencies would be 865, the sum of the values of fd might be -148, and in that case the sum of the values of $f(d + 1)$ should be $865 + (-148) = 717$. But if the total of the last column is not equal to the algebraic sum of the other two totals, some mistake in arithmetic has been made.[1]

5.5. Grouping Error with the Arithmetic Mean.—In computing the arithmetic mean from a frequency table, we have assumed that all the items in any particular class are concentrated at the mid-point of the class. Of course, our results would be the same if no two of the items in the group were the same size, but if they were arranged at equal intervals throughout the class from the lower to the upper class limit. Likewise our assumption would bring no inaccuracy no matter how irregularly the items were scattered in the class if the average of the items within each class were equal to the class mark. Therefore we can say that any one of three assumptions would give us the same results, namely:

1. All items within a class are the same size, each equal to the class mark.
2. No two items in the class are the same size, and the values are spaced equidistantly throughout the class interval.
3. The average of the items within each class is equal to the class mark.

[1] The student who is mathematically inclined will prefer to have this statement proved.

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

In practice, probably none of these assumptions is exactly true. Yet if in one class the items run a little larger than the class mark we expect by the laws of chance that in some other class they will tend to run a little lower than the class mark, and if we have a large enough number of items and a large enough number of classes we should expect any errors to cancel each other out. We find empirically that the arithmetic mean computed by the methods just described is reasonably accurate.

The student can easily test the accuracy of the method by trying it in cases where the actual average is known. For example, on page 24 appear 90 marks received by students on an examination. The arithmetic mean of these 90 marks is $10{,}635/90 = 118.17$. If we group these data in a frequency table with a class interval of 5 and the lower class limits at 40, 45, 50, etc., we can compute the average from the frequency table, in which case we get an answer of 118.44. Table 3.2 shows the data arranged in a frequency table with a class interval of 10. Here the average is again 118.44. If we make up a frequency table with a class interval of 25, with the lower limits at 25, 50, 75, etc., we find an average of 120.28. Table 3.3 shows these data classified with a class interval of 50. Here the arithmetic mean is 121.11. The student will note that our results from the frequency table fall very close to the actual average of the 90 marks as long as we had a reasonably large number of classes. It can be said in general that the grouping error, which arises when all the cases in a class are treated alike, increases as the class interval increases and is less for continuous than for discrete data.

5.6. The Median: Grouped Data.—If we study the data on heights of students with the intention of determining the median, we find the method very similar to that used when the data were not grouped. The problem is still that of discovering a value such that it will divide the distribution into two groups containing the same numbers of items. As before we start by determining how far from the lower end of the distribution we shall have to go to reach such a value. When the data were not grouped, the median value was number $(N + 1)/2$ from either end. When the data are grouped, the problem differs slightly, and the median item is number $N/2$ from either end. Unless we make this change, we shall get one answer when we start from one end and

another answer when we start from the other end. Since we have 1000 items, we must find the $N/2$ item or the $\frac{1000}{2}$ item or the 500th item; that is, we wish to know the size of the 500th item. If the 1000 students are arranged in an array, we wish to know the height of the student who is number 500 from either end.

Our original distribution of heights was as indicated in Table 5.5. Let us start with the shortest men and count until we have

TABLE 5.5.—HEIGHTS OF 1000 HARVARD STUDENTS, AGES 18 TO 25

Height (centimeters)	Number of Students
155–157	4
158–160	8
161–163	26
164–166	53
167–169	89
170–172	146
173–175	188
176–178	181
179–181	125
182–184	92
185–187	60
188–190	22
191–193	4
194–196	1
197–199	1

reached the 500th man. The lowest class contains 4 men; the two lower classes contain 12 men when taken together; the three lower classes together contain 38 men. If we continue thus, we find that the six lowest classes contain 326 students and the seven lowest classes contain 514 men. If we want the man who is 500th from the bottom, he must be farther along than the top of the sixth class, but not so far along as the top of the seventh class; that is, he must be located somewhere in the seventh class. If we assumed, as before, that all items in the seventh class were at the class mark, it would be an easy matter to locate the median. But now we vary our assumption, assuming that the items are evenly distributed throughout the range of the class; in other words, that the 188 items in the seventh class are considered as being equidistant and as being spread over the entire distance

from 172.5 (the actual lower limit of the class) to 175.5 cm. (the actual upper limit of the class). If we follow this assumption it is relatively easy for us to determine the location of the median. The six lower classes contain 326 students, and we wish to locate the 500th student; that is, we have to go 174 items into the seventh class, or $^{174}/_{188}$ths of the way from the lower limit toward the upper limit of the class. Since the class interval is 3 cm., this means that we have to go up from the actual lower limit of the class a distance of $(^{174}/_{188})$ 3 cm. or 2.78 cm. The actual lower limit is 172.5 cm.; so we have the equation

$$172.5 + 2.78 = 175.28 \text{ cm.} = \text{Med.}$$

Note that this carries us almost to the actual upper limit of the class (175.5 cm.). This is to be expected, since we went up from the bottom of the class through 174 of the 188 items in the class. We can compare the median of 175 (which we get by rounding off the computed value of 175.28) with the arithmetic mean, which we have already discovered to be 175 (see page 87). Before the two values were rounded off they were

$$\bar{X} = 175.335$$
$$\text{Med.} = 175.28$$

It must not be thought, however, that it is necessary for the median and the mean to coincide or to be approximately equal. We saw on page 67 a case where they were decidedly different. That the median and the mean are, in this case, so nearly identical is due to the fact that the heights of the students were so symmetrically distributed. This can be seen roughly from the chart on page 95, which shows the heights of the students; but in the main the question of symmetry of curves will be postponed until later.

Let us now summarize the methods used in finding the median from data grouped in frequency tables. The steps are these:

1. Compute $N/2$ to discover the location of the item desired.[1]

[1] The student should apply the method just illustrated for finding the median to the same distribution of heights, counting from the top instead of from the bottom. He will discover that the use of $N/2$ will give the same answer regardless of the end from which he starts, but the use of $(N + 1)/2$ will not. This is the reason for using $N/2$ here instead of the value used when the data are not grouped.

2. Add the class frequencies until the class containing the median item is discovered.

3. Find how many items one must count into this class to reach the median item.

4. Find what fraction this is of the total number of items in the class.

5. Add this fraction of the class interval to the actual lower limit of the class.

When we are dealing with frequency curves, as in this chapter, we can well redefine the median as that particular value on the

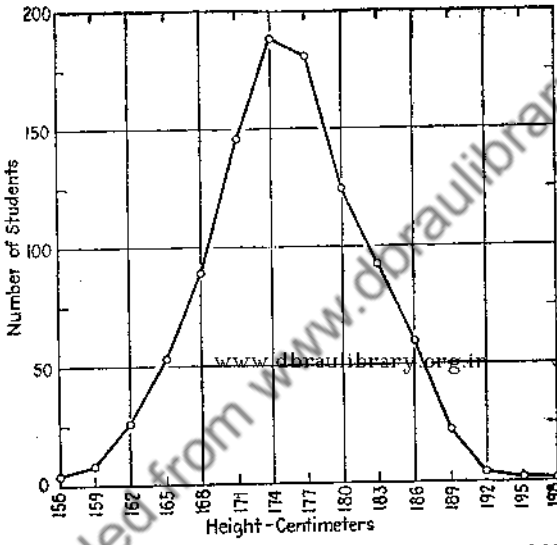


FIG. 5.1.—Number of Harvard students between the ages of 18 and 25 years with various heights, 1914–1916.

base scale of a frequency polygon from which a perpendicular will divide the area under the curve exactly in half. In Fig. 5.1 if we erect a perpendicular from the point on the base scale which corresponds to 175.28 (our value of the median) we shall divide the area under the curve in two, and the two new areas will be equal or approximately so.

We should also note that the definition of the median breaks down in the case of some discrete data. For example, Ernest Thompson Seton counted the number of eggs in each of 77 pelican's nests, and found that 4 nests contained 1 egg each, 65 nests contained 2 eggs each, 5 nests contained 3 eggs each, and 3 nests contained 4 eggs each. In this case there is no

median which meets our definition exactly; that is, there is no number of eggs which is exceeded by as many cases as those which fall short of it. If we pick 2 as the median, we find it exceeded by 8 cases with but 4 cases smaller. If we choose 3 as the median, we find 69 cases smaller and but 3 larger. In such a case the whole idea of the median may break down.

5.7. Finding the Median from an Ogive.—We saw in Sec. 3.11 that every frequency table can be converted in the form of an ogive. This ogive form is often useful for finding the

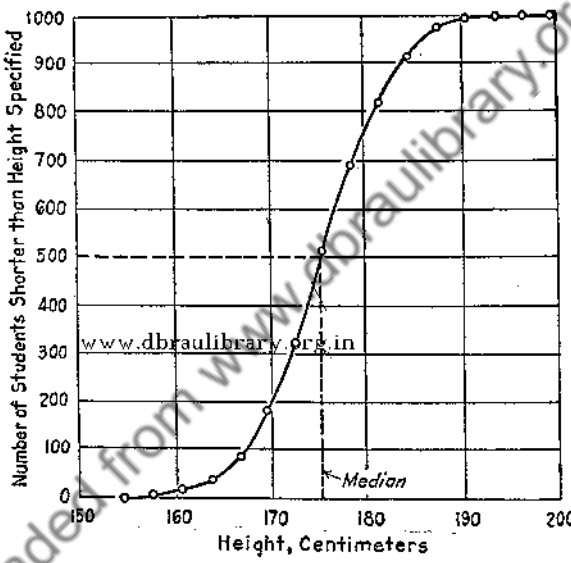


FIG. 5.2.—Determination of the median from an ogive.

approximate values of the "averages of position," such as the median or the quartiles. In Fig. 5.2 the data of Table 5.5 have been put in such form. We notice that the curve takes the peculiar S-shape assumed by mound-shaped frequency curves when they are converted to ogives (see Sec. 3.14). Since there are 1000 cases in this distribution, the median will be the value of the 500th case. If we find the value 500 on the vertical scale, find the point on the ogive curve horizontally opposite it, and then drop a perpendicular from this point on the ogive to the base scale, we can read the approximate value of the median on the base scale. This has been indicated on the diagram by two dotted lines, and we see that the value of the median is approxi-

mately 175. Similarly we could read approximate values for quartiles, deciles, etc. With 1000 cases the third decile would correspond to the 300th case, and we read from Fig. 5.2 a value of approximately 172. The value of the 61st percentile (the 610th case) seems to be approximately 177. Naturally the larger we make our scales, the more accurately we can read our values from the ogive.

5.8. The Mode: Grouped Data.—The mode is the value which occurs most frequently, and it is usually easier to locate when the data are grouped in frequency tables than otherwise. If we look again at Table 5.1, page 82, which shows the distribution of students' heights, we see at once that the most common height is somewhere in the neighborhood of 173 to 178 cm. Sometimes people call the class mark of the most populous class the *mode*. It is better to give this measure the name *crude mode* to distinguish it from the more accurate computed mode.

The mode is the least satisfactory of the measures of central tendency to compute. Some distributions show no mode at all, and other distributions show two or more. Even in those distributions which seem to show one marked spot of concentration, such as the distribution of student heights pictured in Fig. 5.1, we run into the difficulty that different students studying the same data will find very different modes depending on how the data are grouped. One could classify the data on the student heights, for example, into 10 different frequency tables, varying the class interval somewhat and shifting the positions of the class limits even when the class interval remained the same, and the arithmetic average computed from each of these 10 frequency tables would be approximately the same as long as a reasonable number of classes was used. But these 10 different frequency tables, all based on exactly the same original data, would yield significantly different values for the mode were we to compute the mode by any of the commoner and simpler methods usually outlined in textbooks on elementary statistics. For this reason it seems unwise to stress here any of these makeshift and unreliable methods. The best methods will be explained later in Chap. VIII, and here we can point out two methods, either of which the student may use if he is in too much of a hurry to apply the more accurate methods, and if he will remember that they are but makeshifts.

First, in mound-shaped distributions of but moderate skewness there is an approximate relationship between the sizes of the arithmetic mean, the median, and the mode. In distributions which are exactly symmetrical, these three measures coincide, but when distributions are asymmetrical these values differ. Under ordinary circumstances the median will fall between the mean and the mode, and they will be related approximately thus:[1]

$$\text{Mo.} = 3\ \text{Med.} - 2\bar{X}$$

If we have computed the mean and the median, we can substitute them in this formula (if the distribution is only moderately asymmetrical) and discover the value of the mode. We have found the following values of averages of student's heights:

$$\bar{X} = 175.335 \text{ (see page 87)}$$
$$\text{Med.} = 175.28 \text{ (see page 94)}$$

If these values are substituted in our new formula, we have

$$\text{Mo.} = 3(175.28) - 2(175.335) = 525.84 - 350.67 = 175.17$$

The mode found by this method is, then, 175.17. This we would round off to 175.

A second method for finding a rough approximation of the mode involves the use of a formula which is less complicated in use than it appears in print. Let m represent the smallest class mark in the frequency table, let g represent the number of groups or classes in the table, let Ci represent the class interval, and let Mo. represent the mode. Then we have approximately

$$\text{Mo.} = m + \frac{2g(\bar{X} - m) - Ci(g - 1)}{2(g - 1)}$$

[1] This relationship is only approximate. James G. Smith ("Elementary Statistics," p. 119, Henry Holt & Company, New York, 1934) quotes Karl Pearson as having given the exact relationship thus:

$$\text{Mo.} = \bar{X} - \frac{\bar{X} - \text{Med.}}{h}$$

where h is defined as follows:

$$h = 0.3309 - \frac{.0846(\bar{X} - \text{Med.})^2}{\sigma^2 - 9(\bar{X} - \text{Med.})^2}$$

Unless $\sigma = 3(\bar{X} - \text{Med.})$ or more, the last term will be so small that it may be neglected and the formula given above in the text will be accurate enough.

In our problem of student heights (see Table 5.3) the smallest class mark was 156 cm., there were 15 classes in the table, the class interval was 3 cm., and the arithmetic mean was 175.33 cm. Substituting these values in our formula gives

$$\text{Mo.} = 156 + \frac{2(15)(175.33 - 156) - 3(14)}{2(14)}$$

When this is evaluated, we find that Mo. = 175.21 cm. In this particular example, where the skewness is almost negligible, this gives a very accurate value for the mode. It is shown later, on page 206, where more refined methods are used that the mode of this distribution is actually about 175.22 cm.

5.9. The Geometric Mean: Grouped Data.—We saw in the preceding chapter that the geometric mean is usually computed in practice with the aid of logarithms, and a quick review of the work taken up in Sec. 4.6 will show that the logarithm of the geometric mean is the arithmetic average of the logarithms of the original figures. Since the computation of the geometric mean with the aid of logarithms involves the computation of an arithmetic mean, and since we have already learned in Sec. 5.2 how to compute an arithmetic mean of numbers grouped in a frequency table, our present problem really involves little or nothing in the way of new material. Instead of using the class marks, we shall use the logarithms of these class marks; but since, even if the class marks are themselves evenly spaced, their logarithms will not be evenly spaced, we must use the long method of Sec. 5.2 rather than the short method of Sec. 5.3.

Table 5.6 gives the data on student heights, with the class marks from Table 5.2 in the first column and the frequencies from the same table in the second column. In the third column appear five-place common logarithms of the class marks. These, of course, are $\log X$. Our formula becomes

$$\log M_g = \frac{\Sigma f(\log X)}{N}$$

Therefore we multiply each item in column two by the corresponding logarithm in column 3, to get the products which appear in column 4. The sum of column 4 is $\Sigma f(\log X)$, amounting in this particular problem to 2243.56097. We divide this by 1000, the

number of cases, to get, as the logarithm of the geometric mean, 2.24356. This corresponds to a geometric mean of 175.2 cm.

TABLE 5.6.—COMPUTATION OF THE GEOMETRIC MEAN FROM A FREQUENCY TABLE

Class Marks (centimeters) (X)	Number of Students (f)	$\log X$	$f(\log X)$
156	4	2.19312	8.77248
159	8	2.20140	17.61120
162	26	2.20952	57.44752
165	53	2.21748	117.52644
168	89	2.22531	198.05259
171	146	2.23300	326.01800
174	188	2.24055	421.22340
177	181	2.24797	406.88257
180	125	2.25527	281.90875
183	92	2.26245	208.14540
186	60	2.26951	136.17060
189	22	2.27646	50.08212
192	4	2.28330	9.13320
195	1	2.29003	2.29003
198	1	2.29667	2.29667
Totals.....	1000		2243.56097

The geometric mean is, as usual, slightly smaller than the arithmetic mean, but the student will notice that in this case the difference is almost negligible.

5.10. The Harmonic Mean: Grouped Data.—Just as the geometric mean is based on an arithmetic mean of logarithms, so is the harmonic mean based on an arithmetic mean of reciprocals. The formula for this average when found from a frequency table can best be written thus:

$$M_h = \frac{N}{\Sigma \frac{f}{X}}$$

Illustrating again with the student heights, we adapt the "long" method for the arithmetic mean. Table 5.7 illustrates the procedure. The first column shows the class marks from Table 5.2, while the frequencies appear again in the second column. The figures in the third column are found by dividing each fre-

quency by the corresponding class mark. The total number of cases (here 1000) is then divided by the sum of this third column (here 5.71144) to get the harmonic mean, which turns out to be 175.09 cm. We notice here, as we have come to expect, that

TABLE 5.7.—COMPUTATION OF THE HARMONIC MEAN FROM A FREQUENCY TABLE

Class Marks (centimeters) (X)	Number of Students (f)	(f/X)
156	4	0.02564
159	8	0.05031
162	26	0.16049
165	53	0.32121
168	89	0.52976
171	146	0.85380
174	188	1.08046
177	181	1.02260
180	125	0.69444
183	92	0.50273
186	60	0.32258
189	22	0.11640
192	4	0.02083
195	1	0.00513
198	1	0.00506
Totals.....	1000	5.71144

the value of the harmonic mean is smaller than that of either the arithmetic or geometric means, although the values of all three fall very close together in this problem.

5.11. The Quadratic Mean: Grouped Data.—Just as the geometric mean of a series of numbers is based on the arithmetic mean of their logarithms, and as their harmonic mean is based on the arithmetic mean of their reciprocals, so their quadratic mean is based on the arithmetic mean of their squares. We can give the formula for the quadratic mean of numbers grouped in a frequency table as follows:

$$M_q = \sqrt{\frac{\Sigma f(X^2)}{N}}$$

For our problem of student heights this gives us Table 5.8. Here the first column contains the class marks from Table 5.2.

The second column contains the corresponding frequencies. The third column contains the squares of the corresponding numbers in the first column. And in the last column are the products found by multiplying each figure in the second column by the corresponding figure in the third column. The sums of the second and the fourth columns give us N and $\Sigma f(X^2)$, respectively, for use in our formula.

TABLE 5.8.—COMPUTATION OF THE QUADRATIC MEAN FROM A FREQUENCY TABLE

Class Marks (centimeters) (X)	Number of Students (f)	X^2	fX^2
156	4	24,336	97,344
159	8	25,281	202,248
162	26	26,244	682,344
165	53	27,225	1,442,925
168	89	28,224	2,511,936
171	146	29,241	4,269,186
174	188	30,276	5,691,888
177	181	31,329	5,670,549
180	125	32,400	4,050,000
183	92	33,489	3,080,988
186	60	34,596	2,075,760
189	22	35,721	785,862
192	4	36,864	147,456
195	1	38,025	38,025
198	1	39,205	39,205
Totals.....	1000		30,785,716

In this problem we need to find the square root of $\frac{30785716}{1000}$ or the square root of 30785.716. This gives 175.4, cm., which is, as it should be, slightly larger than the arithmetic mean.

5.12. Quartiles, Deciles, and Percentiles: Grouped Data.—The quartiles, deciles, percentiles, and other "averages of position," can either be approximated by inspection of the ogive, as explained in Sec. 5.7, or they may be located in an ordinary frequency table by methods based on the same assumptions as those used in computing the median. It is necessary merely to find the $(N/4)$th item or the $(N/10)$th item or the $(N/100)$th item instead of the $(N/2)$nd item, which we found in the case of

the median. The similarity of the methods makes it unnecessary to give an extended description of them here. The methods of finding the first quartile, the third decile, and the 57th percentile are given below as illustrations. The work here given, when studied in conjunction with the description of the location of the median, is self-explanatory.

Locating the first quartile:

$N = 1000$

$\frac{N}{4} = 250$

We wish to find the size of the 250th item.

It is in class 6.

It is the 70th item up in the class.

It is $\frac{70}{146}$ of the way up in the class.

It is $\left(\frac{70}{146}\right)$ (3 cm.) = 1.45 cm. from the actual lower limit of the class.

$Q_1 = 169.5 + 1.45 = 170.95$

Round off to give $Q_1 = 171$ cm.

Since the first quartile is the 250th item, the third quartile is the 750th item. It would be necessary to find the size of this item as we have found the size of the first quartile.

Locating the third decile:

$N = 1000$

$3\frac{N}{10} = 300$

We wish to find the 300th item.

It is in class 6.

It is 120 items up in the class.

It is $\frac{120}{146}$ of the way up in the class.

It is $\left(\frac{120}{146}\right)$ (3 cm.) = 2.47 cm. from the actual lower limit of the class.

$D_3 = 169.5 + 2.47 = 171.97$ cm.

Round off to give $D_3 = 172$ cm.

Locating the 57th percentile:

$N = 1000$

$57 \frac{N}{100} = 570$

We wish to find the value of the 570th item.
It is in class 8.
It is 56 items up in the class.
It is $\frac{56}{181}$ of the way up in the class.
It is $\left(\frac{56}{181}\right)$ (3 cm.) = 0.93 cm. from the actual lower limit of the class.
$P_{57} = 175.5 + 0.93 = 176.43$ cm.
Round off to get $P_{57} = 176$ cm.

Other quartiles, deciles, and percentiles would be computed similarly. If we interpret the three which we have just computed (using the rounded numbers), we find that Q_1 indicates that $\frac{1}{4}$ of the students are shorter than 170 cm. and $\frac{3}{4}$ are taller than 170 cm. The third decile shows that $\frac{3}{10}$ of the students are below 171 cm. and $\frac{7}{10}$ exceed this height. The 57th percentile shows that 57 per cent of the students are shorter than 176 cm. and 43 per cent are taller. Similar interpretations would be given to other such measures.

5.13. Summary of Averages with Grouped Data.—Just as we summarized in Sec. 4.11 the directions for computing the various averages when each of the original values was given separately, so we summarize here the methods for computing these averages when the data are presented to us in the form of a frequency table.

1. The arithmetic mean ($\bar{X}$). Long method:
   Write down the class marks.
   Write beside each class mark the frequency in the class.
   Multiply each class mark by the corresponding frequency.
   Add the products.
   Divide this sum by the number of cases.
   Formula;

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

2. The arithmetic mean ($\bar{X}$). Short method:
   Select as an origin a class mark near the center of the distribution.

Write down beside each class mark the number of class intervals by which it exceeds (+) or falls short of (−) the origin so selected.
Multiply each frequency by the number written beside it.
Add the products so obtained.
Divide the sum just obtained by the number of cases.
Multiply the quotient by the class interval.
Add the result algebraically to the value of the class mark chosen in the first step.
Formula:

$$\bar{X} = \bar{X}' + \frac{\Sigma(fd)}{N} Ci$$

3. The median (Med.):
Compute $N/2$ to find the location of the desired item.
Add the class frequencies to discover which class contains the median item.
Find how many items one must count into this class to reach the median item.
Divide this number by the number of items in the class which contains the median.
Multiply the decimal so obtained by the class interval.
Add the product to the actual lower limit of the class which contains the median.

4. The mode (Mo.). First approximate method:
Multiply the median by 3.
Multiply the arithmetic mean by 2.
Subtract the latter product from the former.
Formula:

$$\text{Mo.} = 3 \text{ Med.} - 2\bar{X}$$

5. The mode (Mo.). Second approximate method:
Subtract the smallest class mark in the frequency table from the value of the arithmetic mean.
Multiply the remainder by twice the number of classes in the frequency table.
Multiply the class interval by one less than the number of classes in the frequency table.
Subtract the product just obtained from the product obtained in the step next preceding.
Divide half the difference just obtained by a number which is one less than the number of classes in the table.
Add the quotient to the smallest class mark in the table.
Formula:

$$\text{Mo.} = m + \frac{2g(\bar{X} - m) - Ci(g - 1)}{2(g - 1)}$$

6. The geometric mean (M_g):
Write down beside each frequency the logarithm of the corresponding class mark.

Multiply each logarithm by the corresponding frequency.
Add the products.
Divide the sum by the number of cases.
This yields the logarithm of the geometric mean. To find the geometric mean, take the antilog.
Formula:

$$\log M_g = \frac{\Sigma f(\log X)}{N}$$

7. The harmonic mean (M_h):
Divide each frequency by the corresponding class mark.
Add the quotients.
Divide the number of cases by the sum just obtained.
Formula:

$$M_h = \frac{N}{\sum\left(\frac{f}{X}\right)}$$

8. The quadratic mean (M_q):
Square each class mark.
Multiply each square by the corresponding frequency.
Add the products just obtained.
Divide this sum by the number of cases.
Take the square root of the quotient.
Formula:

$$M_q = \sqrt{\frac{\Sigma(fX^2)}{N}}$$

9. The quartiles (Q_1, Q_3):
By adding frequencies in classes find which class contains the item that is number $N/4$ from each end.
Find how many cases one must go into the class containing the quartile.
Interpolate within the class as for the median.

10. The deciles, percentiles, etc. (D_1, D_2, etc.; P_1, P_2, etc.):
Proceed as with the median except that the item wanted is the number $N/10$, $2N/10$, $N/100$, or $2N/100$ instead of $N/2$.

5.14. Characteristics of a Good Average.—Throughout this chapter and the preceding one, we have been considering the detail involved in the methods of computing various averages. It is now time to consider the characteristics of these averages and their advantages and disadvantages. Logically it might seem desirable to have done this first, before we considered the details of computation; but pedagogically it is much easier to discuss the abstractions of advantage and disadvantage after the student has seen the thing discussed than before.

We have seen that an average is a single value selected from a group of values to represent them in some way—a value which is supposed to stand for the whole group of which it is a part, as typical of all the values in the group. If we were to enumerate the qualities that we should desire in such a typical value if we could get a perfect one, it is likely that we should list at least the following seven characteristics:

1. The number should be unequivocally defined, so that there can be no question, in any given distribution, as to just what the value is. It is important that the average be objective —possibly defined by an algebraic formula—so that if 10 different students all work with the same figures they will all (barring arithmetical mistakes) get the same answer. The average should not depend on the whim, caprice, or idiosyncrasy of the computer.

2. The average should be inherently descriptive of the data in such a way that its meaning is easily understood. It should not be such a distant mathematical abstraction that it can be comprehended only by the advanced student. Statistical methods exist to simplify data, not to make them more complex.

3. The average should, if possible, be easy to compute. This, however, is not so important as the ease of understanding. The statistician often performs difficult or tedious processes himself in order to get results that are easily understandable; and ease of computation, while desirable when other things are equal, is not to be sought at the expense of other advantages.

4. The average should depend on every single item in the group, so that if we alter the value of any member of the group we shall alter the value of the average. The average is to be thought of as typifying *all* the members of the group, not merely some of them.

5. Although every item should influence the value of the average, no item or items should influence it unduly. We should not want one or two extremely large or extremely small items to overshadow all the rest. We should prefer the items which make up the group to have approximately equal influence on the average.

6. We should like to get some value which has what the statistician calls "sampling stability." This means that if we pick half a dozen different groups of college students, and compute the average of each group, we should prefer to get

approximately the same value each time. We do not want our answer to depend too much on the particular 1000 students that we have studied, but we should like a value that is dependable—that will be about the same in one sample as in another. We know that there is a considerable difference in practice among the various averages that we have studied in this regard. Also we should prefer to get about the same answer whether we group with class intervals of 5 or 10, or whether we set different class marks with the same class interval. Minor variations in grouping of the items should not affect the average materially.

7. Finally, we should prefer to have an average that can be easily used in further statistical computation. For example, if we have computed an average for freshmen, one for sophomores, one for juniors, and one for seniors, we should like to be able to combine them to get an average for the entire undergraduate body.

5.15. Relationships between the Averages.—Before we take up the various averages for individual discussion and comparison with the criteria listed in the preceding section, it should be pointed out that certain relationships usually obtain among the different averages. The statistician is often almost as much interested in these relationships among two or more of the averages as he is in the averages themselves.

1. If the distribution is symmetrical, the values of the arithmetic mean, the median, and the mode will be identical, and if the distribution is nearly, but not quite, symmetrical their values will be almost identical. In other words, the similarity or divergence in the sizes of these three measures (or any two of them) is to some extent an indication of the symmetry of the distribution.

2. As we have already seen (Sec. 5.8), if the distribution is mound-shaped and only moderately asymmetrical, the median lies between the arithmetic mean and the mode, being approximately twice as far from the latter as from the former.

3. In any distribution where the original items differ at all in size, the following averages will all differ in size and their values will fall in the following order:

$$M_g > \bar{X} > M_a > M_h$$

In the limiting case where all the original items are identical in

size (in which case we would hardly compute an "average"), these four averages would all be equal.

5.16. Advantages and Disadvantages of the Arithmetic Mean.—The arithmetic mean is certainly the most widely used and most commonly understood of all the averages. It is a value so selected out of a group that if all members of the group were uniform in size, and if they retained their actual total size, they would each be equal to the arithmetic mean. Or we can think of the arithmetic mean of N items as a single item made up of $1/N$th part of each of the original items. If we say that the average income of 40 people is \$134.50 per month, we mean that if each of the 40 people contributed $\frac{1}{40}$th of his income to a common fund, this common fund would amount to \$134.50 per month. Thus we see that each value in the distribution plays a part in determining the arithmetic mean, and a change in any item will change the arithmetic mean. It is rigidly defined by a mathematical equation and is easy to compute. Sometimes we can compute it when we cannot compute other averages and when the values of the individual items are not known. For example, if we know that the total consumption of milk in the United States amounts to 51,100,000,000 quarts and that the population is 132,000,000, we do not need to know the facts for any individual family to compute an average consumption of 387 quarts per capita. We could not compute a single one of the other averages from these data. It is also true, as we shall discover later (see Chap. IX), that the arithmetic mean is unusually stable in sampling, running more uniform from sample to sample than any of the other averages. For scientific work this peculiarity of the arithmetic mean is of great importance.

On page 88 we noted that the sum of the deviations of a group of individual items from their arithmetic mean was equal to zero. It is also true that the sum of the squares of these deviations is smaller when taken from the arithmetic mean than when taken from any other number. Let us take, for illustration, the numbers 5, 8, and 14. If we compare these numbers with the number 10, we find that they differ from 10 by -5, -2, and 4. These differences, when squared, give 25, 4, and 16; and the sum of the squares is 45. Had we compared the three numbers with their arithmetic mean 9, we would have found the differences to be -4, -1, and 5. The squares of these

differences are 16, 1, and 25; and the sum of the squares is 42, which is smaller than the sum of the squares when the deviations were measured from 10. The student may try numbers other than 9 or 10, and he will find that the sum is smaller when deviations are taken from 9, the arithmetic mean, than when taken from any other number he can select.[1] This means that the arithmetic mean of a group of numbers is "fitted by least squares," an expression that we shall use later on in considering certain important theorems of probability. We shall discover then that a value which is fitted by least squares has more chance of being correct than any other value if the distribution is what we shall then call "normal." For our present purposes, we need merely point out that no measurement is ever made with complete exactitude (see Chap. II) and that when we measure anything over and over again we get different answers; so we can never be sure what measurement is absolutely correct. But if we measure the speed of light, the length of a line, or the weight of a cubic foot of water over and over again, getting slightly different measurements each time, the arithmetic average of these measurements has a better chance of being the actual speed of light, length of the line, or weight of the water than any other figure which we can take. This "least-squares"

[1] The student of the calculus will be able to prove this fact easily for all cases, rather than relying on experiment. Suppose we are to choose any value M from which to measure the deviations d. Then for each value of our original series X, we have:

$$X + d = M$$
$$d = M - X$$
$$d^2 = (M - X)^2 = M^2 - 2MX + X^2$$
$$\Sigma d^2 = \Sigma(M^2 - 2MX + X^2) = NM^2 - 2M\Sigma X + \Sigma X^2$$

This is the sum of the squares of the deviations, which we wish to minimize. It will have its minimum value when the first differential is equal to zero. If we represent this function by f, we have

$$\frac{df}{dM} = 2NM - 2\Sigma X \text{ which we set equal to zero}$$
$$NM = \Sigma X$$
$$M = \frac{\Sigma X}{N}$$

In other words, we shall get the smallest possible sum of the squared deviations when our value of M is $\Sigma X/N$, that is, when it is equal to the arithmetic mean.

property of the arithmetic mean, though often overlooked, is one of its most important characteristics to the scientist.

As a final advantage of the arithmetic mean, we can point out that it is unusually adaptable when we wish to carry on further mathematical computations with it. Suppose we have three basketball teams made up of five men each. The average weight of the members of Team A is 145 lb., that of Team B is 158 lb., and that of Team C is 162 lb. We can combine these averages directly to find the average weight of the 15 players. Since each average is based on 5 men, we merely take the arithmetic mean of the three means, to get an average of 155 lb. for the 15 men. But we can still compute this mean for the entire group from the means of the subgroups even if the subgroups do not contain equal numbers of cases. If we call our subgroups 1, 2, 3, . . . , N, and represent the totals of the values in the individual subgroups as ΣX_1, ΣX_2, ΣX_3, . . . , ΣX_N, it is evident that the grand total of all the values in all the groups thrown together is the sum of the totals in the subgroups, or, in the form of an equation, if we let ΣX represent the sum of all the items in all the groups together, we get

$$\Sigma X = \Sigma X_1 + \Sigma X_2 + \Sigma X_3 + \cdots + \Sigma X_N$$

But

$$\Sigma X = N\bar{X}, \qquad \Sigma X_1 = N\bar{X}_1, \text{ etc.}$$

Therefore

$$N\bar{X} = N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3 + \cdots + N_N\bar{X}_N$$

The desired value of $\bar{X}$ must be

$$\bar{X} = \frac{N_1\bar{X}_1 + N_2\bar{X}_2 + N_3\bar{X}_3 + \cdots + N_N\bar{X}_N}{N}$$

Thus all we need to do is to take a weighted arithmetic mean of the averages, using as weights the numbers of cases in the individual subgroups. We multiply each subgroup average by the number of cases in that subgroup, add the products, and divide the sum by the grand total number of cases in all groups together. This gives us the average for all groups together. No such computations can be carried out for the median or the mode. If we know the medians or the modes of several subgroups, we cannot find the median or the mode of the whole lot together.

The arithmetic mean thus has so many advantages that it is used far more than all the other averages combined. We might almost say that in case of doubt the arithmetic mean should be used—that other averages should be used only when there is some clear reason for it. Yet the arithmetic mean does have one or two distinct disadvantages. In the first place, it is very sensitive to extremely large or extremely small items (especially so to large ones). The chance inclusion of such an extreme item in the group being studied may give us an arithmetic average which is not really typical of the group. Let us consider the numbers 6, 7, 7, 8, 8, 8, 8, 8, 9, 9, 865. The median of these numbers is 8, and their mode is 8, but their arithmetic mean is 85.7. The latter figure does not depict well either the one large number or the 10 small ones. The inclusion of the single very large number at the upper extreme has thrown our arithmetic mean to a point far from any of the actual items in the group.

Where there is marked skewness in a distribution, so that the arithmetic mean, the median, and the mode differ widely in value, one should always consider the possibility that the arithmetic mean is not a truly representative or typical value, and that the median or the mode should be used in preference to it.

In the United States the distribution of incomes is decidedly asymmetrical, as will be seen from the accompanying chart.[1] If it is said that the arithmetic average income for 1918 was $1690 (which is a rough average of the figures on which the chart is based), the reader is likely to be misled. Between 70 and 75 per cent of the income earners of 1918 earned less than $1690. In other words, $1690 was not only the mean income but one of the higher incomes. The median income was roughly $1170, and the modal income was presumably even lower.[2] For many purposes one is likely to be more interested in the modal income

[1] Based on figures from Warren C. Waite, "Economics of Consumption," p. 22, McGraw-Hill Book Company, Inc., New York, 1928. Waite quotes them from the National Bureau of Economic Research publication "Income in the United States," Vol. I. The figures are for 1918, but there is good reason to believe that the asymmetry still exists.

[2] If we were to compute the mode from the mean and median in accordance with the formula on p. 98, we should find that the modal income was $130. This is obviously altogether too low. It is to be remembered that the formula just mentioned is to be applied only in those cases where the asymmetry is moderate. In this case the asymmetry is great.

in the United States than in the mean income. A few millionaires raise the mean income tremendously without raising the typical plane of living particularly. For this reason, then, we may also prefer to use some average other than the arithmetic one. Similarly we realize that the arithmetic mean of a U-shaped distribution would fall at a point where values were uncommon, and that in such peculiar distributions the arithmetic mean would give a misleading idea of the distribution.

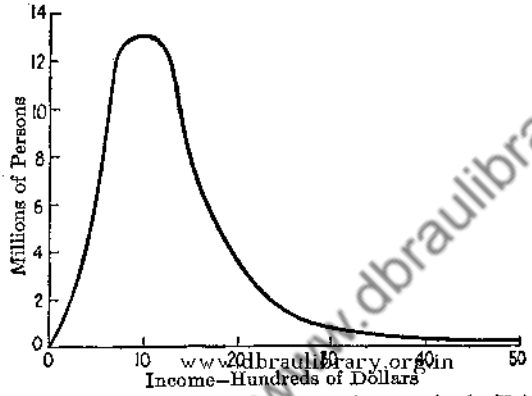


FIG. 5.3.—An approximate distribution of money incomes in the United States in 1918. The curve continues toward the right to incomes of millions of dollars.

5.17. Advantages and Disadvantages of the Median.—The median is rigidly defined (if there is any median at all), and the concept involved is readily understood by anyone even though the term itself may be unfamiliar. If the data are in an array or a frequency table, the median is easy to compute, and items of extreme size have almost no influence on it. It has less sampling stability than the arithmetic mean. If we had 10 groups of newborn babies and found the median weight in each group, we should discover that these medians not only differed in size, but their variation was about a quarter again as large as the variation in the sizes of the arithmetic means of the same ten samples. The median has the advantage that we can compute it even from a frequency table with open-end classes, since we do not need to know the sizes of the extreme items as long as we know that they are extreme items. We can even find the median in cases which are nonmathematical in character, and where numerical measurement is impossible. For example, we might

arrange a number of pieces of blue cloth in order of the intensity of their color. The color of the piece in the middle will then be the median color. Thus the median can be used with data which are nonmathematical in character. A characteristic of the median which is also sometimes useful is that the sum of the absolute deviations (disregarding plus and minus signs) is smaller when measured from the median than when measured from any other value. We showed in the preceding section that the algebraic sum (keeping track of signs) of these deviations was smallest when measured from the arithmetic mean. If we go back to the example which we then used, taking the numbers 5, 8, and 14, the median is 8. The absolute deviations are 3, 0, and 6, giving a sum of 9. Had we chosen the arithmetic mean, which is 9, the absolute deviations would have been 4, 1, and 5, giving a sum of 10. If we chose still other numbers than 8 or 9 from which to measure the absolute deviations, we should find that the sum of these absolute deviations was smaller when measured from 8 than from any other number.[1]

The median has the disadvantages that it is not quite so well known as the arithmetic mean (although easily explained), and that it is necessary to arrange the items in an array before it can be computed. Sometimes there is no median in a discrete series, as was illustrated at the end of Sec. 5.6. Also the median is not adapted to further arithmetical work. As we have seen, if we are told the medians of each of several subgroups, there is no way of finding the median of the group as a whole. Moreover the median is not sensitive to changes in the values of the items that make up the distribution. We can change the sizes of items without influencing the value of the median at all as long as we do not change the size of any item enough to move it to the

[1] The student can readily see that this must be true if he considers the fact that the sum of the distances from two given points to any third point is smallest when the third point lies between them and is always the same no matter where the third point lies on the line between them. The median has as many points on one side of it as on the other, so the points (or values in a distribution) can be taken in pairs, one each side of the median, and for each pair the sum of the distances will be less than for any point which does not lie between them. Therefore the sums of the distances for all such pairs will be smaller when measured around the value in the center than it will be for any point not so located that equal numbers of points lie below and above it.

opposite side of the median. Each item does have a minor and tenuous influence on the size of the median merely by means of the fact that it is larger or smaller than the median, but aside from that the size is immaterial. It is an advantage, as we have seen, for an average not to be overly sensitive, but many workers feel that the median is not sensitive enough.

5.18. Advantages and Disadvantages of the Mode.—Any average is a single value taken to represent a whole group of values. There would be no justification in selecting any such representative value if the items in the original group were not concentrated or clustered about some point. The fact that the mode indicates this point of heaviest concentration makes it in its abstract aspects perhaps the best average of all. We have already seen, in considering an illustrative problem on distribution of incomes (Sec. 5.16), that when a distribution is badly skewed or non-normal we are more likely to be interested in the mode than in the arithmetic mean. In fact, some authorities have suggested that if the mode and the arithmetic mean are significantly different in value, one should use the mode. This is perhaps going too far, yet we must realize that the mode is, as we suggested in Sec. 5.14, "inherently descriptive of the data in such a way that its meaning is easily understood." It is the concept in which the layman is perhaps most often interested, even though he may not be familiar with the name. Moreover the mode is hardly at all influenced by the values of extreme items.

On the other hand, the mode is difficult to compute, and the rough approximations to it which we have so far discussed are unreliable and peculiarly subject to instability of sampling. It is possible to make radical changes in the sizes of items in a distribution without changing the value of the mode at all. While it can be said that the mode depends on the values of all the items to the extent that the mode would have been different if enough of the items had been different, this is almost the same as saying that some of the items have little or no effect on the value of the mode. A distribution may have no mode, and usually there will be no well-defined mode unless the number of cases is large. Or there may be two or more modes, although in such cases the statistician usually investigates to see if his data are really homogeneous. For example, when studying a dis-

tribution of wages one might find two high points on his frequency polygon. He would then ask himself whether this might be because he had lumped together wages of men and of women, or wages of skilled and of unskilled workers, or wages of organized and unorganized workers. And finally, the mode cannot be easily used in further algebraic processes. If we have the modes of several distributions, we cannot combine them to get a new mode of the joint distribution. It is unfortunate that an average which has such an intellectual appeal as the mode happens to be so difficult to compute and so unreliable after it is computed.

5.19. Advantages and Disadvantages of the Geometric Mean. The geometric mean is unusual enough so that it is quite natural for the student to ask why anyone ever bothers to compute it. It is obvious that we *can* multiply 20 numbers together and take the 20th root of the product (or perform the equivalent computations by means of logarithms), but why should we want to do it? The answer is that in certain sorts of problems this is the only way to get the right answer. Just as, in the case of the weighted arithmetic mean, we saw that one weights his average in order to get the right answer, and not because he likes the complications of the method, so in the case of the geometric mean, the method is used in spite of its complications and not because of them.

We can start our discussion of the geometric mean by pointing out that it meets certain of the requirements that we listed in Sec. 5.14. It is rigidly defined by a mathematical formula, so that the result does not depend in the slightest on the whim of the worker who computes it. It depends on the value of every item in the distribution; no single item can be changed in the least without affecting the value of the geometric mean. Its value is not quite so greatly influenced by extreme items as are the values of the quadratic, arithmetic, and harmonic means. And the result can be used in further statistical work. The geometric means of samples can be combined to get the geometric mean of the whole.

In addition to these advantages, however, there are two sorts of cases in which the use of the geometric mean is particularly indicated. First, we have those cases in which we are finding the average of values which are in geometric progression. For example, we might take the progression which runs 1, 2, 4, 8, 16,

etc. Suppose we wish to find the average of these five numbers. If we add them and divide by five, we find their arithmetic mean, 6.2. If, on the other hand, we multiply them and take the fifth root, we get the geometric mean, 4. We note that 4 is actually the number which appears in the middle of the distribution, while 6.2 is not. Moreover, if we make a graph of the data, as in Fig. 5.4, we find that the value 4 really falls on the line, while the value 6.2 does not.

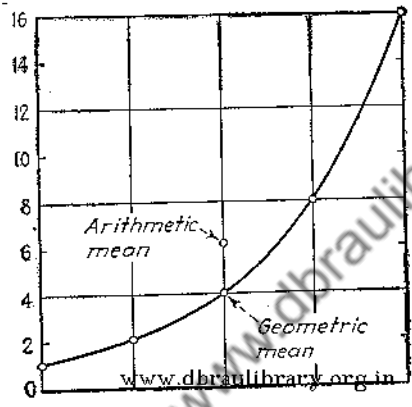


FIG. 5.4.—Arithmetic and geometric means of the numbers 1, 2, 4, 8, and 16.

We have a geometric series whenever the quotient found by dividing any term by the term following is constant throughout the series. In the series above, 1, 2, 4, 8, etc., if we divide any term by the term following we get the quotient ½. We do not insist, of course, that the quotients be exactly equal, as long as they are approximately so. The population of the United States in the first eight censuses was (in millions) as follows:

1790	3.9	1830	12.9
1800	5.3	1840	17.1
1810	7.2	1850	23.2
1820	9.6	1860	31.4

If we divide each of these numbers by the one following, we get the following seven quotients: 0.74, 0.74, 0.75, 0.74, 0.75, 0.74, and 0.74. The approach to uniformity is startling. The series seems to be geometric. If we were to compute the average for this series we should take, not the arithmetic mean of 13.8 million, but the geometric mean 11.1. It will be noted that the

arithmetic mean gives a value even higher than the population of 1830, which is well beyond the middle of the period; while the geometric mean gives a value which not only lies between the populations of 1820 and 1830, but is almost exactly the geometric mean of them.

As another example, if, during a 10-year period, a sum of money at interest grew from \$100 to \$500, how large was the sum at the middle of the period? The natural inclination is to take the arithmetic mean of \$100 and \$500, or \$300. But if the sum increased to three times the original amount in half the period, it should increase to three times three, or nine times the original amount, in the whole period. We must take the geometric mean of the numbers 100 and 500, which is \$223.60. If the original sum was multiplied by 2.236 in the first half, it must have been multiplied by 2.236 twice, or by 2.236^2, or by 5, in the whole period.

Experience shows that many sorts of phenomena in many different sciences tend to grow geometrically. In such cases it is evident that the geometric mean will give the correct answer, while the arithmetic mean will not. We use the geometric mean, not because we like to play with logarithms, nor because it makes us appear sophisticated, but because it is accurate for these data. For other data it would be misleading.

The student might try proving for himself that if the population of a city increases 30 per cent in 10 years it does not increase 3 per cent each year. If it did, and we started with a population of X persons at the beginning, at the end of one year there would be $1.03X$, at the end of the second year there would be 1.03^2X, and at the end of the 10th year there would be $1.03^{10}X$, or $1.34X$, and the population would have increased 34 per cent rather than 30 per cent. This problem calls for another application of the geometric mean, using a formula for geometric progressions which is well known to all students of mathematics as applied to finance. Let us represent the amount at the beginning of our period by B, and the amount at the end by E, while we represent the rate of increase per unit of time by r. Let n represent the number of units of time in our entire period. Then our formula becomes

$$r = \sqrt[n]{\frac{E}{B}} - 1$$

Applied to the problem which we have just attempted, the population at the beginning is X and at the end is $1.30X$. The length of period is 10 years. Substituting in our formula, we get

$$r = \sqrt[10]{\frac{1.30X}{X}} - 1 = \sqrt[10]{1.30} - 1$$

Using logarithms, we find that this yields

$$r = 0.027$$

Instead of an average rate of 3 per cent each year, it turns out that we had an average rate of increase of 2.7 per cent each year. Using the arithmetic mean gives us the wrong answer; using the geometric mean gives us the right answer. If the student will start with any number and multiply it by 1.027 ten times, he will find that he actually ends with a number 30 per cent larger than the one with which he started, proving the correctness of the geometric method.

In addition to these cases of geometric rates of increase or decrease, in which the geometric mean must be used, we shall see in Chap. XII that there are certain theoretical advantages in using the geometric mean when computing index numbers.

Yet the geometric mean has serious disadvantages. Foremost, perhaps, is the fact that most people do not understand the results and are afraid of the method. While it is not really very difficult to compute, it seems so to the student who has a fear of logarithms. The statistician should use it, of course, in those cases to which it is adapted; but even here it would probably be better if he called his answer merely "the average" rather than "the geometric average." If any value in the original series is zero, the geometric mean assumes a value of zero regardless of the sizes of the other items. If any value in the original series is negative, the geometric mean may be either negative or imaginary. In cases where the number of items in the series is even, there are always theoretically two possible values of the geometric mean, one positive and one negative. For example, the square root of 16 is either $+4$ or -4. In such cases, however, we always take the positive value as the geometric mean.

Some authors suggest that when a distribution has considerable positive skewness (as defined in Sec. 8.6), or when there is a

definite lower limit to the values and no definite upper limit (as in a distribution of the numbers of families with various numbers of children, where zero is a lower limit and there is no upper limit) the geometric mean should be used in preference to the arithmetic mean. Other authors suggest that if the frequency distribution with logarithmic frequency classes (see Sec. 3.20) is more symmetrical than the corresponding distribution with equal class intervals the geometric mean should be used in preference to the harmonic mean. We can summarize, however, with the statement that the usual places to use the geometric average are cases involving average rates of increase or decrease, or cases involving the computation of index numbers.

5.20. Advantages and Disadvantages of the Harmonic Mean. The computation of the harmonic mean is somewhat more cumbersome than that of the arithmetic mean, and, as with the geometric mean, we need some explanation of why anyone would want to compute such an average at all. Again the answer is that for certain rather unusual types of problems the harmonic mean is correct and the geometric mean is incorrect. The harmonic mean is ordinarily used only in averaging certain kinds of rates, and even then only under certain conditions. Let us take an example.

Mr. Sedgewick drives his car at the rate of 25 miles per hour, while Mr. Kinsey drives his car at the rate of 50 miles per hour. What is their average speed? The answer obviously depends on whether they drive for the same distance or for the same time. If they both drive the same distance, say 50 miles, Sedgewick takes 2 hr. and Kinsey takes 1 hr., or they take a total of 3 hr. for the 100 miles. This is an average of $33\frac{1}{3}$ miles per hour. If they both drive for the same time, say 1 hr., Sedgewick drives 25 miles and Kinsey drives 50 miles, or they take 2 hr. for 75 miles, or 37.5 miles per hour. It will be noted that in the one case the average speed was $33\frac{1}{3}$ and in the other case 37.5 miles per hour. Both answers are correct. The only difference is in whether we kept the time or the distance constant for the two men. We should note, though, that one of the cases (the second) is the arithmetic mean of the two rates; while the other case (the first one) is the harmonic mean of the two rates. Our original figures were given as 25 and 50 miles per hour. In these original figures the time (hours) was constant and the distance (miles)

varied. If we want the average with both men driving for the same time, we take the arithmetic mean of the rates. If we want the average with both men driving for the same distance, we take the harmonic mean of the two rates.

We can generalize this rule by noticing that in every rate there is a variable and a constant term. For example, in miles per hour the hour is constant and the miles vary; in dollars per dozen the dozen is constant and the dollars vary; in output per man the man is constant and the output varies. If we want to take an average of such rates, we must decide whether it is desired to keep constant in our average the factor that was constant in the rate (in which case we use the arithmetic mean of the rates), or the factor that was variable in the rate (in which case we use the harmonic mean). Two further examples follow as illustrations:

1. Suppose we buy bananas at one store for \$5 per bunch and in another store for \$10 per bunch. What is the average expenditure?

The rate is dollars per bunch, with the number of dollars varying and the bunch constant. If we are to assume that we buy the same number of bunches at each store, we should use the arithmetic mean of \$5 and \$10 (since we are keeping constant the same factor, bunches, which is constant in the rate). If we are to assume that we spend the same amount of money at each store (say \$50), we should use the harmonic mean of \$5 and \$10 (since we are keeping constant the factor, dollars, which was variable in the rates).

2. Mr. Jorgensen gets 12 miles to the gallon of gasoline with his car, and Mr. Gentry gets 18 miles to the gallon. What is the average gasoline consumption?

In this case the gallon is constant and the miles vary. If we assume that both men drive the same distance we should take the harmonic mean of 12 and 18, finding an average of 14.4 miles to the gallon. If we assume that both men use the same number of gallons, we should take the arithmetic mean of 12 and 18, getting an average of 15 miles per gallon. We can prove these statements by taking numerical examples. If the two men drive 36 miles each (keeping the distance the same for both of them) Jorgensen will use 3 gal. and Gentry will use 2 gal. This makes 5 gal. for 72 miles, or 14.4 miles per gallon. On the other

hand, if the men drive until they have used 2 gal. apiece, Jorgensen will have driven 24 miles and Gentry will have driven 36 miles. This will make 60 miles for 4 gal., or 15 miles per gallon. We thus see that in one case the arithmetic mean gives us the right answer, while in the other case we have to use the harmonic mean.

We can summarize again, by saying, then, that the harmonic mean is used in certain cases where we are finding the average of rates. Every rate is stated as a variable number of units of something per constant number of units of something else—as a variable number of miles per single unit of time. If, in taking our average, we keep constant the factor that was variable in the rate, we must use the harmonic mean.

The harmonic mean also has the advantages that it is rigidly defined, and its value depends on the value of each item in the distribution. The results can be used for further mathematical computation.

On the other hand, the concept is an unfamiliar one, difficult for the layman to understand, and somewhat more difficult to compute than the arithmetic mean. It is greatly influenced by extreme items, especially so by extremely small items. It cannot be computed at all if any item in the distribution is equal to zero. The statistician would do well to use some other average save in those cases, just described, where he cannot get the right answer without it.

5.21. Advantages and Disadvantages of the Quadratic Mean. The use of the quadratic mean can best be illustrated in connection with an actual case in the next chapter. In certain probability problems it is theoretically important to deal with the squares of numbers, rather than with the numbers themselves. In such cases the quadratic mean is natural. In other cases, as in dealing with deviations from the arithmetic mean, we cannot take an arithmetic average of the deviations, because the sum of the deviations is always zero—the positive and negative deviations balancing each other (see footnote on page 88). If we square the deviations, however, they are all positive, and we can take the arithmetic mean of the squares. In such a case, also, the quadratic mean is natural. But this average, although rigidly defined and amenable to further algebraic manipulation, is very greatly influenced by extremely large

values, is somewhat more difficult to compute than the arithmetic mean, and is not simple enough to be readily understood by the layman. Therefore the statistician uses it only where there is a real reason for it.

5.22. Summary of the Averages.—The advanced student will learn for himself to use that discrimination in the choice of averages which is a mark of statistical competence. The beginner may well follow the rule of using the arithmetic mean in preference to all other averages except in the following cases:

1. When the distribution is badly skewed, consider the advisability of using the median or the mode. The mode is the harder to find and less reliable than the median, but is perhaps the most natural of all concepts of averages (see Secs. 5.17 and 5.18).
2. When the distribution is U-shaped, use the two modes.
3. When the items form a geometric progression, use the geometric mean (see Sec. 5.19).
4. When finding an average rate of growth or change over a period of time, use the geometric mean (see Sec. 5.19).
5. When logarithmic frequency classes give a more symmetrical frequency polygon than equal class intervals, use the geometric mean (see Sec. 3.20).
6. When averaging rates, and it is desired to keep constant in the average the factor that is variable in the rate, use the harmonic mean (see Sec. 5.20).
7. For certain index-number problems, use the geometric mean (see Chap. XII).
8. Whenever there is any reason to believe that the arithmetic mean would be seriously misleading, on account of undue influence from extreme items or for other reasons, consider the advisability of using the median or the mode.

5.23. Suggestions for Further Reading.—The student will find a complete discussion of the problems here treated, in a great deal greater detail, in Franz Zizek, "Statistical Averages," Henry Holt and Company, Inc., New York, 1913. A short, but rather technical and mathematical, treatment can be found in "Handbook of Mathematical Statistics," edited by H. L. Rietz, Houghton Mifflin Company, Boston, 1924. A number of interesting and useful mathematical theorems with regard to the various averages is treated by John F. Kenney in his "Mathematics of Statistics," Chap. III, D. Van Nostrand Company, Inc., New York, 1939.

EXERCISES

1. If we have a frequency distribution which is almost, but not quite, symmetrical, and the values of the mean and the mode are 27 and 29 lb., respectively, what will be the approximate value of the median?

2. When we are computing the median, why do we make a different assumption as to the location of items within frequency classes from that which we make when computing the mean?

3. Table 5.9 shows the number of laborers in the bread departments of American bakeries who, in 1931, were receiving hourly wages of various amounts.[1] From this frequency table compute the mean and the median hourly wage. Compute the modal wage by each of the two methods explained in this chapter, and compare the results. Compute the quartiles, the seventh decile, and the seventh percentile.

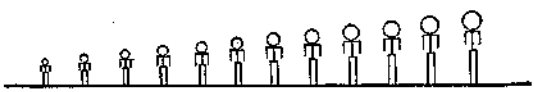
4. Compute the mean of the figures in the preceding exercise by the long and by the short method, timing the processes. Compute by the short method first, so that any advantage which may come from familiarity with the data will accrue to the long method.

5. Compute the mean of the figures in Exercise 3 above by the short method, taking a different guessed mean from that used before. Note the identity of the results. Note also why it is best to take a guessed mean near the center of the table.

TABLE 5.9.—HOURLY WAGES IN BREAD DEPARTMENTS, OF BAKERIES IN UNITED STATES, 1931

Hourly Wage (cents)	Number of Laborers
0– 9.9	1
10–19.9	58
20–29.9	206
30–39.9	442
40–49.9	478
50–59.9	294
60–69.9	79
70–79.9	20
80–89.9	2

6. Below is a diagram representing an array of heights. The figures can be thought of as representing 12 men standing in line and arranged in order of height. Suppose that we were to consider the median as located at the item represented by $N/2$ instead of $(N + 1)/2$. This would be the $1\frac{1}{2}$ item. Locate the $1\frac{1}{2}$, or sixth, item and find how many men are on



each side of it. Locate the $(N + 1)/2$ item and see how many men are on each side of that. Locate likewise the quartiles as they would be if based on $N/4$ and $3\ N/4$ instead of on $(N + 1)/4$ and $3(N + 1)/4$. See how many men each method puts in each quarter. The object of this exercise is to point out why we add unity in the formulas for median, quartiles, deciles, etc.

[1] Data from *U.S. Bureau of Labor Statistics Bulletin* 580, Table 5, p. 11.

7. Try applying the Charlier check to the data of Exercise 4.

8. Find the median of the data in Table 5.9.

9. Find the mode of the data of Table 5.9, using as the basis your figures for the arithmetic mean and the median.

10. Find the mode of the data of Table 5.9, using the method described in number 5 of Sec. 5.13.

11. Make an ogive of the data in Table 5.9.

12. Find the median and the quartiles from the ogive made in the preceding exercise.

13. Find the geometric mean of the data in Table 5.9.

14. Find the harmonic mean of the data in Table 5.9.

15. Find the quadratic mean of the data in Table 5.9.

16. List several other cases, similar to that mentioned in Sec. 5.17, where the median can be found for nonquantitative data.

17. In a certain fraternity house there are 7 seniors whose average weight is 165 lb., 9 juniors with an average weight of 160 lb., 13 sophomores with an average weight of 152 lb., and 20 freshmen with an average weight of 150 lb. What is the average weight of the 49 members of the fraternity? Use the arithmetic mean.

18. If a defense bond costs \$18.75 today, and if it matures in 10 years at \$25, it has increased in value by 33⅓ per cent in 10 years. How much did it increase in value each year? In other words, what was the equivalent annual interest rate?

19. Suppose that Mr. Carter pays 3 cents per kilowatt-hour for his electricity and Mr. Leonard pays 5 cents per kilowatt-hour. Mr. Carter uses 350 kw.-hr. and Mr. Leonard uses 300 kw.-hr. Find the average cost per kilowatt-hour. Note that the answer is neither the simple arithmetic nor the harmonic mean of 3 cents and 5 cents.

20. Explain under what circumstances the average cost in the preceding exercise would have been the arithmetic mean of 3 cents and 5 cents.

21. Explain under what circumstances the average cost in Exercise 19 would have been the harmonic mean of 3 cents and 5 cents.

22. What kind of average can one take of 3 cents and 5 cents in Exercise 19 to get the correct answer? (The correct answer in Exercise 19 is 3.923 cents per kilowatt-hour.)

CHAPTER VI

MEASURES OF DISPERSION

6.1. Variability.—In the preceding chapters we attempted to find single values which could be used to represent whole groups of values. We tried, for example, to summarize the heights of 1000 students by saying that the median height was 175 cm. Yet a moment's consideration will make it plain that two frequency distributions may have averages which are exactly alike, even though the distributions are in other respects decidedly dissimilar. That is to say, the average does not tell the whole story about the characteristics of the distribution.

Suppose that we have three distributions, each containing five values. They are as follows:

Distribution I. 120, 120, 120, 120, 120
Distribution II. 116, 118, 120, 122, 124
Distribution III. 5, 17, 51, 140, 387

The arithmetic mean of each distribution is 120; the medians of the first two distributions are also 120. Yet there are decided differences between the distributions. In the first distribution, either the mean or the median is a perfect figure for representing the values of the group; either average represents each individual item with complete accuracy. In the case of the second distribution, either the mean or the median coincides with but one of the values. If we use it to represent any of the other values in the group, we shall have more or less error. However, the error is not great in any case, and the errors of overstatement are exactly balanced by the errors of understatement. In the third distribution, neither the mean (120) nor the median (51) represents the items particularly well. The items are widely scattered, and many of them lie far from the mean or from the median or from any other single value which we might choose to represent them.

Here, then, are three distributions with the same arithmetic mean, yet the distributions are markedly dissimilar. It would be

quite as easy to illustrate with cases where the median or the mode was the same in a number of radically different distributions.

One of the most noticeable differences between the three distributions we have just used as illustrations is the great difference between the degrees of concentration of the values. In distribution I the five values are identical; there is no divergence at all. In distribution II there is a small scatter of values, but on the whole they are bunched fairly close to each other. In distribution III there is a great dispersion of values with no tendency for items to fall close to any point of concentration. In this chapter we study measures which show the amount of dispersion among data. These measures are variously called measures of *dispersion*, measures of *scatteration*, measures of *variability*, and measures of *variation*. Looked at from the opposite point of view they could, of course, be considered measures of *concentration* or measures of *congregation*. The name is not particularly important, but the concept is. In this book the term "dispersion" is commonly used, since it has the advantage of most general adoption.

6.2. The Range.—On page 24 are listed the marks received in an examination taken by 90 students. The marks are arranged in an array. It is fairly easy to see, by a glance at the array, that there is a considerable dispersion of values. One of the commonest measures of dispersion in popular use is evident from the data as they appear. This measure of dispersion is called the *range*, and is equal to the difference between the largest and the smallest values in the group. In the case of the examination marks, the highest mark was 206 and the lowest was 43. We find the range by subtracting the smallest from the largest, thus:

$$\text{Range} = 206 - 43 = 163$$

When we say that the range of the marks is 163, we obviously say something about the degree of their concentration. If, again, we were to compute the ranges of the three distributions on page 126, we should find these values:

$$\begin{aligned} \text{I.}\quad & 120 - 120 = 0 \\ \text{II.}\quad & 124 - 116 = 8 \\ \text{III.}\quad & 813 - 5 = 808 \end{aligned}$$

It is evident that, *ceteris paribus*, the larger the range, the greater is the scatter of the values in the group.

When we attempt to determine the range of the items in a frequency table, we run into the difficulty of not knowing for certain the size of any item. We do not know the size of the largest or the size of the smallest item; hence we cannot determine accurately the difference between them. We can, however, tell approximately how large they are. If we go back to the figures on heights of Harvard students (page 82), we note that the smallest possible height (taken to the nearest unit) would be 155 cm. and the greatest possible height 199 cm. Thus a rough approximation of the range would be $199 - 155 = 44$ cm. We could, of course, take the two extreme class marks and call the difference between them the range. In this case it would be $198 - 156 = 42$ cm. Either approximation is good enough, since, as we shall now see, the value of the range is at best subject to considerable chance variation.

The value of the range depends on but two items in the distribution: the largest and the smallest. Yet we have already noted (page 112) that it is at the extremes that chance variations are most noticeable and have the greatest effect. The largest item included in any group is largely a matter of chance. If we select groups of 1000 college students at random, there will be much less variation between the medians of the groups than between the extreme items. And the range is dependent entirely on the two extreme items—the two items that are above all subject to chance fluctuation. On this account the range is itself very unstable. If it were not for this fact, the range would be an extremely useful measure of dispersion, because it is easily understood and easy to compute. But its instability is such a serious fault that it is seldom used as the measure of dispersion in work where care and accuracy count. Only if ease of popular comprehension is more important than are exactitude and stability do we use the range.

6.3. The Semi-interquartile Range.—In order to escape from the chance fluctuations which occur toward the extremes of frequency distributions, statisticians are likely to discard the extreme items and find the amount of variation in the central part of the data. It is common for them to discard the upper quarter and the lower quarter of the items, and to measure the

range in the remaining central half of the items. Thus we can find the value of the third quartile and subtract therefrom the value of the first quartile. This will give us the interquartile distance, or the interquartile range. For reasons that will appear later, statisticians more commonly use half of this distance as their measure of dispersion; that is, they subtract the value of the first quartile from that of the third quartile, and take half of the difference as their measure of scatter. Since we have previously computed the quartiles of several distributions, we can immediately determine the value of the interquartile range, and, what is more useful, of one-half of it, that is, of the *semi-interquartile range.*

In the case of students' heights we have seen (page 103) that the first quartile of heights is 170.95 cm. The third quartile turns out to be 179.84 cm. The interquartile range is

$$179.84 - 170.95 = 8.89 \text{ cm.}$$

The semi-interquartile range is one-half of this value, or 4.44. If we round off this value as before, we find that the semi-interquartile range is 4.4 cm.

If we let the letter Q stand for the semi-interquartile range (since it has no subscript, it will not be confused with the symbols for the quartiles themselves), we can summarize our method of computing the semi-interquartile range by the formula

$$Q = \frac{(Q_3 - Q_1)}{2}$$

The interquartile range is obviously the range within which half of the items fall—the central half of the items. In the above example we found that the interquartile range was 9 cm. (8.89 cm). This means that within a range of 9 cm. were to be found half of the students measured. When we divide this figure by 2, it is on the assumption that the distribution is approximately symmetrical. Patently the quartiles of a symmetrical distribution will be equally distant from the median (and from the mean and the mode, since these three averages will coincide in a symmetrical distribution). If we divide the interquartile range in half, we have the distance from the median down to the lower quartile and the distance from the median up to the upper quartile. Thus, if the distribution is symmetrical, the semi-

interquartile range tells the distance we must go above and below the median to include half of the cases. In the present example we may say, since $Q = 4.5$ cm., that by including all students whose heights are within 4.5 cm. of the median we shall include just half of the cases. The other half of the students will be more than 4.5 cm. removed from the average height.

To illustrate again, we discovered on page 75 that the quartiles of the examination marks were as follows: $Q_1 = 88.75$; $Q_3 = 149.25$. In this case the semi-interquartile range is

$$Q = \frac{(149.25 - 88.75)}{2} = \frac{60.5}{2} = 30.25$$

If the distribution were exactly symmetrical, the median would be just halfway between the quartiles, and removed by 30.25 from each of them. Reference to page 75 will show that the median mark was actually 112, which is 23.25 from the lower quartile and 37.25 from the upper quartile. The average of these is $(23.25 + 37.25)/2 = 60.5/2 = 30.25$. This is the semi-interquartile range which we have already computed.

Thus, in distributions where there is not complete symmetry, Q measures the average distance from the quartiles to the median. If we were given merely the median and the semi-interquartile range for these marks (that is, if we are told merely that Med. = 112 and $Q = 30.25$), we should be forced to interpret the latter measure in some such language as this: "If the distribution of marks is symmetrical, half of the marks are within 30.25 of 112 and half of the marks are farther removed from 112 than 30.25. At any rate, regardless of symmetry, if we discard the marks of the lowest quarter and also those of the upper quarter, the marks of the remaining half of the students will fall within a range of 60.5."

6.4. The Average Deviation.—There are, of course, some disadvantages in discarding two quarters of the data in order that we may measure the dispersion of the remaining half. We should usually prefer some measure of dispersion based on all the items. It is obvious that we can get such a measure if we find how far each item is from the average, and then take an average of these deviations. Thus if we have the five values of distribution III on page 126, we can go about the process of measuring dispersion as follows:

Distribution III. 5, 17, 51, 140, 387

We have already noted that the mean of these items is 120. Now the first item differs from the mean by 115, the second item differs by 103, the third by 69, the fourth by 20, and the fifth by 267. If we average these we get

$$\frac{115 + 103 + 69 + 20 + 267}{5} = \frac{574}{5} = 114.8$$

This is the average amount by which the items differ from the mean, and is called the *average deviation.*

Now it will be seen that we neglected the fact that some of the deviations were positive and some negative. As a matter of fact we should have stated the deviation of the first item as -115 and that of the last item as $+267$. Unfortunately, if we kept the signs and added algebraically, the positive values and the negative values would cancel each other, since it is a characteristic of the arithmetic mean of any group of values that the algebraic sum of the deviations from the mean is zero.[1] Hence in computing the average deviation we neglect the signs of the deviations and add their absolute values.

If we are to give in a formula directions for computing the average deviation of ungrouped data, we shall need some symbol to represent the amount by which an item differs from the average. It is customary to represent this deviation by small rather than by capital letters. Thus the amount by which any X differs from the mean of the X's is represented by x. We can define this term by the equation

$$x = X - \bar{X}$$

To summarize the method of computing the average deviation (which is itself symbolized by AD), we have

$$AD = \frac{\Sigma(|x|)}{N}$$

The vertical lines beside the x mean that the signs are to be neglected—that we are to add the values of the deviations as

[1] In fact it is this characteristic of the mean which makes possible the computation of the mean by the short method presented on p. 86. In that method we guess at a mean and calculate the sum of the deviations. If this sum turns out to be zero, we know that our guessed mean coincides with the actual mean. If the sum of the deviations turns out, as it usually does, to be other than zero, we adjust our guessed mean to the point where the sum of the deviations will equal zero.

though they were all positive or zero. The formula would be read as follows.

"The average deviation is equal to the sum of the absolute deviations of the items from their average, divided by the number of cases."

The average deviation can be computed by another method which is somewhat better adapted to computation on calculating machines. This method[1] is summarized by the following formula:

$$AD = \frac{2(B\bar{X} - b)}{N}$$

where B = number of measures below the mean.

b = sum of the items below the mean.

These two symbols are not used in this sense in other statistical formulas, and need not be remembered except insofar as they apply to this specific problem.

The problem of computing the average deviation from data grouped in frequency tables is simple. It can be done by a so-called "short method"; but in the case of the average deviation the time saved by this short method is not large and the method itself is so complicated that it would not pay to master it unless one were doing a good deal of work with the average deviation. We shall confine ourselves here to an exposition of the "long method," the theory of which is easier to follow, and shall leave the interested student to acquaint himself with the other method if he desires.[2]

When we compute the average deviation by the long method, we determine the amount of deviation for the items of each class on the assumption that the items are concentrated at the class

[1] Based on TRUMAN KELLEY, "Statistical Method," pp. 70–75, The Macmillan Company, New York, 1924. The notation is changed in this presentation.

[2] For expositions of the "short" method see, for example: Secrist, "An Introduction to Statistical Methods," pp. 342*ff.*, The Macmillan Company, New York, 1929; Davies and Crowder, "Methods of Statistical Analysis in the Social Sciences," John Wiley & Sons, Inc., New York, 1933; Garrett, "Statistics in Psychology and Education," pp. 32*ff.*, Longmans, Green & Company, New York, 1926; Mills, "Statistical Methods Applied to Economics and Business," pp. 152*ff.*, Henry Holt & Company, New York, 1924; Chaddock, "Principles and Methods of Statistics," pp. 156*ff.*, Houghton Mifflin Company, Boston, 1925.

mark. We then find the average of these deviations, just as we find the average of any values that are grouped in a frequency table. This means, of course, that we must start out by finding the value of the mean in order to find next the amounts of the deviations from the mean.

Let us determine the average deviation of the heights of Harvard students (see Table 6.1). The class marks and the frequencies which we have used before appear in the first two columns. Then we determine how far each class mark is from

TABLE 6.1.—COMPUTATION OF AVERAGE DEVIATION OF HEIGHTS OF HARVARD STUDENTS

Class Mark (X)	Frequency (f)	Deviation from $\bar{X}$ (x)	fx
156	4	−19.335	− 77.340
159	8	−16.335	−130.680
162	26	−13.335	−346.710
165	53	−10.335	−547.755
168	89	− 7.335	−652.815
171	146	− 4.335	−632.910
174	188	− 1.335	−250.980
177	181	1.665	+301.365
180	125	4.665	583.125
183	92	7.665	705.180
186	60	10.665	639.900
189	22	13.665	300.630
192	4	16.665	66.660
195	1	19.665	19.665
198	1	22.665	22.665
Total (neglect signs).....			5278.380

the true mean. We have discovered (see page 87) that the mean height of these students is 175.335 cm. If the 188 students in the class whose class mark is 174 cm. are to be considered as being concentrated at the class mark, each of them has a height of 174 cm. Each of them, that is, falls short of the mean by 1.335 cm., the value which appears opposite this class in the third column. Each figure in the third column shows the deviation of the corresponding class mark from the mean, which is 175.335 cm. In other words, we subtract the true mean from the class mark to find the figures of this column.

If 188 items differ from the mean by an amount of 1.335 cm. each, they deviate a total of 188×1.335 cm. $= 250.98$ cm. This is the figure that appears opposite the class in the fourth column. The figures in the fourth column show for their respective classes the total amount of the deviation when all items in such classes are considered. These figures are the products obtained by multiplying together the figures opposite them in the second and third columns.

If each figure in the last column shows the total amount of deviation for the class in question, then the sum of this column (taken without regard to signs) will show the total amount of deviation in the whole distribution. In this case the total deviation of the 1000 items is 5278.380 cm. The average deviation is found, of course, by dividing this total deviation by the number of cases. Hence, if $\Sigma fx = 5278.380$ and $N = 1000$,

$$AD = \frac{\Sigma|fx|}{N} = \frac{5278.380}{1000} = 5.27838$$

The average deviation, then is 5.3 cm. or 5 cm., depending on the amount to which we round it off.

To summarize the steps involved in computing the average deviation from frequency tables:

1. Compute the mean.
2. Compute the deviation of each class mark from the mean by subtracting the mean from the class mark.
3. Multiply the frequency of each class by the deviation of its class mark from the mean.
4. Add the products just obtained, neglecting signs.
5. Divide the sum just obtained by N.

The summary in shorthand form is

$$AD = \frac{\Sigma|fx|}{N}$$

What is meant when we say that the average deviation of student heights was 5 cm.? It means that the students measured varied in height. Some were above and some below the mean; some were near the mean in height and some were far from it. But these students differed from the mean an average of 5 cm. If, in another group of people, the average deviation of heights was 7 cm., we should say that they varied more on the average

than did the Harvard group. In general, the larger the average deviation, the greater is the dispersion within the group.

It can be demonstrated[1] that the average deviation is smaller when computed from the median than when computed from any other point. Without a rigorous demonstration, we point out that the sum of the distances from any two points to any point between them is constant, and less than the sum of the distances to any point not between them. Since an equal number of cases lie above and below the median, the cases can be paired and the deviations will be smaller in sum than deviations from any point other than this.[2] Now the fact that the sum of the absolute deviations is smaller when taken around the median than when taken around any other value in a good reason for basing the average deviation on the median rather than on the mean. Sometimes this is done. In the vast majority of cases, however, the *AD* is based on the mean, and if any other base is used the fact should be stated. In our illustrative examples here we have based our computations on the mean. The variations which would be involved in basing the measure on the median are obvious.

6.5. The Standard Deviation.—The *standard deviation*, or *root-mean-square deviation*, is by far the commonest and most useful measure of dispersion in technical work. The range, we have seen, is unstable on account of its dependence on items whose size is largely a matter of chance. The semi-interquartile range arbitrarily excludes half of the items from consideration. The average deviation neglects the fact that some deviations are negative and some positive, and it treats them all as positive. Although the average deviation is an extremely useful measure of dispersion and is easily explained to the layman, nevertheless the neglect of the signs of the deviations makes this measure of dispersion almost useless in further mathematical work. We desire some measure of variation which escapes these several faults, and to a considerable extent the standard deviation does so.

In scientific work the standard deviation is always represented by the small Greek letter sigma (σ), and it is so commonly used

[1] See, for example, KELLEY, *op. cit.*, p. 74.

[2] See LOVITT and HOTZCLAW, "Statistics," p. 109, Prentice-Hall, Inc., New York, 1929.

that the statistician forms the habit of reading the symbol σ as "standard deviation" rather than as "sigma." In some instances those who are more familiar with Greek than with statistics go to the other extreme of using the word "sigma" when they mean "standard deviation." But if someone says that the "sigma" of a distribution is 14, it is safe to interpret his statement to mean that the standard deviation of the distribution is 14. At any rate it would commonly be written

$$\sigma = 14$$

The standard deviation, like the average deviation, is based on deviations from the mean. Of course, it was this base that got us into trouble in the case of the average deviation, since we had to neglect the signs of some of the deviations. In computing the standard deviation, however, we get around this difficulty by taking the quadratic mean of the deviations rather than their arithmetic mean. The student will recall that when we take a quadratic mean we square all the original figures. This makes them all positive, and we need not neglect signs. The standard deviation, then, is the square root of the arithmetic mean of the squares of the deviations. This description of it alone is enough so that the student should be able to go ahead with the computation by himself. We give, however, examples of the computation.

TABLE 6.2.—COMPUTATION OF STANDARD DEVIATION

X	x	x^2
5	−4.9	24.01
8	−1.9	3.61
13	+3.1	9.61
12	+2.1	4.41
11.5	+1.6	2.56
49.5		44.20

6.6. Standard Deviation: Ungrouped Data.—When data are not grouped, we proceed exactly in accordance with the directions given above. Suppose we take the following items:

5; 8; 13; 12; 11.5

The total of these five items is 49.5, and their average is

$$\frac{49.5}{5} = 9.9$$

Let us compute their standard deviation (see Table 6.2). In the second column is given the difference between each value and the average, and the squares of these differences appear in the third column. We have thus found the sum of the squared deviations to be

$$\Sigma(x^2) = 44.20$$

Since there are five deviations, the average of the squared deviations is found by dividing by 5, thus:

$$\frac{\Sigma(x^2)}{N} = \frac{44.20}{5} = 8.84$$

And the square root of the average is

$$\sigma = \sqrt{\frac{\Sigma(x^2)}{N}} = \sqrt{8.84} = 2.97$$

The standard deviation, then, is 2.97. The process of finding it can be summarized thus:

1. Find the mean.
2. Find for each item the deviation from the mean by subtracting the mean from the item.
3. Square these deviations.
4. Add the squares just obtained.
5. Divide the sum just obtained by N.
6. Take the square root of the quotient just found.

Or, if we want the directions in shorthand form,

$$\sigma = \sqrt{\frac{\Sigma(x)^2}{N}}$$

This method of finding the standard deviation from ungrouped data is correct, but another method is usually somewhat shorter in application and gives exactly the same results. The directions for this preferred method are

1. Square the original figures.
2. Add these squares.
3. Divide this sum by N.

4. Subtract from this quotient the square of the mean.
5. Take the square root of the difference.

The formula in this case becomes[1]

$$\sigma = \sqrt{\frac{\Sigma(X^2)}{N} - \bar{X}^2}$$

TABLE 6.3.—COMPUTATION OF STANDARD DEVIATION

X	X^2
5	25
8	64
13	169
12	144
11.5	132.25
49.5	534.25

$$\Sigma(X^2) = 534.25$$

$$\frac{\Sigma(X^2)}{N} = 106.85$$

$$\bar{X}^2 = 9.9^2 = 98.01$$

$$\frac{\Sigma(X^2)}{N} - \bar{X}^2 = 106.85 - 98.01 = 8.84$$

$$\sigma = \sqrt{8.84} = 2.97$$

[1] The equivalence of these two formulas for σ may be seen from the following:

$$\sigma = \sqrt{\frac{\Sigma(x^2)}{N}}$$

But, since x is the deviation of X from the mean, we have

$$x = X - \bar{X}$$

$$\sigma = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N}}$$

$$= \sqrt{\frac{\Sigma(X^2 - 2X\bar{X} + \bar{X}^2)}{N}}$$

$$= \sqrt{\frac{\Sigma X^2 - 2\bar{X}\Sigma X + N\bar{X}^2}{N}}$$

$$= \sqrt{\frac{\Sigma X^2}{N} - 2\bar{X}^2 + \bar{X}^2}$$

$$= \sqrt{\frac{\Sigma(X^2)}{N} - \bar{X}^2}$$

If we illustrate with the same five values that we used before, the process becomes as shown in Table 6.3, page 138. This is precisely the same answer that we found before. The work of computation involved here seems as long as that of the earlier method. For such a short example, and for one in which the mean happens to be a number with but two digits, it is perhaps as long. But let the student try the ordinary example, in which the mean turns out to be a number with 5 to 50 decimals, such as 12.396724. Try taking the deviation of each item from such a mean. Try squaring these deviations. Try adding the squares and taking the square root. In such a case any method which involves merely the squaring of the original figures without the taking of deviations is a blessing. It will be noted that capital X's rather than small x's are used in the second formula to indicate that it is based on the original values rather than on the deviations of these values from the mean.

6.7. Standard Deviation: Grouped Data.—When we have data in frequency tables, we can compute the standard deviation by either the long or the short method. In this case the short method lives up to its name and is a considerable timesaver. Below the long method is explained first, so that the student may see the reasons for the steps involved. We then illustrate and explain the short method as applied to the same data, so that the student may see where the savings in time are made. In order that we shall have all measures of dispersion on the same data for purposes of comparison, we illustrate again with the frequency table showing the heights of Harvard students. The pertinent parts of the table appear again in Table 6.4 with other information which is now needed.

It is necessary first to compute the mean of the heights. We have discovered earlier that the mean height is 175.335 cm. (see page 87). Hence we state our class marks at the left of the table, and in the second column we state each class mark as a deviation from the mean. For example, the first class contains items whose values are 156 cm. (under our assumption that these items are concentrated at the class mark). But 156 cm. falls short of the mean by 19.335 cm. This value is listed in the second column. Opposite each other class mark is listed also its deviation from the mean, found by subtracting

the mean from the class mark. These figures show for each class the amount by which each item in the class deviates from the mean.

TABLE 6.4.—COMPUTATION OF STANDARD DEVIATION FROM GROUPED DATA (LONG METHOD)

Class Mark (X)	Deviation from Mean (x)	x^2	f	fx^2
156	−19.335	373.842225	4	1,495.368900
159	−16.335	266.832225	8	2,134.657800
162	−13.335	177.822225	26	4,623.377850
165	−10.335	106.812225	53	5,661.047925
168	− 7.335	53.802225	89	4,788.398025
171	− 4.335	18.792225	146	2,743.664850
174	− 1.335	1.782225	188	335.058300
177	1.665	2.772225	181	501.772725
180	4.665	21.762225	125	2,720.278125
183	7.665	58.752225	92	5,405.204700
186	10.665	113.742225	60	6,824.533500
189	13.665	186.732225	22	4,108.108950
192	16.665	277.722225	4	1,110.888900
195	19.665	386.712225	1	386.712225
198	22.665	513.702225	1	513.702225
Totals.....			1000	43,352.774600

We have seen that the standard deviation is based on the squares of such deviations; hence the third column shows the squares of the values in the second column. In other words, the square of the deviation for each item in the class is stated opposite each class. Then follows a column, with which we are familiar, showing the number of cases in each class. In the first class there are four items, and the squared deviation of each is 373.842225. Hence the total squared deviation of these four items is $4 \times 373.842225 = 1495.368900$. Similarly we find the total of the squared deviations for the other classes by multiplying the squared deviation of column 3 by the number of cases listed in column 4. This gives us the last column, which is headed fx^2. (In adding this column to get the sum of the squared deviations for the distribution, it is not necessary to neglect signs, since all the values became positive

when we squared the values of column 2 to get the values in column 3.) The sum of the last column, then, gives us the sum of the squared deviations for the distribution. We discovered earlier that we must now divide this sum by the number of cases and take the square root of the quotient. These operations give

$$\frac{43{,}352.774600}{1000} = 43.3527746$$

$$\sigma = \sqrt{43.3527746} = 6.584$$

We have thus found the standard deviation of the heights of students to be 6.584 cm. Rounding it off, we have 6.6 cm. or 7 cm.[1]

TABLE 6.5.—COMPUTATION OF STANDARD DEVIATION FROM GROUPED DATA (SHORT METHOD)

Class Mark (X)	Frequency (f)	Class Deviation (d)	fd	fd^2
156	4	−7	−28	196
159	8	−6	−48	288
162	26	−5	−130	650
165	53	−4	−212	848
168	89	−3	−267	801
171	146	−2	−292	584
174	188	−1	−188	188
177	181	0	0	0
180	125	1	125	125
183	92	2	184	368
186	60	3	180	540
189	22	4	88	352
192	4	5	20	100
195	1	6	6	36
198	1	7	7	49
Totals.....	1000		−555	5125

The process through which we have just derived the standard deviation is tedious and time-consuming. Fortunately an alternative process is quick and easy. This short method of

[1] In the reference from which these figures are taken we are told that $\sigma = 6.56$ cm. The reader will recall (see p. 85*n*.) that our figure for the mean also differed from that of the original reference.

computing the standard deviation will now be explained, the same data being used for purposes of comparison. The short process is much like the short process of discovering the mean of grouped data, and, since we are using the same data which we then used, it may pay the student to review the section explaining that process in conjunction with the present exposition. We repeat here the necessary figures on height and add those data which are necessary to the computation of the standard deviation.

The student will remember that the computation of the mean by the short method was based on the practice of guessing at the mean, taking deviations from the guessed mean in units of the class interval, and making the necessary adjustment to compensate for the error in the guessed mean. In the short method for computing the standard deviation we follow a parallel procedure, and in our illustration we shift our guessed mean one class from its former position so that the student may see that the position of the guessed mean is of no importance (save as it minimizes work if it is near the large frequencies).

If the mean had not already been computed from these data we could compute it now, although it is not necessary.[1] In the short method we proceed directly to the computation of the standard deviation itself. Having listed the class marks and the frequencies as before, we next guess at a mean, selecting always one of the class marks near the center of the distribution where the frequencies are large. In this case we have assumed that the mean is 177 cm. We have then, in the third column, stated the deviations in class intervals from the assumed mean. The first class is seven classes below the assumed mean, so we label it -7; the next class is -6; etc. Our figure for the first class means that each of the four items in that class is seven classes below the assumed mean; the figure for the second class means that each of the eight items in that class is six classes below the assumed mean; etc. We now multiply these class deviations by the frequencies in their respective classes. Each

[1] Substituting the values from the table into the formula for the mean (p. 89), we get

$$\bar{X} = 177 + 3\left(\frac{-555}{1000}\right) = 177 - 1.665 = 175.335$$

This is the identical result that we found when we took the guessed mean at another point.

of the four items in the first class is -7 deviations from the mean, so that this class totals $-7 \times 4 = -28$ class deviations from the mean. Similarly for the other figures in the fourth column. It is important that the computer keep track of signs in this work, for each class whose class mark is smaller than the assumed mean has a negative deviation.

Finally we get the last, or fifth, column in the table by multiplying the figures in the fourth column by the figures in the third (or the d) column. Since these figures are already the product of f and d, and since we now multiply them by d again, they are equal to fd^2. It is evident that this second multiplication by d will make all the signs positive, so that we neglect no signs.

We now add the three columns headed f, fd, and fd^2. These totals are needed for the computation of the standard deviation. The total of the f column we know already is $\Sigma f = N = 1000$. In adding the column of fd's to get $\Sigma(fd)$, it is important to keep track of the signs. In this case we find that $\Sigma(fd) = -555$. We also find that $\Sigma(fd^2) = 5125$. To find the standard deviation from these figures we go through the following steps:

1. Divide $\Sigma(fd^2)$ by N.
2. Divide $\Sigma(fd)$ by N, and square the quotient.
3. Subtract the second result from the first.
4. Take the square root of this difference.
5. Multiply the square root by the class interval.

In formula form this is

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

Substituting the values of our present problem, we have

$$\begin{aligned}\sigma &= 3\sqrt{\frac{5125}{1000} - \left(\frac{-555}{1000}\right)^2} \\ &= 3\sqrt{5.125 - 0.308025} = 3\sqrt{4.816975} \\ &= 3(2.194) = 6.582\end{aligned}$$

Thus we find that the standard deviation of the students' heights is 6.582 cm. Comparison with the answer obtained by the long method shows a discrepancy in the third decimal place: this results from the fact that we have dropped decimal places.

In fact, we drop more decimal places in the long than in the short method, since in the former there are more decimal places to drop.

The notation used in the formula for the short method is the same as that already used in computing the mean (see page 89).

6.8. Checking Accuracy of Computations.—We noted in Sec. 5.4 (see page 90) that there are ways in which the statistician is able to check the accuracy of his arithmetic in some computations. We applied such a method in the case of the arithmetic mean. The Charlier check can also be applied in the case of the standard deviation. Table 6.6 is exactly the same as Table 6.5 except that a new column has been added at the extreme right. This new column contains values of $f(d+1)^2$. To find these values, we add the number 1 to each value in the column headed (d). This gives us values of $(d+1)$. We square these values and multiply the squares by the frequencies in the column headed (f). For example, the top figure in the (d) column is -7. If we add 1 we get -6. This value squared gives us 36. We multiply 36 by 4 (the value of f) to get 144, the first figure in the new last column. Similarly the fifth figure from the end of the column (960) is found by adding 1 to the value of d to get $3+1=4$, squaring to get 16, and multiplying by 60, the frequency, to get 960.

Having obtained the numbers in the last column, we add them, getting a total of 5015. We next apply the Charlier check, which consists of the equation

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

This means that the sum of the last column should be equal to the sum of the (f) column plus the sum of the (fd^2) column plus twice the sum of the (fd) column.[1] In our example $\Sigma[f(d+1)^2]$ is 5015; $\Sigma(fd^2)$ is 5125; $\Sigma(fd)$ is -555; and Σf is 1000. Substituting these values in the Charlier equation, we get

$$\begin{aligned} 5015 &= 5125 + 2(-555) + 1000 \\ 5015 &= 5125 - 1110 + 1000 \\ 5015 &= 5015 \end{aligned}$$

This proves that our arithmetical work was correct.

We shall discover in a later chapter that our standard deviation computed from a frequency table is inaccurate for another reason—namely, because of our assumption that the items within any given frequency class are all equal to the class mark of that class. This assumption involves relatively little error when one is computing the arithmetic mean, but it involves a biassed error, always in the same direction, when one computes the standard

[1] The proof is simple.

$$\begin{aligned} (d+1)^2 &= d^2 + 2d + 1 \\ f(d+1)^2 &= fd^2 + 2fd + f \\ \Sigma[f(d+1)^2] &= \Sigma(fd^2) + 2\Sigma(fd) + \Sigma f \end{aligned}$$

deviation, and always gives a value for the standard deviation which is too large. This error is discussed in Sec. 8.4, page 195. There it will be seen that the error is usually a very small one, and that the answer obtained by the methods discussed here is reasonably dependable.

TABLE 6.6.—CHARLIER CHECK FOR THE STANDARD DEVIATION—SHORT METHOD

Class Mark (X)	Frequency (f)	Class Deviation (d)	(fd)	(fd^2)	$f(d+1)^2$
156	4	−7	− 28	196	144
159	8	−6	− 48	288	200
162	26	−5	−130	650	416
165	53	−4	−212	848	477
168	89	−3	−267	801	356
171	146	−2	−292	584	146
174	188	−1	−188	188	0
177	181	0	0	0	181
180	125	1	125	125	500
183	92	2	184	368	828
186	60	3	180	540	960
189	22	4	88	352	550
192	4	5	20	100	144
195	1	6	6	36	49
198	1	7	7	49	64
Totals.....	1000		−555	5125	5015

6.9. Meaning of the Standard Deviation.—We have found that the standard deviation of the heights of 1000 Harvard students is 6.6 cm. (rounded off from 6.582 cm). As in the other measures of dispersion, the larger the value of the standard deviation, the less closely grouped are the items. A large standard deviation means that the items are widely scattered. Under ordinary circumstances the range, the semi-interquartile range, the average deviation, and the standard deviation differ in size. The semi-interquartile range is usually the smallest, followed by the average deviation, the standard deviation, and the range, in the order named. In those cases where we have what is called a "normal" distribution[1] the sizes of the measures of dispersion bear a definite and known relationship, and in these distributions the

[1] See Chap. VII for a description of the normal distribution and for a more complete description of the interrelationships of the measures of dispersion.

semi-interquartile range is about two-thirds of the standard deviation and the average deviation is about four-fifths of the standard deviation. (More exact values are $Q = 0.6745\sigma$; $AD = 0.7979\sigma$.) If we compare the measures which we have now computed for students' heights, we find the following:

$$Q = 4.44 \text{ cm. (page 129)}$$
$$AD = 5.28 \text{ cm. (page 134)}$$
$$\sigma = 6.58 \text{ cm. (page 141)}$$

It will be noted that the values appear in the order we have just indicated. Moreover, we see that in this case

$$Q = \left(\frac{4.44}{6.58}\right)\sigma = 0.6746\sigma$$
$$AD = \left(\frac{5.28}{6.58}\right)\sigma = 0.803\sigma$$

Thus while these measures do not have exactly the relative size that they would have in a normal distribution, they have approximately that relative size.

It is also true that in a normal distribution about two-thirds of all the items in the distribution[1] will fall within one standard deviation, and practically all the items within three standard deviations of the mean. (We have seen that 50 per cent of the items fall within Q of the mean, and in a normal distribution 57.5 per cent of the items fall within AD of the mean. This gives us another basis for comparing these measures of dispersion.) Thus if the heights of the Harvard men are normally distributed, we should expect that two-thirds of them have heights between $175.335 + 6.582$ and $175.335 - 6.582$ cm; that is, between the points which lie at a distance of one standard deviation on each side of the mean. In our problem this will mean between 181.917 cm. and 168.753 cm. The entire range of a distribution will ordinarily, then, lie within the three standard deviations above and the three standard deviations below the arithmetic mean—an over-all distance of six standard deviations. We discussed in Sec. 3.15 the problems involved in deciding how many classes to use in a frequency table, and how large the class

[1] Actually 68.27 per cent of the cases will fall within 1σ and 99.7 per cent within 3σ. See Chap. VII for a more complete discussion.

interval should be. Fisher states[1] that, while grouping in frequency classes brings perforce some inaccuracy, nevertheless the error in estimating values from a normal population will be less than 1 per cent if the class interval does not exceed one quarter of a standard deviation. If we think of the entire distribution as being spread over six standard deviations, with a class interval of one quarter of a standard deviation, we see that this rule would require the use of approximately 24 classes to include the bulk of the cases in many distributions. In practice, however, the number of classes is seldom so large.

We can, then, interpret the standard deviation in this way. When we are told that the standard deviation of heights is 6.6 cm., we know that the dispersion is less than it would be in a group where $\sigma = 10$ cm. and more than in a group where $\sigma = 2$ cm. We know that, if the distribution of heights is about normal, approximately two-thirds of the items in the group will be within one standard deviation of the mean, or, in this case, within 6.6 cm. of the mean. We know that practically all the cases will be within three standard deviations, or 19.8 cm., of the mean. Practically never will we find a height less than 175.335 − 19.8, and practically never one more than 175.335 + 19.8. An inspection of the original data on heights will show that these statements on extremes of height hold good in this distribution.

6.10. Variance.—In advanced statistical work a great deal of use is made of what is called the *variance* of a distribution. The variance is the square of the standard deviation. If we represent it by the small letter v we can define it thus:

$$v = \sigma^2$$

We saw in Sec. 5.16 that it is possible to combine a number of distributions, and to compute the arithmetic mean of the combined groups on the basis of the arithmetic means of the subgroups. Similarly it is possible, when we combine a number of subgroups, to compute the variance of the combined group on the basis of the variances of the subgroups. Let n_1 be the number of cases in the first subgroup and n_2 be the number of cases in the second subgroup, with N the number of cases in the combined group ($N = n_1 + n_2$.) Let $\bar{X}_1$ be the arithmetic mean of the items in the first subgroup, $\bar{X}_2$ in the second subgroup, and $\bar{X}$ in the combined groups. Let d_1 be the difference between the arithmetic mean of the first subgroup and the arithmetic mean of the combined groups ($d_1 = \bar{X}_1 - \bar{X}$), and let d_2 be the corresponding

[1] R. A. Fisher, "Statistical Methods for Research Workers," 3d ed., p. 50, Oliver & Boyd, Edinburgh, 1932.

difference between the mean of the second subgroup and the mean of the combined groups. Then we have the relationship

$$v = \frac{n_1 v_1 + n_2 v_2 + n_1 d_1^2 + n_1 d_2^2}{N}$$

where v_1 and v_2 are the variances of the first and the second subgroups and v is the variance of the combined group. It will be seen that the variance of the large group is the weighted arithmetic mean of the variances of the subgroups plus the weighted arithmetic mean of the squared differences between the averages of the subgroups and the large group. This can be put in another form to show that the variance of the large group is the sum of two parts.

1. The weighted arithmetic mean of the variances of the subgroups.
2. The variance of the means of the subgroups themselves.

This fact is extremely important in the *analysis of variance*, one of the most powerful of the recently perfected statistical tools. The subject is too advanced for us to take up in an elementary textbook, but it ties in directly with what we have been studying here about the standard deviation.

Before we leave the subject, let us illustrate the computation of the standard deviation on a major group from the data on the subgroups. Suppose we have in a given school 72 boys with an average height of 68 in. and a standard deviation of 3 in. In the same school are 38 girls with an average height of 61 in. and a standard deviation of 2 in. What is the standard deviation in the heights of all 110 people in the school? We find that the average height for all the people in the school is 65.58 in. (using the method explained in Sec. 5.16). The average for the boys exceeds this by 2.42 in., while the average of the girls falls below it by 4.58 in. We therefore substitute in the formula as follows (remembering that if the standard deviation of the boys' heights is 3 in. the variance is 9 in., etc.):

$$v = \frac{72(9) + 38(4) + 72(2.42^2) + 38(-4.58^2)}{110}$$

Carrying out the required computations, we find that $v = 18.35$ or that $\sigma = \sqrt{18.35} = 4.28$ in. Thus we know that if we throw the two groups together the standard deviation of the combined groups will be 4.28 in. We can carry this process out to any number of subgroups, merely adding in our numerator for any new group, x, the values $n_x v_x$ and $n_x d_x^2$, and increasing our denominator to include the sum of all the cases in all the subgroups. We could give the directions for any number of subgroups as follows:[1]

1. Multiply the variance in each subgroup by the number of cases in that subgroup.
2. Add these products for all subgroups.
3. Square the difference between the mean of each subgroup and the mean of the large group; then multiply this square for each subgroup by the number of cases in the group.

[1] For proof, see John F. Kenney, "Mathematics of Statistics," pp. 95–97, D. Van Nostrand Company, Inc., New York, 1939.

4. Add these products for all subgroups.
5. Add the sums found in steps 2 and 3 above.
6. Divide the sum in step 5 by the total number of cases in all groups. The quotient will be the variance of the large group. Its square root will be the standard deviation of the large group.

6.11. Measurement of Relative Dispersion.—The measures of dispersion which we have treated are called "absolute" measures of dispersion. The results are expressed in the same units as the original data; that is, the standard deviation is 6.6 *centimeters* or 7.4 *dollars* or 534 *foot-pounds*. The standard deviation is expressed (as are the other measures of dispersion as well) in the units in which the X values were originally stated. There is nothing in the answer to show whether the standard deviation is large or small. We might well have two distributions with the same standard deviation, say a standard deviation of 1 ft., and yet in the one case this might be a very large dispersion and in the other case a very small one. How is this possible?

Suppose we illustrate. Imagine that we measure the lengths of the main-line track of the railroad systems of the United States. We find the length of each line and then compute the standard deviation in the lengths. A standard deviation of 1 ft. would be unbelievably small. It would mean that a considerable number of the railroads were within 1 ft. of the average length, and that almost no railroad differed from the average length by more than 1 yd. Suppose, on the other hand, that we measure the lengths of the noses of 500 college seniors and find a standard deviation of 1 ft. Is this large or small? It is obviously large. It means that we might expect about one-third of our seniors to have noses which differed from the average length by as much as 1 ft.! In both of these cases the standard deviation is 1 ft., yet in one case it is unbelievably small and in the other impossibly large. This example illustrates the fact that the absolute size of the measure of dispersion does not tell us in itself whether the dispersion is large or small.

But if these measures of dispersion cannot tell us what we want to know, how can we find out? Let us set up another problem. Suppose you were told that some keeper of a zoo had weighed, at one time or another, 150 newborn black bears. The weights were analyzed, and it was found that there was a standard deviation in the weights of ½ lb. Is this a large or a

small dispersion? Suppose we were told that the standard deviation in the weight of newborn babies is also ½ lb.[1] Would you think that the dispersion in the weights of the bear cubs was greater or smaller than that in the weights of the babies? In other words, the question becomes a relative one. Is a ½-lb. dispersion *relatively* large or *relatively* small? To what shall we relate the dispersion?

In practice we use as measures of relative dispersion a comparison between the mean and the measure of dispersion. A standard deviation of 1 ft. in lengths of railroad systems is small *when compared with the average length of railroad systems;* but a standard deviation of 1 ft. in the length of noses is large *when compared with the average length of noses.* If we are to know whether a standard deviation of ½ lb. in weights at birth is large or small, we must know the average weight at birth. It is said that the average male human weighs about 7.5 lb. at birth[2] and the average black bear cub comes into the world weighing about 10.5 oz.[3] Thus if the young of these two animals have the same dispersion in weights, the human babies are *relatively* much less variable than the bear cubs.

The simplest and most obvious method of stating a measure of dispersion in relative terms which compare it with the mean is to state it as a percentage of the mean. This is the way in which all measures of relative dispersion are computed. We have discovered that the semi-interquartile range of student heights is 4.44 cm. (page 129). The mean height is 175.335 cm. (page 87). If we wish to compute relative dispersion based on the semi-interquartile range for this distribution, we get it in this manner:

$$\frac{100(Q)}{\bar{X}} = \frac{444}{175.335} = 2.53 \text{ per cent}$$

Measures of relative dispersion are always given in percentage terms and always show the percentage which the measure of absolute dispersion is of the average. The average used is

[1] Neither of the standard deviations given in this paragraph is based on actual figures. Both are hypothetical.

[2] L. E. Holt, "Care and Feeding of Children," p. 33, D. Appleton-Century Company, Inc., New York, 1928.

[3] E. T. Seton, "Lives of Game Animals," Vol. II, Pt. I, p. 174, Doubleday, Doran & Company, Inc., New York, 1929.

almost always the mean; if any other average is used it should be specified.

Any measure of absolute dispersion can be converted into a measure of relative dispersion by stating it as a percentage of the mean. The formula would be

$$\frac{100 \text{ (absolute dispersion)}}{\text{Average}} = \text{relative dispersion}$$

A large relative dispersion does not mean that the values are widely scattered absolutely, but that they are widely scattered as compared with the mean.

Although any measure of dispersion can be used in conjunction with any average in the computing of relative dispersion, statisticians in fact almost always use the standard deviation as their measure of dispersion (see page 135) and the arithmetic mean as their average. When the relative dispersion is stated in terms of the arithmetic mean and the standard deviation, the resulting percentage is known as the *coefficient of variation*, or the *coefficient of variability*. This coefficient is symbolized by the letter V, defined thus:

$$V = \frac{100\sigma}{\bar{X}}$$

If we take the hypothetical cases of bear and human weights which we used above, we can now compute the coefficients of variation:

Human babies:	Bear cubs:
Mean weight = 7.5 lb.	Mean = 0.656 lb.
σ of weights = 0.5 lb.	σ = 0.5 lb.
$V = \frac{50}{7.6} = 6.7$ per cent	V = 76.2 per cent

By comparing the two coefficients of variation, we discover that bear cubs are (in this hypothetical case) relatively much more variable in weights at birth than are human babies, even though their absolute variabilities are identical.

For a final illustration let us compute from the problem we have been studying the coefficient of variation of student heights. Here we have

$$\bar{X} = 175.335 \text{ cm. (page 87)}$$
$$\sigma = 6.582 \text{ cm. (page 143)}$$
$$V = \frac{658.2}{175.335} = 3.75 \text{ per cent}$$

Even now one does not know whether a standard deviation which is 3.75 per cent of the mean shows a large or a small scatter. One can judge this only by comparing it with other scatters. In order that the student may get some idea of the usual sizes of the coefficient of variation, a table is presented on page 153 which shows these measures for a large number of types of data.[1]

In commenting on these figures Warren and Pearson say: "In the biological field, coefficients of variability above 30 are not common. In the economic field, such a low variability is very uncommon."[2] The student will note that Pearl's figure for the variability of stature ($V = 3.60$ per cent) is approximately the same as the figure we found for the variability of heights of Harvard students (3.75 per cent).

Coefficients of relative (rather than absolute) variability are used when:

1. The series to be compared are stated in different and noncomparable units. For example, if the standard deviation of heights is 6.6 cm. and the standard deviation of weights is 11.9 kg., which represents the greater variability? We cannot compare centimeters and kilograms. But we can say that the coefficient of variation in height is 3.75 per cent and in weight is 18.1 per cent.[3] This comparison would show considerably more variability in weight than in height, at least in this group of students.

2. The series, although stated in the same units, differ so in their average magnitudes that we should ordinarily expect much more absolute variation in the one than in the other. We have pointed out that one should expect more variation in the lengths of railroads than in the lengths of noses, even though both are measured in the same units (feet).

6.12. Suggestions for Further Reading.—A good mathematical treatment of the problems involved in dispersion is found in John F. Kenney, "Mathematics of Statistics," Chap. V, D. Van Nostrand Company, Inc., New York, 1939. George R. Davies and Walter F. Crowder, in their "Methods of

[1] Raymond Pearl, "Medical Biometry and Statistics," pp. 347*ff*., W. B. Saunders Company, Philadelphia, 1930.

[2] Warren and Pearson, in *Farm Economics*, No. 34, p. 456, Cornell University, Ithaca, N.Y.

[3] Castle, from whom the figures are quoted, gives these figures for mean weight and σ of weights, but gives $V = 11.9$ per cent. This is an obvious error. See Castle, "Genetics and Eugenics," pp. 61 and 65, Harvard University Press, Cambridge, 1916.

I. Coefficients of Variation in Man[1]

	Per Cent
Visual acuity	39.12
Weight of healthy spleen	38.21
Keenness of sight	28.68
Strength of grip, right hand	25.93
Body weights (Bavarian)	21.32
Intelligence quotient	18.01
Respiration rate	17.80
Weight of healthy heart	17.71
Breathing capacity	16.6
Auditory acuity	15.84
Pulse rate per minute	14.89
Body weights (American)	13.16
Mouth breadth (American)	8.69
Chest circumference	8.45
Length of forearm	5.24
Length of foot (English)	4.59
Stature (American)	3.60
Oral temperature	0.49

II. Coefficients of Variation on 680 Illinois Dairy Farms, 1912[2]

	Per Cent
Profits	456.4
Labor income	190.7
Number of men hired by the year	104.4
Value of machinery per farm	79.0
Yield of timothy hay	48.7
Number in family over 16 years of age	46.0
Size of farm	42.5
Yield of corn per acre	38.3
Number of horses per farm	35.0
Crop acres per man	34.7
Yield of oats per acre	34.0
Taxes per dollar of capital	27.9
Age of farm operator	25.2
Value of land per acre	25.1
Per cent of milk produced in the winter months	17.2

[1] These figures are selected from a large group of such coefficients compiled by Raymond Pearl and published in his "Medical Biometry and Statistics," pp. 347*ff.*, W. B. Saunders Company, Philadelphia, 1930. For details and further examples see this book.

[2] These figures, based on a study of farms in Illinois, are published by Warren and Pearson in *Farm Economics*, No. 34, p. 456, Cornell University, Ithaca, New York. Many other coefficients are given there.

Statistical Analysis in the Social Sciences," John Wiley & Sons, Inc., New York, 1933, discuss variations in the computation of the standard deviation which are skewed in logarithmic form. Truman L. Kelley, in his "Statistical Method," The Macmillan Company, New York, 1924, described certain theoretical advantages in using the range between the 10th and the 90th percentiles as a measure of dispersion. For a discussion of the analysis of variance, the student is referred to George W. Snedecor, "The Analysis of Variance" and "Statistical Methods," Collegiate Press, Ames, Iowa, 1934 and 1937; R. A. Fisher, "Statistical Methods for Research Workers, Oliver & Boyd, Edinburgh, 1938; or D. D. Paterson, "Statistical Technique in Agricultural Research," McGraw-Hill Book Company, Inc., New York, 1939.

EXERCISES

1. Compute the standard deviation and the coefficient of variation of the wages given in Table 5.9, page 124.

2. Measurements of 1017 freshmen women at Hollins College from 1920 to 1927 show that the mean height was 63.86 in. and the standard deviation of heights was 2.09 in. The mean weight of these same students was 115.65 lb., with a standard deviation of 15.78 lb. Compute the two coefficients of variation. Were these students more variable in height or in weight? Were they more or less variable in height than the Harvard students?[1]

3. A group of 100 selected Smith students averaged 163.8 cm. in height, with a coefficient of variation of 3.3 per cent.[2] What was the standard deviation in their heights?

4. A study of 129 mothers showed that the average age of the mother at the time her first child was born was 23.9 years. The standard deviation in ages was 5.39 years.[3] What was the coefficient of variation? Was there more or less variation in mothers' ages at the birth of first-born than in heights of Harvard students?

5. The average number of offspring in 55 completed families was 3.55. The standard deviation was 1.79. What was the coefficient of variation?[4]

6. A study of 22,498 divorces which took place in Wisconsin from 1887 to 1906 shows that the average duration of the marriage which preceded the divorce was 10.37 years. The standard deviation was 8.39 years. The corresponding figures for the 2,651 divorces of 1929 were $\bar{X} = 9.83$ years and $\sigma = 8.26$ years. Had there been an increase or a decrease in the variability of marriage duration?[5]

[1] Data from PALMER, Physical Measurement of Hollins Freshmen, *Journal of The American Statistical Association*, Vol. 24, No. 165, March, 1929, pp. 42–45.

[2] PALMER, *op. cit.*, p. 42.

[3] CONRAD and JONES, Field Study of Differential Birth Rate, *Journal of the American Statistical Association*, Vol. 27, No. 178, June, 1932, p. 158.

[4] *Ibid.*

[5] YOUNG and DEDRICK, Variation of Duration of Marriages Which End in Divorce, *Journal of the American Statistical Association*, Vol. 27, No. 178, June, 1932, p. 161.

7. A group of men were tested with respect to the strength of grip in their right hands. The average was 48.9 kg., and the standard deviation was 1.94 kg.[1] Compute the coefficient of variability.

8. Apply the Charlier check to your computation of the standard deviation in Exercise 1.

9. The 85 girls who entered Hollins college in 1920 had an average height of 63.24 in. with a standard deviation of 2.35 in. The 125 girls who entered in 1921 had an average height of 63.74 in. with a standard deviation of 1.77 in.[2] What was the standard deviation in the entire group of 210 girls for the two years combined?

[1] BENEDICT *et al.*, Human Vitality and Efficiency under Prolonged Restricted Diet, *Carnegie Institution Publication* 280, p. 583.

[2] PALMER, *loc. cit.*

CHAPTER VII

SIMPLE PROBABILITY AND THE NORMAL CURVE

7.1. Probability.—Suppose that you have a bag in which there are 25 white balls and 75 black balls. Suppose that the balls are well mixed, and you draw one ball at random from the bag. What is the probability that the ball selected will be white? There are evidently 25 chances that you will be successful and 75 chances that you will fail, or 100 chances in all. If we let s represent the number of ways in which you can succeed and f the number of ways in which you can fail, and if these ways are equally likely, then we say that the probability of success is

$$\frac{s}{s+f} = \frac{s}{n}$$

and the probability of failure is

$$\frac{f}{s+f} = \frac{f}{n}$$

In our illustration the probability of success would be

$$\frac{s}{n} = \frac{25}{100} = 0.25$$

and the probability of failure would be $f/n = {}^{75}\!/_{100} = 0.75$. In other words the probability of the occurrence of an event is the relative number of times which we would expect it to occur in an infinitely large number of trials.

The probability of success is usually symbolized by the letter p, and the probability of failure by the letter q. It should be obvious that

$$p + q = \frac{s}{s+f} + \frac{f}{s+f}$$
$$= \frac{s+f}{s+f} = 1$$

In other words, the probability that an event will either happen or fail to happen is represented by the figure 1, which therefore

stands for absolute certainty. Impossibility would be represented by the figure 0. Chances between absolute certainty and impossibility would be represented by some decimal between 0 and 1. It is also evident that if we know either p or q the value of the other can be calculated at once from the relationship $p + q = 1$.

We have illustrated the probability concept with a case (the drawing of balls from a bag) in which one can reason out the probable results without experiment. To be sure, the reasoning depends on the past experience of the reasoner, and to this extent it would be incorrect to say that the result is based on reason rather than on experience. But it is true that one can come to some conclusions with regard to probabilities in such cases without carrying out experiments for the specific purpose of measuring the probability. In such cases, where we state the probability as a product of our reasoning, we call the result the *a priori probability*.

In statistical work we have little contact with problems involving a priori probability except in those cases where we are deriving and illustrating theory. Most actual statistical problems are so complicated that no one can reason out the expected results. For example, what is the probability that a child under one year of age who has whooping cough will recover? No amount of reasoning will tell us the answer. There are too many variables involved, and their relationships are too obscure. In such cases we fall back on the experience which we have had with the problem. The Minnesota State Department of Health stated that 50.5 per cent of children under one year of age recover from whooping cough and 49.5 per cent die.[1] Thus we can say that the probability of recovery is 50.5 cases out of 100, or 50.5/100, or 0.505. The probability is usually stated in the latter form. The likelihood of failure to recover (death) would similarly be 0.495. These facts would be stated thus:

$$p = 0.505$$
$$q = 0.495$$

Probability of this kind, which is based on records of past performance rather than on pure reasoning, is called *statistical*

[1] Quoted in FAEGRE and ANDERSON, "Child Care and Training," p. 48, University of Minnesota Press, Minneapolis, 1930.

probability or *empirical probability.* One cannot rely on such probability except on the assumption that the past performances which form the basis of calculations were typical, and similar to what can be expected in the future. Thus one would have to be sure, before one used this figure for the probability of recovery from whooping cough, that the figures of past performance on which the estimate of probability is based were records of typical cases. If these figures were taken during an exceptionally severe or unusually light epidemic, or if the children were subjected to some particular type of medical care, or if in any way the cases differed from other cases to which we might wish to apply the probabilities, then these statistical probability figures might lead us astray.

7.2. Mean and Standard Deviation of Probability Data.—If, on the other hand, we can assume that the basic data from which we compute statistical probability are typical, then probability figures will be very useful in the solving of statistical problems. Suppose that an epidemic of whooping cough breaks out in our community, and suppose that we can take the statistical probability worked out from the Minnesota cases ($p = 0.505$) as being applicable to local conditions. There are, let us say, 55 children in the community who are afflicted and who are under one year of age. How many will recover? We cannot tell with certainty, of course; sometimes more will recover and sometimes less. But on the average we should expect that 0.505(55) will recover; that is, the average number of recoveries will be $np = 0.505(55) = 27.775$. In the average occurrence of 55 cases, therefore, we should expect 28 children to recover and 27 to die.

We have, then, a very simple way of finding the average occurrence when the probability is known. If 10 cards are drawn at random from a well-shuffled pack of 52 cards, how many black cards will be among them? Sometimes we shall find more and sometimes less. Table 4.1, page 68, shows that when the experiment was actually tried 102 times, the number of black cards varied from 1 to 10. But what should one expect on the average in such cases? The total number of cards in the pack is 52. Of these the 13 spades and the 13 clubs, making a total of 26 cards, are black. Thus the probability (a priori) of drawing a black card is $^{26}\!/_{52} = 0.5$. We are to draw 10 cards. N, then, is 10. On the average we should expect to draw

$$np = (10)(0.5) = 5 \text{ black cards}$$

A glance at the table on page 68 will show that in these trials the average was very close to 5 black cards out of 10.

But to be told that we should expect 27 children with whooping cough to die and 28 to recover, on the average, under the circumstances previously mentioned, is not enough. We have just seen that one can expect to draw 5 black cards out of 10 *on the average,* but the table also shows that on one of the drawings 10 black cards were drawn. Is it not well within the realm of chance, then, that all the children will recover from whooping cough, or that they will all die? We see that, on the average, the recoveries and deaths will almost balance, but what are the chances of departure from this average?

This is the same question that was raised in the preceding chapter on Dispersion. We saw there (page 126) that we do not by any means obtain a complete description of a frequency distribution from the mean. We need to know also something about the dispersion. In the case of deaths from whooping cough we want to know not only the average number that may be expected to live, but the dispersion of the numbers that will live. We have seen that, for a sample of size n, the average number of successes will be np. It can be demonstrated that the standard deviation of the number of successes will be $\sqrt{npq}$.[1] Thus if we take our most recent example, in which 55 babies were afflicted with whooping cough, we have already seen that on the average 28 of them (27.775) would recover. It is now apparent that the standard deviation of recoveries will be $\sqrt{npq} = \sqrt{(55)(0.505)(0.495)} = \sqrt{13.75} = 3.7$. We can therefore say that in about two-thirds of such cases the number of recoveries will not differ from the average by more than 3.7, and that practically never will the number of recoveries differ from the average by over 3(3.7), or 11.1. This means, then, that, in two-thirds of the cases when 55 babies have whooping cough, between $27.775 + 3.7$ and $27.775 - 3.7$ will recover. Carrying through the computations, we discover that the number of recoveries will run between 24.1 and 31.4 in two-thirds of the cases. The chances are two to one that the number of recoveries will be between 24 and 31. And we have also discovered that

[1] For proof see Richardson, "Introduction to Statistical Analysis," pp. 228–229, Harcourt, Brace and Company, New York, 1934.

one almost never finds a value over 3σ from the mean. Here $3\sigma = 3(3.7) = 11.1$. We should almost never get more recoveries than $27.775 + 11.1$, and almost never fewer than $27.775 - 11.1$ recoveries. Practically, then, the greatest number of recoveries that can reasonably be expected (if these cases are like those on which the statistical probabilities were computed) is 38.9, and the smallest number that can reasonably be expected is 16.7. We now know a great deal more about the likelihood of recoveries than was known when we knew merely that the average outcome would be 28 recoveries and 27 deaths. We shall come back to this problem again at a later point in this chapter.

7.3. Elementary Theorems.—Up to this point we have been talking about the likelihood that one single thing will happen or fail to happen. What are the chances when two or more things are combined? Here we have two or three simple theorems which are demonstrated in every book on elementary algebra. They are merely listed and illustrated here; the student whose memory of them is hazy can refresh his mind from any good algebra.

1. Events are said to be *independent* if the occurrence of one of them does not affect the occurrence of others. They are said to be *dependent* if the occurrence of the others is affected by the occurrence of the one. They are said to be *mutually exclusive* when, if one of them happens on a particular occasion, the other cannot happen.

2. The probability that two or more independent events will all happen on a given occasion is the product of their separate probabilities. Thus, if we toss two pennies the chance that either will come up heads is $\frac{1}{2}$. The probability that both will come up heads is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$.

3. The probability that one or another of several mutually exclusive events will happen on a given occasion is the sum of their separate probabilities. Thus the probability of drawing an ace from a shuffled deck of cards on a single draw is $\frac{4}{52} = \frac{1}{13}$. The chance of drawing a king is likewise $\frac{1}{13}$, and this is also the probability of drawing a queen. What are the chances of drawing an ace *or* a king *or* a queen on a single draw? The probability is the sum of the separate probabilities:

$$\frac{1}{13} + \frac{1}{13} + \frac{1}{13} = \frac{3}{13}$$

If p is the chance of success on any trial, and we make n trials, the probability that the event will occur exactly r times (and fail $n - r$ times) is

$$\frac{n!}{r!(n-r)!}p^r q^{n-r}$$

If we draw a card from a shuffled pack, reinsert it, shuffle, draw a second card, and repeat the process until we have drawn 4 cards in this manner, what is the probability that we shall get exactly 2 black cards in the 4 draws? Substituting in the formula, we get

$$\frac{4!}{2!(2!)}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = \left(\frac{24}{4}\right)\left(\frac{1}{4}\right)\left(\frac{1}{4}\right) = \frac{6}{16} = \frac{3}{8}$$

Three times out of 8 (on the average) we should get exactly 2 black cards in 4 draws.

7.4. Expansion of the Point Binomial.—Suppose we toss a single coin. There are two possible ways for it to fall (excluding the possibility that it will fall on its edge), and these we can symbolize by H for heads and T for tails. The possible results are, then

$$1H \quad 1T$$

If we throw two coins, they both can fall heads (this we can represent by HH); or the first can fall heads and the second tails (HT); or the first can fall tails and the second heads (TH); or both can fall tails (TT). Unless we had the coins numbered or otherwise distinguished, the second and third of these possible occurrences would appear to be identical; that is, we would have two ways in each of which we could get one tail and one head. We could summarize our possible results thus:

$$HH \quad \begin{matrix} HT \\ TH \end{matrix} \quad TT$$

Or, to put them in another form, we could write

$$1HH + 2HT + 1TT$$

If we throw three coins, the possible results are (using similar symbols)

$$HHH \quad \begin{matrix} HHT \\ HTH \\ THH \end{matrix} \quad \begin{matrix} HTT \\ THT \\ TTH \end{matrix} \quad TTT$$

In the other form this would become (if H^2T means 2 heads and 1 tail) $1H^3 + 3H^2T + 3HT^2 + T^3$. With four coins the possibilities are

HHHH	*HHHT*	*HHTT*	*HTTT*	*TTTT*
	HHTH	*HTTH*	*THTT*	
	HTHH	*TTHH*	*TTHT*	
	THHH	*THHT*	*TTTH*	
		HTHT		
		THTH		

That is, the results are

$$1H^4 + 4H^3T + 6H^2T^2 + 4HT^3 + 1H^4$$

Finally, if we try the experiment with five coins we discover these possibilities:

HHHHH	*HHHHT*	*HHHTT*	*HHTTT*	*HTTTT*	*TTTTT*
	HHHTH	*HHTTH*	*HTHTT*	*THTTT*	
	HHTHH	*HTTHH*	*HTTHT*	*TTHTT*	
	HTHHH	*TTHHH*	*HTTTH*	*TTTHT*	
	THHHH	*HHTHT*	*THHTT*	*TTTTH*	
		HTHTH	*THTHT*		
		THTHH	*THTTH*		
		THHHT	*TTHHT*		
		THHTH	*TTHTH*		
		HTHHT	*TTTHH*		

This becomes $H^5 + 5H^4T + 10H^3T^2 + 10H^2T^3 + 5HT^4 + 1T^5$.

The observing reader will note that the summary formulas which we are obtaining are the same results that would be obtained by raising the binomial $(H + T)$ to higher and higher powers. Thus:

$$
\begin{aligned}
(H + T) &= H + T \\
(H + T)^2 &= H^2 + 2HT + T^2 \\
(H + T)^3 &= H^3 + 3H^2T + 3HT^2 + T^3 \\
(H + T)^4 &= H^4 + 4H^3T + 6H^2T^2 + 4HT^3 + T^4 \\
\text{etc.}
\end{aligned}
$$

Thus by expanding the binomial we can get the same results at once that we would get from long experiment.

Elementary books on algebra give rules for the expansion of the binomial to higher powers.[1] By following these rules one obtains

[1] See, for example, Fite, "College Algebra," p. 150, D. C. Heath and Company, Boston, 1913; Rietz and Crathorne, "College Algebra," p. 93,

the proper coefficients and exponents for any power of the binomial.

Perhaps the simplest of these rules is the following:
To find the terms of the expansion of $(q + p)^n$:
a. The first term is q^n.
b. The second term is $nq^{n-1}p$.
c. In each succeeding term the power of q is reduced by 1 and the power of p is increased by 1.
d. The coefficient of any term is found by multiplying the coefficient of the preceding term by the power of q in that preceding term, and dividing the product so obtained by one more than the power of p in that preceding term.

Example:

$$(q + p)^6 = q^6 + 6q^5p + 15q^4p^2 + 20q^3p^3 + 15q^2p^4 + 6qp^5 + p^6$$

We notice that, in accordance with rule *a*, the first term is q^n or q^6. We notice that, in accordance with rule *b*, the second term is $nq^{n-1}p$ or $6q^5p$. The third term finds the power of q reduced by 1 and the power of p increased by 1 to give p^4q^2, and the coefficient is found in accordance with rule *d*; namely, we multiply the coefficient of the preceding term (6) by its power of q (5) and divide by one more than the power of $p(1 + 1 = 2)$ to get $6(5)/2 = 15$.

We can also get these results quickly from Pascal's arithmetical triangle, part of which is given in Table 7.1. It will be noted

TABLE 7.1.—COEFFICIENTS OF THE BINOMIAL EXPANSION

Number of Terms	Coefficients of Binomial Expansion							
1	1							
2	1	1						
3	1	2	1					
4	1	3	3	1				
5	1	4	6	4	1			
6	1	5	10	10	5	1		
7	1	6	15	20	15	6	1	
8	1	7	21	35	35	21	7	1

Henry Holt and Company, New York, 1919; WILCZYNSKI and SLAUGHT, p. 142, "College Algebra," Allyn & Bacon, Boston, 1916; GRIFFIN, "Introduction to Mathematical Analysis," p. 431, Houghton Mifflin Company, Boston, 1921.

that the figures in this table proceed in accordance with a definite rule. The first column consists of nothing but 1's. The second column is the arithmetical progression 1, 2, 3, 4, . . . , and starts at the second row. Each number in the table is the sum of the number above it and the number to the left of that number. In other words, we add to a given number the number at its left and put the sum below the given number in the triangle. In the next-to-last row of the table, for example, appears the number 20. It is found by adding the number above it (10) and the number to the left of that number (10). Note that the rows in this triangle give the coefficients of the coin-tossing experiment. There is always one more term in the expanded binomial than the number of coins tossed (or the number of equally likely independent events). With two coins there are three possible occurrences: two heads, one head, or no heads. Hence we look for the row in the table with one more term than the number of coins. We note that the expansion with three terms has the coefficients 1, 2, and 1. Thus we know that the relative number of occurrences of the possible outcomes of tossing two coins are: two heads once, one head twice, and no heads once. To be sure, these results would be experienced only *in the long run.*

It will be noted that as we add more and more terms to the binomial expansion [that is, as we raise $(H + T)$ to higher and higher powers], we continue to have values which are small toward the extremes, get larger and larger as we approach the center, and exhibit absolute symmetry. If we raise the binomial to the 14th power, giving 15 terms, they are 1, 14, 91, 364, 1001, 2002, 3003, 3432, 3003, 2002, 1001, 364, 91, 14, and 1. If we plot these on a frequency graph, we get the chart shown in Fig. 7.1. It will be noted that the chart exhibits absolute symmetry and regularity, and that there is a peak of high frequencies at the center from which the frequencies fall away toward the ends. The slope of the curve is at first gentle as we leave the peak, gets steeper and steeper for a time, and then slowly tends to level out. It never becomes quite level, but as it approaches the base line it becomes more and more nearly so.

7.5. The Normal Curve.—If the binomial were raised to an infinitely high power, the number of coefficients would become infinitely large, and the short straight lines of Fig. 7.1 would merge into a continuous, smooth curve. This curve, which is

the limit approached as the binomial is raised to higher and higher powers, is called the *normal curve of error*, or more usually merely the *normal curve*. It is likewise variously called the Gaussian curve, the Laplacian curve, the probability curve, and the normal distribution curve. The general expansion of

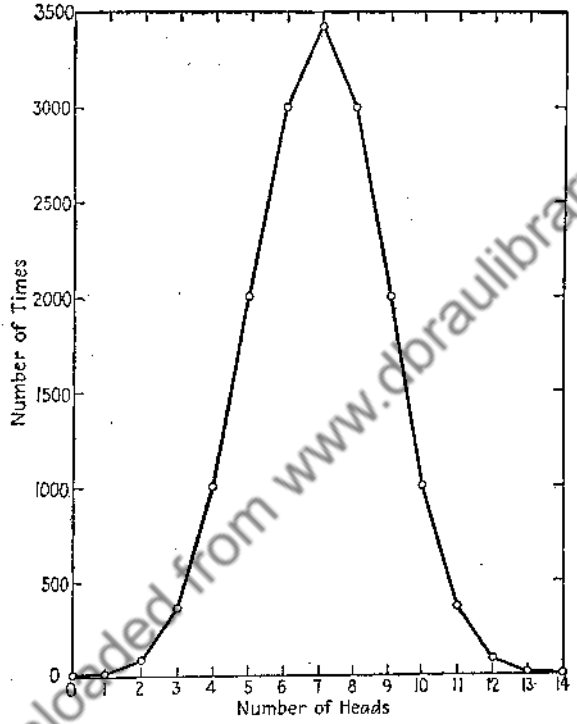


FIG. 7.1.—Coefficients of $(a+b)^{14}$, giving the numbers of times that various numbers of heads would be expected to appear in 16,384 throws of 14 coins.

$(p+q)^n$ is called the *point binomial*, and in the special case where $p = q = \frac{1}{2}$ and n is infinitely large we get the normal curve. In other words, the normal curve is a special case of the point binomial which we have when an infinitely large number of forces are operating, each of which is equally likely to happen or to fail.

It has been found in practice that the point binomial describes tolerably well many natural occurrences. It has been found especially that many phenomena of biology, economics, psychology, education, etc., even though not exactly normal in

distribution, can be described roughly by the normal curve or some other point-binomial curve. To be sure, one seldom meets an actual distribution that is exactly symmetrical or is exactly normal in any other way—but likewise one seldom sees a trend that is perfectly described by a straight line or by a second-degree parabola. The normal curve is found in practice to be a convenient method of smoothing out chance irregularities which occur in a frequency distribution, without departing in too great a degree from the underlying characteristics of the original data.

We have already noted the fact that many frequency distributions tend to have small numbers of cases near the extremes and many cases toward the center (see page 39). The heights of Harvard students, with which we have now become so familiar, were so distributed. It is common for such data as physical measurements to be so arranged. In fact, this type of distribution is so common that some people have come to look on it as normal and call it the "normal distribution." It should be emphasized that in most statistical problems there is no a priori reason for expecting normality of distribution—no reason for believing in advance that the data will be distributed as are the coefficients of the expansion $(\frac{1}{2} + \frac{1}{2})^n$. But so many groups of data are distributed in this manner that the characteristics of such a distribution become especially important. It becomes worth our while to study this "normal distribution" so that we shall know what it is like. Then, in those many cases where the binomial expansion does approximately describe the data, we shall know better how to handle the problem. In more advanced statistical work, other forms of the point binomial become important (cases where $p \neq q$), but we shall confine ourselves in this chapter to a discussion of the most important case, that where $p = q = \frac{1}{2}$, which we call the normal curve.

First let us describe the normal curve. It is pictured in Fig. 7.2. It should be noted that it is entirely symmetrical bilaterally. There is a high point exactly at the center, and the heights (frequencies) grow less and less toward the extremes. The slope grows steeper and steeper for a time as we progress toward the ends, and then the slope becomes less and less. We say technically that there is a "point of inflection" on each side of the curve—that is, a point where the slope ceases to become

steeper and begins to become more gradual. Students of the calculus will recognize this peculiarity better if we say that there is a change in the sign of the second derivative of the curve at these points. If we let the height of the curve be represented by y, and distances along the horizontal axis measured from the

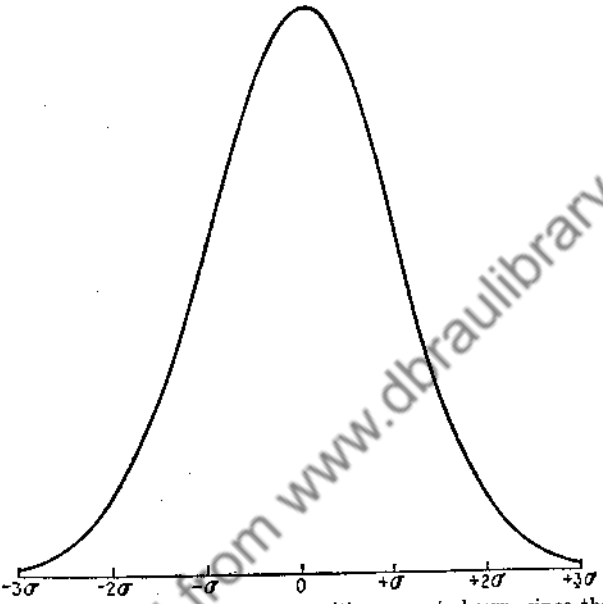


FIG. 7.2.—The normal curve. The extremities are not shown, since the curve continues in either direction indefinitely.

mean of the X's be represented by x (that is, x is a deviation from the mean), the mathematical equation of the curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

In this equation σ represents the standard deviation of the X's, π is the ratio of the circumference of a circle to its diameter, and e is the basis of the Napierian system of logarithms and is equal to approximately 2.71828. This curve is asymptotic at the base; that is, it approaches closer and closer to the base line, but never quite reaches it. The horizontal distance from the center of the curve (which represents the mean, the median, and the mode) to either point of inflection is equal to one standard deviation. If we drop from the points of inflection lines perpendicular to the

base, these two lines, the base line, and the curve will enclose 68.27 per cent of the entire area under the curve. Perpendiculars erected at twice this distance from the mean (that is, a distance of 2σ) will, together with the base and the curve, enclose 95.5 per cent of the area under the curve. If the perpendiculars are moved to points which are 3σ each side of the mean, the area referred to will be 99.7 per cent of the total area under the curve. It is on the basis of these facts that the statements on page 146, relative to the interpretation of the standard deviation, were made.

7.6. Areas under the Normal Curve.—It is possible to compute the percentage of the total area under the curve which will be cut

TABLE 7.2.—RELATIVE AREAS UNDER THE NORMAL CURVE BETWEEN THE MEAN AND VARIOUS NUMBERS OF STANDARD DEVIATIONS

Tenths of a σ	Whole Standard Deviations			
	0	1	2	3
0	0.0000	0.3414	0.4773	0.4986
1	0.0398	0.3643	0.4821	0.4990
2	0.0793	0.3849	0.4861	0.4993
3	0.1179	0.4032	0.4893	0.4995
4	0.1554	0.4192	0.4918	0.4997
5	0.1915	0.4332	0.4938	0.4998
6	0.2258	0.4452	0.4953	0.4998
7	0.2580	0.4554	0.4965	0.4999
8	0.2881	0.4641	0.4974	0.4999
9	0.3159	0.4713	0.4981	0.5000

off by perpendiculars erected at any number of standard deviations from the mean. The number of cases, the mean, and the standard deviation give a complete description of any curve which is really normal, and if we know these three values we can reconstruct the entire curve. This has made it possible to construct tables showing the percentage of the total area which falls within given numbers of standard deviations from the mean. Table 7.2 is a short example of this kind. A somewhat longer one appears in Appendix I (see page 509).

To find the portion of the area under the curve which lies between the mean and any other point we proceed as follows: Suppose we desire to find the portion of the area between the mean and a point which is removed from the mean by 1.7 stand-

ard deviations. We look for the column headed "1 standard deviation," and we look in the row opposite the entry 7 in the left-hand column (which lists tenths of standard deviations). We find the entry 0.4554. This means that 45.54 per cent of the total area of the curve lies between the mean and a point either 1.7σ above the mean or 1.7σ below the mean. Hence 2(45.54) or 91.18 per cent of the area will be within 1.7σ of the mean.

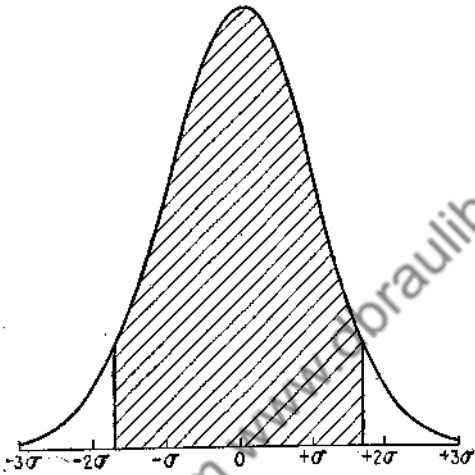


FIG. 7.3.—The normal curve with perpendiculars erected at points 1.7 standard deviations each side of the arithmetic mean. The shaded area, enclosed by the basic line, the perpendiculars, and the curve, is 91 per cent of the total area under the curve.

Since the area of the curve represents the total number of cases in the distribution, we can say that if the values are normally distributed 91 per cent of them will lie within 1.7σ of the mean (see Fig. 7.3).

This table is of great help in the interpretation of statistical conclusions. We shall, therefore, use it to aid in interpreting two more examples.

We discovered (page 87) that the mean height of Harvard students is 175.335 cm. The standard deviation of their heights is 6.6 cm. (page 141). How likely is it that a student chosen at random will exceed 185 cm. in height? We attack this problem thus: The question is, what is the probability that a value will exceed the mean by 9.665 cm.? That is, how likely is it that a value will be as much as $9.665/6.6 = 1.46\sigma$ above the mean? If

the heights are normally distributed, 50 per cent of them will fall short of the mean. And our table tells us that between the mean and a point 1.5σ from the mean will be another 43.32 per cent of the cases. (We could get somewhat more accurate figures from the table in Appendix I, which shows that 42.8 per cent of the area falls between the mean and a point 1.46σ from the mean. We shall use the shorter table here, however, and round off our deviation from 1.46σ to 1.5σ.) Thus if we include all the area from a point 1.5σ above the mean on down, we include $50 + 43.32 = 93.32$ per cent of the cases. We can say, then, that in only 7 per cent of the cases will a student chosen at random exceed a height of 185 cm.

Let us go back to the whooping-cough problem which we met early in this chapter (page 157). We discovered that, when 55 babies less than a year old are afflicted, on the average 27 deaths and 28 recoveries will result. We also discovered that the standard deviation in the number of recoveries is 3.7. How likely is it that as few as 22 babies will recover?

Our procedure is just as before. We shall outline it here.

1. What is the mean? (28 recoveries)
2. What is the standard deviation? (3.7)
3. What is the point about which we want information? (22 recoveries)
4. How far is it from the mean? $(28 - 22 = 6)$
5. How many standard deviations is it from the mean? $(6/3.7 = 1.62)$
6. What per cent of the cases lie between this point and the mean? (44.52 per cent)[1]
7. What per cent of the cases lie the other side of the mean? (Always 50 per cent)
8. This makes a total of what per cent of the cases? $(50 + 44.5 = 94.5$ per cent)
9. How likely is the occurrence mentioned? It will happen in 5.5 per cent of the cases and fail in 94.5 per cent of the cases; that is, in 55 cases out of 1000 we should find fewer than 22 recoveries. In 945 cases out of 1000 we should find more recoveries. (This answer can be found directly from Appendix IV, page 512.)

It is now seen that the standard deviation is an extremely valuable measure in conjunction with a frequency distribution *if the distribution is normal or approximately so.*[2] Under such

[1] Taken from the table on p. 168 and as 1.6σ from the mean. Actually there are 44.7 per cent of the cases between the mean and 1.62σ (see Appendix I, p. 509).

[2] The figures here given are for perfectly normal distributions, but Salvosa

circumstances we can tell what percentage of the total cases will fall within given numbers of standard deviations from the mean. It must be remembered that deviations are always measured in units of the standard deviation.

This fact—that all normal curves can be described in such terms—makes it possible to compare some measures which could not otherwise be compared. We give but one example, but others will quickly suggest themselves. Suppose that John scores 127 on a test on which the average score is 112 and the standard deviation of scores is 15. Robert scores 98 on a test on which the average score is 90 and the standard deviation is 3. Who makes the better score?

It is immediately obvious that we cannot say that John makes the better score, merely because his score was higher. He took the easier test, as is shown by the fact that the average score was higher. We note next that John is 15 points above the average for his test, while Robert is but 8 points above the average for the test that he took. But again we cannot say that this proves that John is better, because there was a great deal more variation on John's test than on Robert's: the standard deviation is much higher. We must find out how far each deviates from the mean in *standard units*, that is, in units of the standard deviation. If we do this we find that John is $^{15}/_{15}\sigma = 1\sigma$ above the mean on his test, and Robert is $^{8}/_{3} = 2.67\sigma$ above average on his test. This shows considerably better performance on Robert's part than on John's. If the distributions of marks are normal, John's mark is exceeded by 16 per cent of those who took his test and Robert's score is exceeded by but 0.35 per cent of those who took his test. The student should verify these figures for himself, using Table 7.2 or Appendix I.

If we wish to find the distance which, when laid off above and below the mean, will include half the area under the curve, we look in the table on page 168 or in the table in Appendix I, page 509. We hunt for the point which, when laid off on one side of the mean, will include 25 per cent of the cases (because, since the curve is symmetrical, this distance *both* sides of the mean will

has published tables similar to these showing areas under the ordinates of curves of varying degrees of asymmetry. See Luis R. Salvosa, Tables of Pearson's Type III Function, *Annals of Mathematical Statistics*, Vol. 1, May, 1930, pp. 191*ff*.

include 2×25 per cent = 50 per cent of the cases). We discover that we need to go between 0.67σ and 0.68σ to reach this point. As a matter of fact, it is necessary to go 0.6745σ from the mean in each direction in order to enclose half of the area. But it is to be remembered that the semi-interquartile range, when laid off on each side of the mean in a symmetrical distribution, includes half the area.[1] We thus see that $Q = 0.6745\sigma$, as we

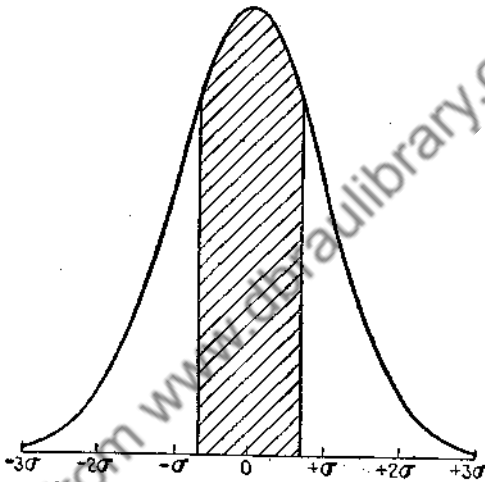


FIG. 7.4.—Perpendiculars erected under the normal curve at distances of 0.6745 standard deviation on each side of the mean. The area enclosed by the base line, the perpendiculars, and the curve is one-half of the total area under the curve.

discovered on page 146. It can also be shown that, when the distribution is normal, $AD = 0.7979\sigma$, as stated on page 146 also. One should remember, then, that a distance equal to about two-thirds of the standard deviation laid off on each side of the mean will include half the cases in a normal distribution (see Fig. 7.4).

These various relationships hold true strictly only when the distribution is exactly normal. It is seldom that empirical data show absolute normality, just as it would be unusual for a man throwing five pennies 32 times to get exactly one case where no heads turned up, 5 cases in which one head turned up, 10 cases of two heads, 10 cases of three heads, 5 cases of four heads, and one case when all the pennies turned up heads. With an infinite

[1] See pp. 128 and 129.

number of throws of five pennies, one should expect these proportions,[1] but in any finite number there might be some deviation from it. Thus also, even though we might get an exactly normal distribution of heights if we had an infinitely large number of cases, when we take a finite number such as 1000 cases we must expect some deviation from normality. Hence one can never interpret the standard deviation exactly as if the data were normal. We get approximations only, and the closeness of the approximation depends on the closeness with which the normal curve describes the data. When, however, the data are approximately normal, we can interpret the standard deviation with a fair degree of exactitude.

7.7. Preliminary Tests for Normality.—How can we discover whether or not a curve is approximately normal? There are many methods. We can group the data in a frequency table and see whether or not there tend to be large frequencies in the central classes and small frequencies in the end classes. We can plot the data in a frequency curve and see whether it looks roughly like the normal curve shown in Fig. 7.2, page 167. We can see if the description of the normal curve given on page 164 seems to fit the data. We can investigate to learn if about 68 per cent of the cases are included within 1σ. We can see if Q is approximately two-thirds of σ. Or perhaps even better, we can plot the ogive of the data on a special sort of graph paper called "probability paper" to see if it "straightens out." We noticed in Sec. 3.14 that the graph of an ogive assumes a typical S-shape when the data are normally distributed. This characteristic S-shape appears in Fig. 5.2, page 96. Yet the S-shape indicates only that the original frequency distribution was mound-shaped, and not necessarily that it was normally distributed. If we convert the data of our ogive into percentage form, and plot them on probability paper, the ogive will turn into a straight line if, and only if, the distribution was normal; and if the distribution is almost, but not quite, normal, the ogive on probability paper will fall almost into a straight line.

The first step in the use of probability paper is to compute the data for a percentage ogive. These computations appear, starting with the data on student heights, in Table 7.3. The first two columns are those of Table 5.5, page 93, except that

[1] See p. 162.

the first column lists the lower class limits rather than the class marks. The third column, found by adding the items in the second column, shows the number of students who had heights greater than those listed in the first column. It will be noticed that our figures in this column start off with 1000, the total number of students, since all the students had heights greater than 154.5 cm. Since, however, there were 4 students with heights between 154.5 and 157.5 cm., there were only 996 whose

TABLE 7.3.—DATA PUT IN FORM FOR PROBABILITY PAPER

Height (centimeters) (class limit)	Number of Students	Number with Greater Heights	Percentage with Greater Heights
154.5	4	1000	100.0
157.5	8	996	99.6
160.5	26	988	98.8
163.5	53	962	96.2
166.5	89	909	90.9
169.5	146	820	82.0
172.5	188	674	67.4
175.5	181	486	48.6
178.5	125	305	30.5
181.5	92	180	18.0
184.5	60	88	8.8
187.5	22	28	2.8
190.5	4	6	0.6
193.5	1	2	0.
196.5	1	1	0.1
199.5	0	0	0.0

heights were greater than 157.5 cm. Starting at the bottom of this third column, we find that no one had a height greater than 199.5 cm., the actual upper limit of the tallest class. But there was one man whose height lay between 196.5 and 199.5, so we list one person taller than 196.5 cm. in the third column. There was also one person whose height was between 193.5 and 196.5 cm. so we have two people taller than 193.5 cm. Any figure in the third column can be found by adding to the number at the left in the second column all the numbers farther down in the second column. The fourth column is found by dividing the third column by the number of cases and then multiplying by 100. In

this case, since the total number of cases is 1000, this can be done easily by pointing off one place.

Now we transfer the data of Table 7.3 to probability paper, as in Fig. 7.5. The vertical lines are evenly spaced, but the hori-

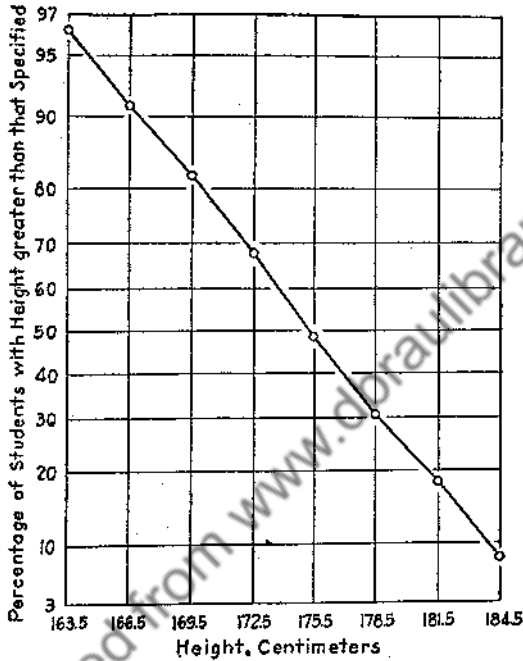


FIG. 7.5.—Ogive plotted on probability paper, showing a distribution which is nearly normal.

zontal lines are bunched closely together in the center and spread farther apart toward the top and the bottom. When we put the data of Table 7.3 on the chart, we find that the points fall almost, although not exactly, along a straight line. Thus we know that the students' heights were distributed almost, but not exactly, in a normal curve.

Probability paper can be purchased from some dealers in draftsman's supplies, but it is easy to make, and since most stores do not carry it, it may be worth while to give here the directions for making it. From the sample in Fig. 7.5, we see that the first thing to do is to lay out the required number of vertical lines, spacing them at convenient equal intervals. Next we locate the line marked 50 per cent, which is at the center of the vertical lines. The other lines are arranged symmetrically around this center line.

TABLE 7.4.—BASIC DATA FOR USE IN CONSTRUCTING PROBABILITY PAPER

Line Number	Units from 50% line	Line Number	Units from 50% Line
50%	0	15.5 or 84.5	1015
49 or 51	25	15.0 or 85.0	1036
48 or 52	50	14.5 or 85.5	1058
47 or 53	75	14.0 or 86.0	1080
46 or 54	100	13.5 or 86.5	1103
45 or 55	126	13.0 or 87.0	1126
44 or 56	151	12.5 or 87.5	1150
43 or 57	176	12.0 or 88.0	1175
42 or 58	202	11.5 or 88.5	1200
41 or 59	228	11.0 or 89.0	1227
40 or 60	253	10.5 or 89.5	1254
39 or 61	279	10.0 or 90.0	1282
38 or 62	305	9.5 or 90.5	1311
36 or 63	332	9.0 or 91.0	1341
36 or 64	358	8.5 or 91.5	1372
35 or 65	385	8.0 or 92.0	1405
34 or 66	412	7.5 or 92.5	1440
33 or 67	440	7.0 or 93.0	1476
32 or 68	468	6.5 or 93.5	1514
31 or 69	496	6.0 or 94.0	1555
30 or 70	524	5.5 or 94.5	1598
29 or 71	553	5.0 or 95.0	1645
28 or 72	583	4.5 or 95.5	1695
27 or 73	613	4.0 or 96.0	1751
26 or 74	643	3.5 or 96.5	1812
25 or 75	674	3.0 or 97.0	1881
24 or 76	706	2.5 or 97.5	1960
32 or 77	739	2.0 or 98.0	2054
22 or 78	772	1.5 or 98.5	2170
21 or 79	806	1.0 or 99.0	2326
20 or 80	842	0.8 or 99.2	2409
19 or 81	878	0.6 or 99.4	2512
18 or 82	915	0.4 or 99.6	2652
17 or 83	954	0.2 or 99.8	2878
16 or 84	994	0.1 or 99.9	3090

That is, the distance to the line marked 30 per cent is the same as the distance to the line marked 70 per cent. The distances are given in Table 7.4. The first column of this table shows the line in question. The second column shows how many units the given line lies above or below the 50 per cent line. For example, the line which represents 75 per cent (and also the line which represents 25 per cent) lies 674 units from the center. These units are entirely arbitrary. Suppose, for example, that we are laying out a piece of probability paper on an ordinary sheet of 8½- by 11-in. notebook paper. We might lay off our vertical lines at half-inch intervals. We might then decide that we wanted our horizontal lines to cover a distance of, say, 7 in. out of the total of 11½ in.; that is, from the top horizontal line to the bottom horizontal line is to be 7 in. The 50 per cent line will be established first, exactly in the center, or 3½ in. from either the top or the bottom. Since the normal curve runs limitless distances either side of the center, we cannot show it all. Suppose we decide to show 98 per cent of the cases, from 1 to 99 per cent, leaving off the two extremes. Then we know that the 3½ in. from our 50 per cent line to the bottom (or to the top) represents the distance to the 1 per cent (or the 99 per cent) line. We look in Table 7.4 and see that these lines lie 2326 units from the center. In other words, we let 3½ in. represent 2326 units. Now if we want to find the 70 per cent line we note that it lies 524 units from the center, or $^{524}/_{2326}$ as far as the 99 per cent line, or $^{524}/_{2326}$ths of 3½ in., or 0.79 in. from the center. Other lines are found similarly. We decide first what line we shall select for our top (or bottom) line—whether 90, 95, or 98 per cent, etc. We note in the table how many units it is from the center. Then we locate the other lines by proportion. In this way we lay out whatever horizontal lines we want, keeping in mind the size of the sheet of paper on which we are working. The student can verify the fact that in the illustrative case just given the 90 per cent line (and the 10 per cent line) will be 1.93 in. from the center.

One of the best criteria is to compute the normal curve which corresponds to the data themselves and to note how it agrees with the data. We have already noted that if we know the number of cases, the mean, and the standard deviation, we can find the normal curve. We can determine these constants from our original data and "fit" a normal curve to them. Let us try this procedure in the problem of the heights of Harvard students.

The normal curve can be fitted by either of two methods. In both we make use of tables that describe a normal curve which has a standard deviation of one class interval and in which the number of cases is one, that is, in the curve described in the tables

$$N = 1$$
$$\sigma = 1 \text{ class interval}$$

This is called a *unit normal curve.* We find from the tables (see

Appendix, pages 509 and 510) the values for such a curve, and then convert these values to fit our particular problem. The procedure will be clearer if we work out examples. We shall use first the method of ordinates and then the method of areas.

7.8. Fitting the Normal Curve: Method of Ordinates.—In any actual problem we should ordinarily start by plotting our data in a frequency curve, letting the ordinates represent frequencies and the abscissas represent the values of X. If we take the case of student heights, for example, the abscissas will represent

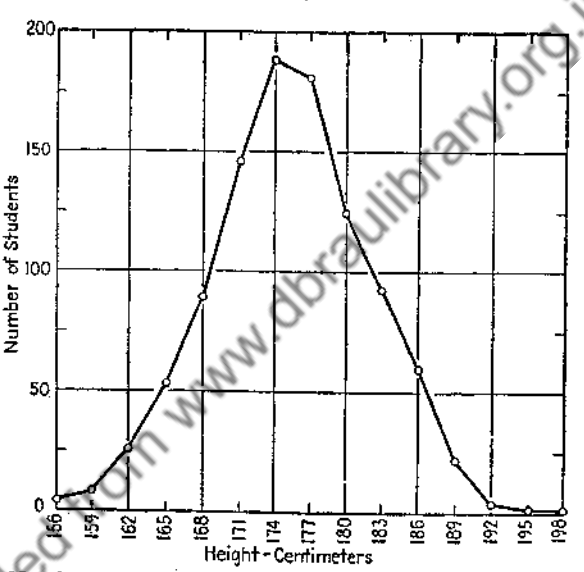


FIG. 7.6.—Numbers of Harvard students between the ages of eighteen and twenty-five years with various heights, 1914–1916.

heights. Such a diagram appears in Fig. 7.6. It will be noted that this diagram does look roughly like the normal curve.

Our next step would be to compute the necessary constants: the mean, the standard deviation, and the number of cases. These we have already computed for other reasons in the case of students' heights. They are

$$N = \Sigma f = 1000$$
$$X = 175.335 \text{ cm. (see page 87)}$$
$$\sigma = 6.582 \text{ cm. (see page 143)}$$

The two latter can be rounded off to 175.3 and 6.6.

Since the normal curve is symmetrical, we know that the arithmetic mean, the median, and the mode will coincide; that is, the highest point on the curve (the mode) will be located at that point on the horizontal axis which represents 175.3 cm. (the mean). This tells us that the maximum ordinate will be located at 175.3 on the horizontal scale, but it does not tell us how high the highest point will be. The maximum ordinate (greatest height) of a normal curve can always be found from the equation

$$y_o = \frac{0.3989(Ci)(N)}{\sigma}$$

where y_o is the magnitude of the maximum ordinate, Ci the class interval, etc.[1] If we substitute the values of our problem, we get

$$y_o = \frac{(0.3989)(3)(1000)}{6.6} = 181.3$$

In other words, our normal curve will reach its maximum height at a point opposite 175.3 on the horizontal scale, and that maximum height will correspond to 181.3 cases on the vertical scale.

We could compute the heights of other points on the curve by further solutions of the general formula,[2] but the use of tables will

[1] We have seen (p. 167) that the formula for the normal curve is

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}$$

We want the value of y when $x = 0$ (since x is the deviation from the mean and we want the value of y at the mean). But if we put $x = 0$ in this formula, we find that

$$e^{-x^2/2\sigma^2} = e^0 = 1$$

Therefore at this particular point (the mode or the mean) the formula becomes

$$y = \frac{N}{\sigma\sqrt{2\pi}} = \frac{0.3989N}{\sigma}$$

Since the deviations from the mean (x) in this equation are in terms of the class interval (Ci), we must multiply our result by the class interval to get an answer in the units of our original problem. This gives the value of y_o given above.

[2] Let us illustrate for one additional case. If we take the general formula for the curve, and let $N = 1$ and $\sigma = 1$ as in the unit normal curve, our formula becomes

$$y = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} = 0.3989e^{-x^2/2}$$

What is the height of the curve at a point one standard deviation from the

save us considerable time. These tables show the height of the unit normal curve at various distances from the mean, and also the percentage of the total area under the curve which lies between a perpendicular erected at the mean and a perpendicular erected at various distances from the mean. We make our computations for the unit normal curve, and then convert the units of our own problem. The necessary tables are given in Appendixes I and II, pages 509 and 510.

Reference to the table of ordinates will show, for example, that the unit normal curve has a height of 0.1295 at a distance of 1.5σ from the mean. In any particular problem we find the height in the original units by multiplying the tabular value by $(N)(Ci)/\sigma$. In the problem we have studied, this means that we must multiply any tabular value by $(1000)(3)/6.6 = 454.5$. At 1.5σ from the mean, therefore, the height will be $0.1295(454.5) = 58.9$.

Usually we are interested in the height of the curve at particular points: the class mid-points. These are the points for which the frequencies are known in the original problem, and we should like to know the theoretical frequencies at these points so that we can compare them. We find these easily, and the computations are summarized in Table 7.5 shown on page 181. The first three columns of this table are taken directly from Table 6.4, page 140, the values in the third column being rounded off. The third column represents the distance of each class mark from the mean, and is found by subtracting the value of the mean (175.3) from each of the figures of the first column. The fourth column states these distances in units of the standard deviation. Since the standard deviation in this problem is 6.6, we get the figures of the fourth column by dividing each figure of the third column by

---

mean; that is, when $x = 1$? Our formula then becomes

$$y = 0.3989e^{-\frac{1}{2}} = \frac{0.3989}{\sqrt{e}} = \frac{0.3989}{\sqrt{2.71828}} = 0.242$$

This tells us that the height of the curve at a distance of one standard deviation from the mean is 0.242 if $N = 1$, $\sigma = 1$, and $Ci = 1$. In any particular problem we must multiply our answer by $(N)(Ci)/\sigma$. Here this is

$$\frac{(1000)(3)}{6.6} = 454.5$$

Carrying out the multiplication, we have $454.5(0.242) = 110$. At one standard deviation from the mean the height is 110 cases.

6.6. The fourth column, then, tells us how many standard deviations from the mean each class mark lies. We now look in the table of ordinates of the unit normal curve (Appendix II, page 510) and find the height of this curve at each of these deviations. These values appear in the fifth column of the table. For example, at a distance of 2.02σ from the mean the unit curve has a height of 0.0519.

TABLE 7.5.—FITTING THE NORMAL CURVE BY ORDINATES

Class Mark (X)	Observed Frequency (f)	Deviation from Mean (x)	Deviation in σ Units (x/σ)	Tabular Value	Computed Frequency (f')
156	4	−19.3	−2.92	0.0056	2.5
159	8	−16.3	−2.47	0.0189	8.6
162	26	−13.3	−2.02	0.0519	23.6
165	53	−10.3	−1.56	0.1182	53.7
168	89	− 7.3	−1.11	0.2155	97.9
171	146	− 4.3	−0.65	0.3230	146.8
174	188	− 1.3	−0.20	0.3910	177.7
177	181	1.7	0.26	0.3857	175.3
180	125	4.7	0.71	0.3101	141.0
183	92	7.7	1.17	0.2012	91.4
186	60	10.7	1.62	0.1074	48.8
189	22	13.7	2.08	0.0459	20.9
192	4	16.7	2.53	0.0163	7.4
195	1	19.7	2.98	0.0047	2.1
198	1	22.7	3.44	0.0011	0.5

Finally we must convert the data of the unit normal curve into the units of our own problem. As we have just discovered, this is done by multiplying each ordinate of the unit curve by $(N)(Ci)/\sigma$. In our problem this means that each figure must be multiplied by (1000)(3)/6.6, or 454.5. The products are entered in the last column of the table. Thus each figure in the last column is the product found by multiplying the corresponding figure in the preceding column by 454.5.

It is now possible to compare the frequencies which actually did occur (in the second column) with those which would have occurred in a corresponding normal distribution. The figures in the last column are those that would be found in a normal distribution whose mean was 175.3, whose standard deviation was 6.6, and

which had 1000 cases. In other words, if we had an exactly normal distribution with mean, dispersion, and number of cases the same as those in our actual distribution, the cases would be distributed as they are in the last column of the table. The cases actually were distributed as in the second column. In the class from 187.5 to 190.49 (the class with a mid-point of 189) we did get 22 cases; in a normal distribution we should expect 20.9 cases.[1] Similarly comparisons may be made at other points.

7.9. Fitting the Normal Curve: Method of Areas.—We shall now fit the normal curve to the same data by the alternative method of areas. In this method we start with the lower limits of our classes rather than with the mid-points. As we discovered in Chap. III, if the class interval is 3 and the mid-point is 156, the lower limit of the class will be 154.5. Similarly we find the lower

TABLE 7.6.—FITTING THE NORMAL CURVE BY AREAS

Lower Class Limit	Observed Frequency (f)	Deviation from Mean	Deviation in σ Units	Per cent of Total Area	
				Below This Class	In This Class
154.5	4	−20.8	−3.15	0.1	0.3
157.5	8	−17.8	−2.70	0.4	0.8
160.5	26	−14.8	−2.24	1.2	2.5
163.5	53	−11.8	−1.79	3.7	5.5
166.5	89	− 8.8	−1.33	9.2	9.5
169.5	146	− 5.8	−0.88	18.7	15.0
172.5	188	− 2.8	−0.42	33.7	17.5
175.5	181	0.2	0.03	51.2	17.2
178.5	125	3.2	0.48	68.4	14.2
181.5	92	6.2	0.94	82.6	9.2
184.5	60	9.2	1.39	91.8	5.0
187.5	22	12.2	1.85	96.8	2.1
190.5	4	15.2	2.30	98.9	0.8
193.5	1	18.2	2.76	99.7	0.2
196.5	1	21.2	3.21	99.9	0.1
199.5	0	24.2	3.67	100.0	

[1] The student may find the decimals of the last column confusing. What do we mean when we say that we should expect 20.9 students in a particular class? Actually if the heights were normally distributed and we took many groups of 1000 students, we should sometimes find 19 students in this height group, sometimes 20, sometimes 21, etc. *On the average* we should find 20.9 students in the class.

limits of the other classes. These are entered as the first column of our summary table.

The figures in the second column are the actual frequencies with which heights occurred in the various classes. In the third column are given the distances of each class lower limit from the mean. These are found, of course, by subtracting 175.3 (the value of the mean) from each of the figures in the first column. In the fourth column these distances are expressed in terms of the standard deviation; that is, each entry in the fourth column is found by dividing the corresponding figure in the third column by the standard deviation (6.6).

The figures in the fifth column are derived from those in the tables of areas under the unit normal curve (see Appendix I, page 509). This table shows the percentage of the total area under a unit normal curve (that is, the percentage of the total number of cases in such a normal distribution) which lies between a perpendicular erected at the mean and another perpendicular erected at any given number of standard deviations from the mean. We discover from this table that 49.9 per cent of the area lies between the mean and a perpendicular erected 3.15σ below the mean. Always, of course, 50 per cent of the area lies above the mean (since the normal curve is symmetrical). Therefore a total of 49.9 per cent + 50 per cent = 99.9 per cent lies above this class limit. We conclude that 0.1 per cent must lie below this class limit. This is the first figure in the fifth column.

Similarly our table tells us that 40.8 per cent of the area lies between the mean and a point 1.33σ below the mean. Therefore 90.8 per cent of the area (50 per cent + 40.8 per cent = 90.8 per cent) must lie above this point and 9.2 per cent must lie below it. This gives us the fifth figure in the fifth column. When we come to the tenth figure in the column, we find that we are 0.94σ above the mean. The table in Appendix I tells us that 32.6 per cent of the area lies between this point and the mean. Since another 50 per cent lies below the mean, we know that 82.6 per cent of the total area (50 per cent + 32.6 per cent = 82.6 per cent) lies below the lower limit of this class. This is the tenth entry in the fifth column. The other figures in the column are found similarly.

Each figure in column 6 is found by subtracting the corresponding figure in column 5 from the figure below it. The column is a column of differences. For example, if we subtract the first figure

in column five (0.1) from the second figure (0.4), we get the first figure in column six (0.3). If we subtract the seventh figure in column 5 from the eighth, we get the seventh figure in column six. (51.2 − 33.7 = 17.5). The reason for this taking of differences is evident after a moment's thought. The first figure in column 5 tells us that 0.1 per cent of the total area lies below the bottom of the first class. The next figure tells us that 0.4 per cent lies below the bottom of the second class. The difference, 0.3 per cent, must lie in the first class. Similarly, if 33.7 per cent of the area lies below the bottom of the seventh class and 51.2 per cent below the bottom of the eighth class, the difference, or 17.5 per cent of the area, must lie in the seventh class.

Since the total number of cases is 1000, it is easy to convert this last column to actual numbers of cases expected. In any such problem we should find the percentages of the total number of cases as listed in the sixth column. In our problem the expected numbers of cases are 3, 8, 25, 55, etc. These expected frequencies will be compared with the actual frequencies shown in the second column. It is evident that the results obtained by the method of areas and those obtained by the method of ordinates

TABLE 7.7.—ACTUAL DISTRIBUTION OF STUDENT'S HEIGHTS COMPARED WITH ESTIMATES OF THE CORRESPONDING NORMAL DISTRIBUTION AS COMPUTED BY THE METHOD OF ORDINATES AND BY THE METHOD OF AREAS

Class Mark	Actual Number Observed	Number Expected by Method of:	
		Ordinates	Areas
156	4	2.5	3
159	8	8.6	8
162	26	23.6	25
165	53	53.7	55
168	89	97.9	95
171	146	146.8	150
174	188	177.7	175
177	181	175.3	172
180	125	141.0	142
183	92	91.4	92
186	60	48.8	50
189	22	20.9	21
192	4	7.4	8
195	1	2.1	2
198	1	0.5	1

are not exactly identical. The two methods seldom show entire agreement. We can compare the actual frequencies with which various heights occurred with those which would have occurred in a normal distribution by arranging the actual figures and the estimates in parallel columns, as shown in Table 7.7 on page 184.

It is obvious that the method of areas might also have been used to estimate to the nearest tenth of a case, as was the method

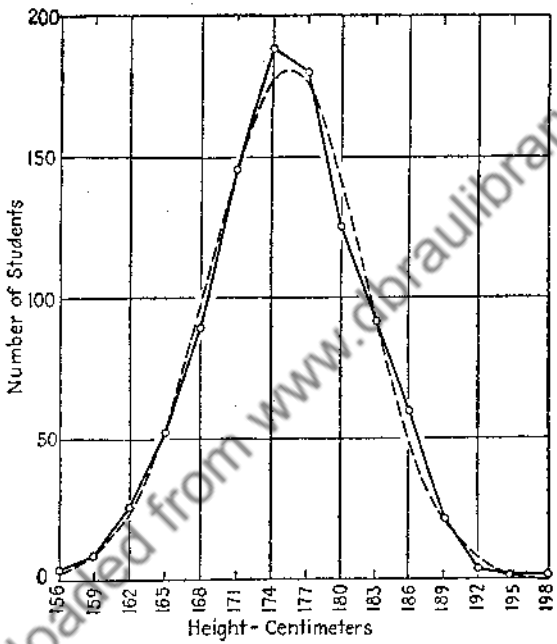


FIG. 7.7.—Distribution of the heights of 1000 Harvard students, showing the actual and the normal distributions. The smooth curve drawn with a broken line is the normal curve fitted by the method of ordinates.

of ordinates. It would have been necessary only to take our figures from the Appendix tables to one more place. However, as we noted in Chap. II, the adding of decimal places would give us only seeming increases in accuracy. Figure 7.7 shows graphically the agreement between the actual frequencies and those expected when we estimate by the method of ordinates.

We have now discovered a number of ways in which data can be tested to see whether or not they are approximately normal in their distribution. In the next chapter we shall study

further, more precise, and more accurate tests for normality, as well as learning something about other types of frequency curves which differ in their characteristics from the normal curve. At the end of that chapter will be found suggestions for further reading on the subject of frequency curves in general, and the references there given (see page 230) can be used by the reader who wishes to pursue further the work of this chapter as well.

EXERCISES

1. Give two original examples of statistical probability; of a priori probability.

2. A baseball player has made 28 hits in 117 times at bat. How many hits is he expected to get in his next 25 times at bat? What are the chances that he will make 10 or more? 3 or less?

3. What is the probability of drawing 3 hearts in succession from a pack of 52 cards, each card drawn being reinserted and the pack being shuffled before the next draw?

4. What is the probability that we shall get exactly 4 heads in 6 throws of a penny? that we shall get 4 heads or more?

5. Using the scheme used on page 162 to show the results of throwing 5 pennies, diagram the possible results when 6 pennies are thrown. Compare your result with that shown in Table 7.1, page 163.

6. Continue Table 7.1, page 163, as it would be if two more rows were added at the bottom.

7. Which is the more general term: "point binomial" or "normal curve"? Distinguish between them.

8. We erect a perpendicular from the base line to a normal curve at a point 1.7σ above the mean, and another 0.6σ below the mean. What per cent of the total area under the curve is within the space bounded by the two perpendiculars, the base line, and the curve? What per cent of the area is above the upper perpendicular?

9. What odds would you offer that a Harvard student chosen at random will be between 185 cm. and 170 cm. tall? Use the results discovered in the text as a basis for your answer.

10. Three students take different tests. *A* gets a score of 72, *B* of 85, and *C* of 17. The average marks received on the three tests are 85, 90, and 25, respectively. The three standard deviations are 7, 2, and 7, respectively. Arrange the three students in order of excellence as you would judge them by these results.

11. In a normal distribution $\bar{X} = 17$ and $\sigma = 3$. What are the values of Q, AD, Q_1, Q_3, Mo., and Med.?

12. Fit a normal curve to the data given in Exercise 3, page 124.

13. Fifty-three of 625 cases of diphtheria in Providence, Rhode Island, in 1915 resulted in death.[1] Let us suppose that this ratio of deaths to total

[1] G. C. Whipple, "Vital Statistics," p. 377, John Wiley & Sons, Inc., New York, 1923.

cases is correct for the universe of diphtheria cases. Suppose we have 300 cases of diphtheria in an epidemic. What are the chances that there will be as many as 27 deaths? If there were 60 deaths, what would you conclude? If there were 15 deaths? If you knew that in the universe one would get an average of 53 deaths in 625 cases, how many deaths could occur in an epidemic of 300 cases before you would rule out chance as the cause of the increased fatalities and decide that the cases must be fundamentally different from those of the universe mentioned?

14. An instructor's records show that he has, in the past, turned in failing grades for 12 of 140 students in elementary statistics. His present class numbers 20. How likely is it that every member of the present class will pass? That as many as 4 will fail? If you are sixth from the bottom of the present class, and if this class is comparable to past ones, what are the chances that you will fail?

15. Make a sheet of probability paper on a sheet of 8½- by 11-in. notebook paper. Put the 98 per cent line at the top and the 2 per cent line at the bottom. Let the distance between these two lines be 10 in. Put in the following horizontal lines: 95, 90, 85, 80, 75, 70, 60, 50, 40, 30, 25, 20, 15, 10, and 5 per cent.

16. Plot on the probability paper made in the preceding exercise, or on a piece furnished, the data of Table 5.9, page 124. Are the wages normally distributed?

17. In Appendix II is a table showing the height of the normal curve at various distances from the mean. Using the data in this table, draw on a sheet of graph paper a picture of the normal curve. Locate the heights of the curve at each fifth of a standard deviation from the center, and connect them by a smooth, freehand curve.

18. Find the median and the quartiles of the heights of students from the diagram of Fig. 7.5, page 175. Compare these answers which were found graphically with the computed answers on pages 94 and 129.

19. Show from their formulas that the standard deviation of a probability distribution is always less than the square root of the arithmetic mean except in the limiting case where it may be as large as the square root of the arithmetic mean.

CHAPTER VIII

MOMENTS, FREQUENCY CURVES, AND THE CHI-SQUARE TEST

In the preceding chapter we have studied what is, perhaps, the most common and most useful of all frequency curves—the so-called "normal curve." We have discovered that this normal curve can be described completely in terms of the number of cases, the arithmetic mean, and the standard deviation. If we are told these three things we can draw the curve, or we can tell what number of cases will fall within any given area under the curve.

So much emphasis is given to the normal curve that the student sometimes draws the conclusion that it is the only frequency curve there is, or, at least, the only important one. We shall see in this chapter, however, that there are many other important frequency curves, and we shall learn that in order to identify and to describe them we usually need some information in addition to the values of N, $\bar{X}$, and σ. While the normal curve can be used to describe with reasonable accuracy a good many distributions, the statistician soon learns that there are also many distributions which differ in character so much from the normal that a normal curve fitted to the data would be misleading rather than informative. In this chapter, we shall study some of these non-normal curves.

8.1. The Higher Moments of a Frequency Distribution.—We have used the symbol x to represent the deviation of any item in a distribution from the arithmetic average of that distribution. We find the value of x by subtracting the value of the arithmetic mean from the value of the item (see pages 131 to 132). The arithmetic means of the various powers of these deviations in any distribution are called the *moments* of the distribution. If we take the mean of the first power of the deviations, we get the first moment about the mean; the mean of the squares of the deviations gives us the second moment about the mean; the

mean of the cubes of the deviations yields the third moment about the mean; etc. These moments can be defined by the following formulas if we let v_1 represent the first moment about the mean, v_2 the second moment, etc.:

$$v_1 = \frac{\Sigma x}{n}$$

$$v_2 = \frac{\Sigma x^2}{n}$$

$$v_3 = \frac{\Sigma x^3}{n}$$

$$v_4 = \frac{\Sigma x^4}{n}$$

etc.

The 0th moment about the mean will, of course, be equal to $\Sigma x^0/n$. But as in the case of each item, regardless of the amount of the deviation, the 0th power of the deviation will equal 1, this is equivalent to $n/n = 1$. In any distribution, then, the 0th moment equals 1. We have discovered also that $\sigma = \sqrt{\Sigma(x^2)/n}$ (page 137). But this is the square root of the second moment about the mean, as will be seen from the formula above. We can thus say that the second moment about the mean $= \sigma^2$. This will be true of any distribution.

When we were studying the average deviation we discovered that in any distribution the sum of the deviations of the items from the mean was equal to zero (page 131). But it will be noted that the formula for the first moment about the mean involves Σx and that it reduces to zero, since $\Sigma x = 0$. We can thus say some things about the moments of all curves in advance:

(1) $v_0 = 1$
(2) $v_1 = 0$
(3) $v_2 = \sigma^2$

There are one or two other things that can be deduced about the moments. If the curve is symmetrical there will be a deviation below the mean which exactly equals each deviation above the mean. This is what we mean by symmetry. If this is true, then these positive deviations and negative deviations will exactly balance each other and when added will cancel out. Of course, if the deviations are raised to even powers their signs will all be

positive, and they will no longer cancel out. But the sums of the odd powers will all be equal to zero on account of the cancellations. We thus know in advance that *in any symmetrical curve* the odd moments, being based on the sums of odd powers of deviations, will equal zero. That is, in symmetrical distributions

$$v_3 = 0$$
$$v_5 = 0$$
$$v_7 = 0$$
etc.

This does *not* hold true in asymmetrical distributions. The rules which we laid down for v_0, v_1, and v_2 hold true for *any* distribution. The rules just enunciated for odd-powered moments above the first hold true only if the distribution is symmetrical. For this reason we can use them, and do use one of them, as measures of asymmetry (see page 204).

8.2. Computation of the Higher Moments.—We could, of course, compute the higher moments directly from their formulas. Since the third moment about the mean is $\Sigma x^3/n$, we could find the deviation of each item from the mean, cube it, and divide by n. Following our earlier practice where data are grouped in frequency tables we should, in such cases, use the formula $\Sigma(fx^3)/n$. But, as before, it pays here to use a short method in which we guess at a mean, take our deviations in units of the class interval, and carry on our computations, finally adjusting our results to take care of the difference between our guessed mean and the true mean. This method has become familiar to us in computing the mean and the standard deviation (pages 86 and 141), and we shall not go into the details of the theory of it here. We shall, however, give an example. Still using our data on the heights of Harvard students, let us compute the third and the fourth moments about the mean. (Higher moments would be handled similarly.) The process is illustrated in Table 8.1.

In adding totals be careful to keep track of signs. The first five columns of this table are taken directly from page 6.5, where we used these same figures in computing the standard deviation. The computation of the figures in the remaining two columns is evident. Each figure in column 6 is the product of the corresponding figures in columns 3 and 5; that is, $(d)(fd^2) = (fd^3)$.

Similarly, each figure in column 7 is the product of the corresponding figures in columns 3 and 6. Thus $(d)(fd^3) = (fd^4)$. Had we not computed the mean and the standard deviation, it is obvious that we could do it directly from the figures given here, since this is the method heretofore used for their computation.

We cannot compute the moments about the mean directly from these figures, since these figures show deviations about an assumed mean. (Here the assumed mean is 177 cm.) Hence we compute first the moments about the assumed mean. Just as we symbolize the assumed mean by $\bar{X}'$ instead of by $\bar{X}$, in order that it may be distinguished from the true mean, so we shall

TABLE 8.1.—COMPUTATION OF THE HIGHER MOMENTS: HEIGHTS OF HARVARD STUDENTS

Class Mark (X)	Frequency (f)	Class Deviation (d)	fd	fd^2	fd^3	fd^4
156	4	−7	− 28	196	−1,372	9,604
159	8	−6	− 48	288	−1,728	10,368
162	26	−5	−130	650	−3,250	16,250
165	53	−4	−212	848	−3,392	13,568
168	89	−3	−267	801	−2,403	7,209
171	146	−2	−292	584	−1,168	2,336
174	188	−1	−188	188	− 188	188
177	181	0	0	0	0	0
180	125	1	125	125	125	125
183	92	2	184	368	736	1,472
186	60	3	180	540	1,620	4,860
189	22	4	88	352	1,408	5,632
192	4	5	20	100	500	2,500
195	1	6	6	36	216	1,296
198	1	7	7	49	343	2,401
Totals....	1,000		−555	5,125	−8,553	77,809

represent the first, second, third, and fourth moments about the assumed mean by v'_1, v'_2, v'_3, and v'_4, respectively, in order that they can be distinguished from the moments about the mean. The formulas for the moments about the assumed mean follow:

$$v'_1 = \frac{\Sigma fd}{n} = \frac{-555}{1000} = -0.555$$

$$v_2' = \frac{\Sigma fd^2}{n} = \frac{5125}{1000} = 5.125$$
$$v_3' = \frac{\Sigma fd^3}{n} = \frac{-8553}{1000} = -8.553$$
$$v_4' = \frac{\Sigma fd^4}{n} = \frac{77{,}809}{1000} = 77.809$$

The general formulas appear at the left, and at the right we have substituted the values found for this particular problem.

Now comes the problem of shifting from the assumed mean to the true mean. The formulas for the moments about the mean in terms of the moments about the assumed mean follow:

$$v_1 = \frac{Ci(\Sigma fd)}{n} - \frac{Ci(v_1')(\Sigma f)}{n} = 0$$
$$v_2 = Ci^2(v_2' - v_1'^2)$$
$$v_3 = Ci^3(v_3' - 3v_2'v_1' + 2v_1'^3)$$
$$v_4 = Ci^4(v_4' - 4v_3'v_1' + 6v_2'v_1'^2 - 3v_1'^4)$$

If we substitute in these equations the values of our problem and solve, we get the following results:

$$v_1 = \frac{3(-555)}{1000} - \frac{3(-0.555)(1000)}{n} = 0$$

The first moment about the mean must always equal zero. It is worth while to substitute the proper values in the equation for v_1 and solve as a check on the arithmetic, since unless a mistake has been made the result must equal zero.

$$v_2 = 9(5.125 - 0.308) = 9(4.817) = 43.353$$

It will be remembered that this is the square of the standard deviation. Had we not computed σ before we should now compute $\sqrt{43.353} = 6.58 = \sigma$. Compare this with the σ found before on page 143. The value of v_2 is always the square of σ.

$$v_3 = 27[-8.553 - 3(5.125)(-0.555) + 2(-0.555^3)] = -9.774$$
$$v_4 = 81[77.809 - 4(-8.553)(-0.555) + 6(5.125)(0.308) - 3(0.094864)] = 5508.567$$

These are the moments of the distribution about the mean. If we gather together our results relative to the distribution of

students' heights, we find

$$v_1 = 0$$
$$v_2 = 43.353$$
$$v_3 = -9.774$$
$$v_4 = 5508.567$$

It is obvious that the curve is not exactly symmetrical, for if it were the value of v_3 would be zero. It is in fact -9.774. But we cannot tell whether this is a large or a small deviation from symmetry merely by the size of v_3. In the case of v_3 we are dealing with the third power of deviations from the average. To judge the degree of asymmetry we must relate v_3 to the standard deviation, and since the deviations are cubed we relate it to the cube of the standard deviation. Similarly we can relate the fourth moment to the fourth power of the standard deviation. The various moments divided by the proper power of the standard deviation give us another group of useful coefficients which we represent by the Greek letter α (alpha). We can define them thus:

$$\alpha_1 = \frac{v_1}{\sigma} = 0$$

$$\alpha_2 = \frac{v_2}{\sigma^2} = 1$$

$$\alpha_3 = \frac{v_3}{\sigma^3}$$

$$\alpha_4 = \frac{v_4}{\sigma^4}$$

etc.

These measures are read as "alpha one," "alpha two," "alpha three," etc.

It can be demonstrated[1] that the values of α_3 and α_4 for the normal curve are always 0 and 3, respectively. Thus we can test the curve of students' heights by computing these constants. The computation follows:

$$\alpha_3 = \frac{v_3}{\sigma^3} = -\frac{9.774}{6.6^3} = -0.034$$

$$\alpha_4 = \frac{v_4}{\sigma^4} = \frac{5508.567}{6.6^4} = 2.926$$

[1] RIETZ *et al.*, "Handbook of Mathematical Statistics," p. 97, Houghton Mifflin Company, Boston, 1924.

If we compare these two figures with those which would have occurred had the heights been normally distributed, we find that this curve is approximately normal.

8.3. Checking Accuracy of Computations.—We have learned earlier (see Secs. 5.4 and 6.8) that it is possible to check the accuracy of arithmetical computations by means of what is called the "Charlier check." This check really consists, as can be seen by looking back to the earlier examples, in choosing another guessed mean as a starting point at the class mark of the next smaller class, so that the values of d are each increased by unity. We use the same general method when computing the higher moments, although it is probably easier to go through the process once, as we did in Table 8.1, and then set up an entirely new second table with a new arbitrary zero point, as in Table 8.2. It will be noticed that the first two columns of Table 8.2 are exact duplicates of the first two columns of Table 8.1, but in the third column each value of d is greater by one than it was in the earlier table. The zero point is taken in the preceding class where the class mark is 174 cm. instead of at 177 cm.

TABLE 8.2.—CHARLIER CHECK FOR ACCURACY OF COMPUTATIONS—THE MOMENTS

Class Mark (X)	Frequency (f)	Class Deviation ($d+1$)	$f(d+1)$	$f(d+1)^2$	$f(d+1)^3$	$f(d+1)^4$
156	4	−6	− 24	144	− 864	5,184
159	8	−5	− 40	200	−1,000	5,000
162	26	−4	−104	416	−1,664	6,656
165	53	−3	−159	477	−1,431	4,293
168	89	−2	−178	356	− 712	1,424
171	146	−1	−146	146	− 146	146
174	188	0				
177	181	1	181	181	181	181
180	125	2	250	500	1,000	2,000
183	92	3	276	828	2,484	7,452
186	60	4	240	960	3,840	15,360
189	22	5	110	550	2,750	13,750
192	4	6	24	144	864	5,184
195	1	7	7	49	343	2,401
198	1	8	8	64	512	4,096
Totals....	1000		445	5015	6,157	73,127

as in Table 8.1. We then go through exactly the same processes which we used in the preceding section. In order to show the connection between the two tables, we label the third column $d + 1$, since each number in it is found by adding one to the corresponding entry in Table 8.1.

We now make use of the following equations, which the student can easily derive for himself after the fashion of the derivations in the footnotes on pages 91 and 144:

$$\begin{aligned}
\Sigma f(d+1) &= \Sigma(fd) + N \\
\Sigma f(d+1)^2 &= \Sigma fd^2 + 2\Sigma fd + N \\
\Sigma f(d+1)^3 &= \Sigma fd^3 + 3\Sigma fd^2 + 3\Sigma fd + N \\
\Sigma f(d+1)^4 &= \Sigma fd^4 + 4\Sigma fd^3 + 6\Sigma fd^2 + 4\Sigma fd + N
\end{aligned}$$

The first two of these equations have been used heretofore in checking our computation of the arithmetic mean and the standard deviation, but are repeated here so that we may have all the customary Charlier equations together. The last two equations are an obvious extension for the third and the fourth powers of $d + 1$. If we substitute the values from Table 8.2 in the left-hand members of these equations, and the values from Table 8.1 in the right-hand members, we get the following:

$$\begin{aligned}
445 &= -555 + 1000 \\
5015 &= 5125 + 2(-555) + 1000 \\
6157 &= -8553 + 3(5125) + 3(-555) + 1000 \\
73{,}127 &= 77{,}809 + 4(-8553) + 6(5125) + 4(-555) + 1000
\end{aligned}$$

Since these four equations all check out when we evaluate the right-hand members, we know that our arithmetical work has been accurate. We could, of course, compute the values of the arithmetic mean, standard deviation, and the alphas quite as well from Table 8.2 as from Table 8.1. The computations in the tables are very rapid, and hence the check consumes relatively little time.

8.4. Grouping Error.—Our computation of the moments has been carried out on the assumption that the items in each class are all concentrated at the mid-point of the class. It would be surprising if the data were really so arranged. If the data are normally distributed they will not be so arranged, and our assumption that they are congregated at the class marks has introduced a slight error which is called the *grouping error*. In

those cases where continuous data have been grouped in frequency tables (especially if the class intervals are large) and where the curve approaches closer and closer to the base line at each extremity without reaching it, we can apply correction factors to eliminate this error. The corrections are small, and there is little use in making them if our original figures contain very much error. But where we have accurate continuous data with the characteristics just described, we may well apply Sheppard's corrections. Since our data on students' heights approximately meet these requirements, we may illustrate the application of the corrections.

The moments that we have computed, which have not been corrected by Sheppard's process, are called the *crude moments*, to distinguish them from the *adjusted moments* which we get by applying Sheppard's corrections.

The first and third moments need no correction. If we let μ_2 stand for the adjusted second moment and μ_4 for the adjusted fourth moment, we apply the corrections thus:

$$\mu_2 = v_2 - \frac{Ci^2}{12}$$

$$\mu_3 = v_3$$

$$\mu_4 = v_4 - \frac{v_2 Ci^2}{2} + \frac{7Ci^4}{240}$$

If the class interval is one, the application is obviously simplified. If we correct the moments by Sheppard's correction, we use the corrected moments rather than the crude moments in computing the values of σ, α_3, and α_4.

Applying Sheppard's corrections to our problem of the distribution of students' heights, we get the following adjusted moments:

$$\mu_2 = 43.353 - \frac{9}{12} = 43.353 - 0.75 = 42.603$$

$$\mu_3 = v_3 = -9.774$$

$$\mu_4 = 5508.567 - \frac{(43.353)(9)}{2} + \frac{(7)(81)}{240} = 5315.84$$

If, now, we use these corrected moments in computing σ, α_3, and α_4, we have

$$\sigma = \sqrt{\mu_2} = \sqrt{42.603} = 6.53$$
$$\alpha_3 = \frac{\mu_3}{\sigma^3} = -\frac{9.774}{278.4} = -0.035$$
$$\alpha_4 = \frac{\mu_4}{\sigma^4} = \frac{5315.84}{1818.2} = 2.93$$

If the adjusted and the crude results are compared, we have

	Crude	Adjusted
2d moment	43.353	42.603
4th moment	5508.567	5315.84
α_3	−0.034	−0.035
α_4	2.926	2.93
σ	6.58	6.53

It will thus be seen that the corrections bring but minor changes in the value of the α terms.

8.5. Moments of Probability Distributions.—We learned in Sec. 7.2 that the arithmetic mean and standard deviation of probability distributions could be computed quickly and easily by means of the formulas

$$\bar{X} = np$$
$$\sigma = \sqrt{npq}$$

Now that we have studied the higher moments, we can add two similar useful formulas[1] for the values of the alphas:

$$\alpha_3 = \frac{q - p}{\sqrt{npq}} = \frac{q - p}{\sigma}$$
$$\alpha_4 = \frac{1}{npq} - \frac{6}{n} + 3 = \frac{1}{\sigma^2} - \frac{6}{n} + 3$$

These formulas will hold for any point binomial distribution found by evaluating $(q + p)^n$. The student will note that, as n grows extremely large, the value of α_3 approaches zero and the value of α_4 approaches 3. But these are the values in a normal distribution. Hence we see that the point binomial distribution approaches the normal distribution as n gets extremely large.

[1] For proof of these formulas, as well as proofs of the formulas for the arithmetic mean and standard deviation, see John F. Kenney, "Mathematics of Statistics," Vol. II, pp. 11–15, D. Van Nostrand Company, Inc., New York, 1939.

It is also apparent that α_3 is zero whenever $q = p$, and therefore in such cases we get symmetrical distributions.

Perhaps it is not quite so evident that point binomial distributions are entirely fixed in terms of their arithmetic means and their standard deviations. Yet the student will note that the formulas for α_3 and α_4 can be stated in the alternative form:

$$\alpha_3 = \frac{2\sigma}{\bar{X}} - \frac{1}{\sigma}$$

$$\alpha_4 = \frac{1}{\sigma^2} - \frac{6(\bar{X} - \sigma^2)}{\bar{X}^2} + 3$$

Here it is evident that if we know the values of $\bar{X}$ and σ we can find the values of α_3 and α_4 immediately. Figure 7.1, page 165, shows the values obtained when we raise $q + p$ to the 14th power if both p and q equal $\frac{1}{2}$; that is, we have the values of the terms of $(\frac{1}{2} + \frac{1}{2})^{14}$. We now see from our formulas that for this distribution the values are

$$\bar{X} = 14(\tfrac{1}{2}) = 7$$

$$\sigma = \sqrt{14(\tfrac{1}{2})(\tfrac{1}{2})} = 1.87$$

$$\alpha_3 = \frac{0.5 - 0.5}{1.87} = 0$$

$$\alpha_4 = \frac{1}{3.5} - \frac{6}{14} + 3 = 2.86$$

We also note immediately from Fig. 8.1 that when p and q are not equal, the point binomial will be skewed. Suppose we test this for the case, say, where $p = 0.3$ and $q = 0.7$. As always, we have $p + q = 1$. If we raise this binomial to the eighth power, we get

$$\begin{aligned}(q + p)^8 &= (0.7 + 0.3)^8 \\ &= 0.7^8 + 8(0.7^7)(0.3) + 28(0.7^6)(0.3^2) + 56(0.7^5)(0.3^3) \\ &\quad + 70(0.7^4)(0.3^4) + 56(0.7^3)(0.3^5) + 28(0.7^2)(0.3^6) \\ &\quad + 8(0.7)(0.3^7) + 0.3^8\end{aligned}$$

If, now, we evaluate each of these terms, we find the following:

$$\begin{aligned}(0.7 + 0.3)^8 &= 0.5764801 + 0.19765032 + 0.29647548 \\ &\quad + 0.25412184 + 0.13613670 + 0.04667544 + 0.01000188 \\ &\quad + 0.00122472 + 0.00006561\end{aligned}$$

If we plot these terms, as in the lower left-hand section of Fig. 8.1, we obtain an asymmetrical or skewed distribution, as contrasted with the symmetrical distribution of Fig. 7.1 or the symmetrical distribution in the center of Fig. 8.1. In Fig. 8.1 we see the point binomials obtained by raising $(q + p)$ to the eighth power,

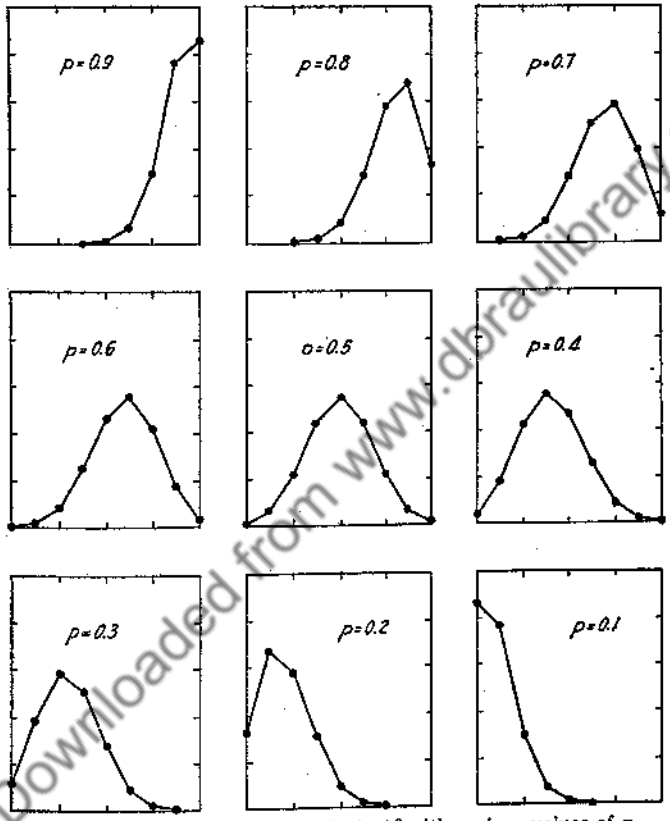


FIG. 8.1.—Point binomials for $(q + p)^8$ with various values of p.

with varying values of p and q. When p and q are both equal to $\frac{1}{2}$, we get the symmetrical point binomial in the center of the figure, and the more p (or q) differs from $\frac{1}{2}$, the more skewness becomes evident.

In the particular case which we have just worked out, where $p = 0.3$ and $q = 0.7$, we may try applying our simple formulas

for probability distributions. We find that

$$\bar{X} = np = 8(0.3) = 2.4$$
$$\sigma = \sqrt{npq} = \sqrt{(8)(0.3)(0.7)} = \sqrt{1.68} = 1.30$$
$$\alpha_3 = \frac{(q - p)}{\sigma} = \frac{(0.7 - 0.3)}{1.30} = 0.308$$
$$\alpha_4 = \frac{1}{npq} - \frac{6}{n} + 3 = \frac{1}{1.68} - \frac{6}{8} + 3 = 2.845$$

It is the fact that p and q are unequal which has been responsible for the skewness in eight of the nine sections of Fig. 8.1. Let us note in the above formulas the effect of changing p or q when we hold n constant. Let us start with $p = q = \frac{1}{2}$, and then increase the size of p slowly, reducing the size of q always so that $p + q = 1$. We note that $\bar{X}$ will increase. But if we increase p and decrease q, keeping their sum equal to 1, their product pq will diminish, as the student will discover immediately by experiment. Therefore σ will diminish as p moves away from 0.5 in either direction. Since α_3 is based on $q - p$, it will be negative when q is smaller than p and positive when q is larger than p; and the farther p or q is from 0.5 the greater will be the difference between them; so the greater will be the absolute size of α_3. Finally, as p or q gets farther from 0.5, the value of pq in the denominator of the last formula will diminish, thus increasing the value of the fraction and of α_4. To summarize, we note that $\bar{X}$ grows larger whenever p increases (if we hold n constant). We note also that, if n does not change, the sizes of the other three values depend on the amount of difference between the values of p and q. The greater the difference between p and q, the smaller the value of σ the larger the value of α_3 and the larger the value of α_4.

It is now time for us to learn more definitely how to interpret these results, as we do in the next two sections.

8.6. Measures of Skewness.—Thus far we have confined our description of frequency curves in the main to measures of central tendency (averages) and measures of dispersion. These two types of measures tell us a good deal about the character of the distribution. For example, when we have discovered that the average height of a group of students is 175.3 cm. and the standard deviation of heights in the group is 6.6 cm., we know (if the

distribution is roughly normal) that about two-thirds of the students have heights between 168.7 and 181.9 cm. We know also that almost never should we find a student shorter than 155.5 cm. or taller than 195.1 cm. (the mean plus and minus 3σ).

We have discovered, however, that the normal curve is symmetrical. We could be somewhat more confident in our interpretation of the mean and the measures of dispersion if we knew that our distribution was symmetrical. We have seen that the distribution of incomes in the United States is not symmetrical (page 112), and with any distribution we may well wish to test the symmetry. When a distribution is asymmetrical we usually call it a *skewed* frequency distribution, and the measures of asymmetry are usually called *measures of skewness.*

Many measures of skewness have been proposed, and none has been uniformly adopted. For this reason, when one gives a measure of skewness it is necessary to indicate the method by which it was computed. The commoner methods are given here.

If our distribution is mound-shaped (that is, if it has small frequencies at the extremities and larger frequencies toward the center) and symmetrical, the mean, the median, and the mode will coincide. If the curve is skewed, these measures will not coincide. Thus it is possible to acquire some idea of the absolute amount of the skewness by noting the amount of the divergence between any two of these measures of central tendency. If we wish to measure *relative skewness,* we shall have to compare the displacement of the averages with some standard measure of dispersion. Usually we should use the standard deviation for the latter measure. Karl Pearson has suggested the following as a measure of relative skewness:

$$\text{Sk.} = \frac{(\bar{X} - \text{Mo.})}{\sigma}$$

We have discovered the following values for these constants in the case of the heights of Harvard students:

$$\bar{X} = 175.3 \text{ (page 87)}$$
$$\text{Mo.} = 175.2 \text{ (page 99)}$$
$$\sigma = 6.6 \text{ (page 143)}$$

Substituting these in the equation for skewness, we have

$$\text{Sk.} = \frac{(175.3 - 175.2)}{6.6}$$

$$= \frac{0.1}{6.6} = 0.015$$

It is evident from the formula that the skewness may be either positive or negative. It will be positive when the mean exceeds the mode and negative when the mean is smaller than the mode. Such cases are illustrated in Fig. 8.2.[1] The upper part of the figure shows a distribution in which the mean is pulled toward the right by the few extremely high cases: this is positive skewness. In the lower part of the figure the mean is pulled toward the left by the few very small cases: this is negative skewness.

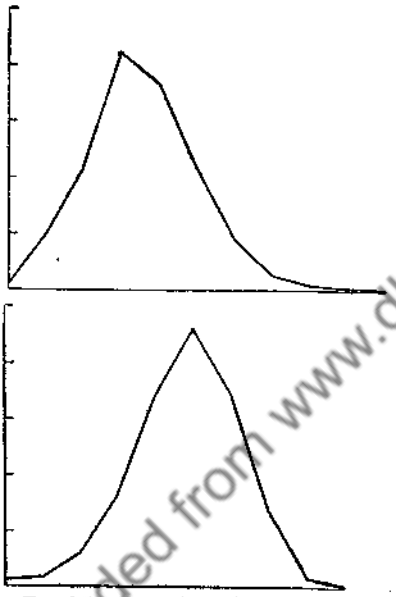


FIG. 8.2.—Two skewed curves. The upper curve exhibits positive skewness, and the lower curve negative skewness.

We have discovered earlier (page 97) that the mode is difficult to find, and that different methods of locating it give different results. For that reason Pearson's formula, given above, is not entirely satisfactory. We have seen also that, when the asymmetry is not great, the averages have the following relationship:[2]

$$\text{Mo.} = 3\ \text{Med.} - 2\bar{X}$$

[1] The upper part of Fig. 8.2 shows the distribution of hourly earnings of 2960 employees of filling stations in the United States in 1931. Data from *United States Bureau of Labor Statistics Bulletin* 578, p. 9.

The lower part of Fig. 8.2 shows the distribution of the annual egg productions of 3131 white Leghorn hens. Data from *Storrs Agricultural Experiment Station Bulletin* 147, p. 246.

[2] See p. 98.

If this value of Mo. is substituted in our equation for skewness, it becomes

$$\text{Sk.} = \frac{(\bar{X} - 3\ \text{Med.} + 2\bar{X})}{\sigma}$$
$$= 3\,\frac{(\bar{X} - \text{Med.})}{\sigma}$$

If we substitute the values of the Harvard student problem in this equation, remembering that the median was found to be 175.3 (page 94), we have

$$\text{Sk.} = \frac{3(175.3 - 175.3)}{6.6} = 0$$

If we take the figures for the mean and the median as originally computed before rounding off (see pages 87 and 94), we have

$$\text{Sk.} = \frac{3(175.335 - 175.28)}{6.6} = 0.025$$

Again we find that the skewness is very small in this distribution. And again the skewness, what there is of it, is positive.

In a symmetrical distribution the quartiles would, of course, be equidistant from the median; that is, Med. $- Q_1 = Q_3 -$ Med., if the distribution is symmetrical. If the distribution is not symmetrical, the quartiles will not be equidistant (unless the entire asymmetry is located in the extreme quarters of the data, or unless there is some very peculiar arrangement of the data within the central quartiles).

These facts have led Bowley to suggest the following as a measure of skewness:

$$\text{Sk.} = \frac{(Q_3 - \text{Med.}) - (\text{Med.} - Q_1)}{Q_3 - Q_1}$$
$$= \frac{Q_3 + Q_1 - 2\ \text{Med.}}{Q_3 - Q_1}$$

If we apply this measure of skewness to our data, we find

$$Q_3 = 179.84 \text{ (page 129)}$$
$$Q_1 = 170.95 \text{ (page 103)}$$
$$\text{Med.} = 175.28 \text{ (page 94)}$$
$$\text{Sk.} = \frac{(179.84 + 170.95 - 350.56)}{(179.84 - 170.95)}$$
$$= \frac{0.23}{8.89} = 0.026$$

Better yet as an estimate of the mode, but requiring still more computation, is the following:[1]

$$\frac{\bar{X} - \text{Mo.}}{\sigma} = \frac{\sqrt{\beta_1}\,(\beta_1 + 3)}{2(5\beta_1 - 6\beta_1 - 9)}$$

where $\beta_1 = \alpha_3^2$ and $\beta_2 = \alpha_4$.

This, again, can be used as a measure of skewness, being positive if the mean exceeds, and negative if the mean falls short of, the mode. Computing the value of this measure of skewness from our illustrative data, we have

$$\begin{aligned}
\bar{X} &= 175.335 \text{ (page 87)} \\
\sigma &= 6.582 \text{ (page 143)} \\
\beta_1 &= 0.00117 \text{ (page 193)} \\
\beta_2 &= 2.931 \text{ (page 197)} \\
\text{Sk.} &= \frac{\sqrt{0.001117}\,(2.935 + 3)}{2(14.675 - 0.00702 - 9)} \\
&= \frac{0.203}{11.34} = 0.0179
\end{aligned}$$

This shows an extremely slight positive skewness. If we use this value for computing the mode, we have

$$\begin{aligned}
\frac{175.335 - \text{Mo.}}{6.582} &= 0.0179 \\
\text{Mo.} &= 175.217
\end{aligned}$$

This value for the mode is almost identical with the value discovered by the other method based on moments. It is probably the best estimate we can make of the modal height of the 1000 Harvard students.

The student will now understand why he was told on page 201 that it is always necessary to accompany any measure of skewness with a statement of the method of computation. We have computed the skewness of the student heights by several methods, and have found varying answers. Let us collect the answers for purposes of comparison:

$$\begin{aligned}
\text{Sk.} &= +0.015 \text{ (page 202)} \\
\text{Sk.} &= +0.025 \text{ (page 203)} \\
\text{Sk.} &= +0.026 \text{ (page 203)}
\end{aligned}$$

[1] MILLS, "Statistical Method," p. 546, Henry Holt and Company, Inc., 1924.

Sk. = −0.035 (page 205)
Sk. = +0.0179 (page 206)

It will be noted that, although there is some variation among these measures, as is to be expected since they have been computed by decidedly different methods, nevertheless the five results are nearly identical in size. The fact that the measures differ in sign, one being negative and the other four positive, is unimportant, since they are all approximately equal to zero. The extreme difference, between the third and the fourth measures, amounts to but 0.061; that is, the differences are confined to the second decimal place.

8.7. Measures of Kurtosis.—We have studied measures of central tendency, measures of dispersion, and, just now, measures of skewness. There remains but one more common type of measure of the characteristics of a frequency distribution. These measures are called *measures of kurtosis*. They show the extent to which the distribution is more peaked or more flat-topped than the normal curve. If the items are more closely bunched around the mode than normal, making the curve unusually peaked, we say that the curve is *leptokurtic*. If, on the other hand, the curve is more flat-topped than normal, we say that it is *platykurtic*. The normal curve itself is *mesokurtic*. The condition of peakedness or of flat-toppedness itself is known as *kurtosis* or *excess*.

The principal measure of kurtosis is the value which we have already computed and called α_4. It is also sometimes symbolized by β_2, the two being identical, and, as we have seen, being defined thus:

$$\alpha_4 = \beta_2 = \frac{v_4}{\sigma_4}$$

In the normal curve, α_4 and β_2 equal 3. When they are greater than 3, the curve is more peaked than the normal curve, and is said to be *leptokurtic*. When they are less than 3, the curve has a flatter top than the normal curve, and is said to be *platykurtic*. The normal curve, and other curves with α_4 and β_2 equal to 3, are said to be *mesokurtic*. In the distribution of student heights, α_4 is 2.93. The curve is slightly flatter than the normal curve—slightly platykurtic.

A very peaked curve has kurtosis greater than 3. As we flatten the curve the value of α_4 decreases, and when it has reached 3 the curve is mesokurtic. If we flatten it still more the

curve becomes platykurtic. Ultimately, of course, the curve will flatten out entirely into a straight line, with the various frequency classes containing equal numbers of cases. This is what we have called a *rectangular distribution* (see Sec. 3.14, page 41). The value of α_4 for a rectangular distribution depends on the number of frequency classes, being 1.8 when the number of classes is infinite, and approaching 1.8 rather rapidly for finite numbers of classes. Table 8.3 shows the values of α_4 for rectangular distributions with small numbers of classes.

TABLE 8.3.—VALUES OF α_4 IN RECTANGULAR DISTRIBUTIONS WITH VARIOUS NUMBERS OF CLASSES

Number of Classes	Value of α_4	Number of Classes	Value of α_4
1	0.0000	12	1.7832
2	1.0000	13	1.7857
3	1.5000	14	1.7877
4	1.6400	15	1.7893
5	1.7000	16	1.79059
6	1.7314	17	1.79167
7	1.7500	18	1.79257
8	1.7619	19	1.79333
9	1.7700	20	1.793985
10	1.7758	25	1.796153
11	1.7800	30	1.797330

If we continue to push down the middle of the distribution still further, the value of α_4 will fall below that given in Table 8.3, and the distribution will become U-shaped. We can, then, judge something of the shape of the frequency curve by the value of α_4. If α_4 is greater than 3, the curve is more peaked than the normal curve. If the value of α_4 lies between 3 and the value given in Table 8.3, the curve is flatter than the normal curve, but still mound-shaped. If α_4 has the value given in Table 8.3, the distribution is rectangular. If α_4 has a value lower than that given in Table 8.3, the distribution is U-shaped. If the student does not have a copy of the table handy for reference, he can remember that in most actual frequency tables the critical value of α_4 which separates mound-shaped from U-shaped curves is between 1.75 and 1.8.

8.8. Interpretation of Frequency Statistics.—In general we can describe any frequency distribution quite satisfactorily in terms of five statistical measures:

1. The number of cases.
2. An average, or measure of central tendency.
3. A measure of dispersion.
4. A measure of skewness.
5. A measure of kurtosis.

By far the commonest among the last four of these measures are the arithmetic mean, the standard deviation, α_3, and α_4. Any given distribution has some definite five values of these measures, yet in some other distribution all five measures may be different. Such a number, used to describe a frequency distribution, being a constant for any particular distribution but a variable as we shift from distribution to distribution we call a *statistic* of the distribution. We can say, then, that a frequency distribution can be described with reasonable accuracy in terms of five statistics.

The student who remembers our use of the normal curve to describe distributions (see Secs. 7.8 and 7.9 and Fig. 7.7) may feel that the first three of those statistics are enough—that we

TABLE 8.4.—FOUR FREQUENCY DISTRIBUTIONS ILLUSTRATING THE USE OF COMMON FREQUENCY STATISTICS

Class Limits	f_1	f_2	f_3	f_4
20– 29	1			
30– 39	4	2		
40– 49	6	5	12	
50– 59	8	10	12	34
60– 69	10	16	12	12
70– 79	16	17	12	6
80– 89	18	18	12	4
90– 99	16	12	12	6
100–109	10	10	12	12
110–119	8	7	12	34
120–129	6	5	12	
130–139	4	3		
140–149	1	1		
150–159	..	1		
160–169	..	1		

can get a complete and satisfactory description of a distribution in terms of its frequency, arithmetic mean, and standard deviation. Let us look, therefore, at Table 8.4. In this table the first column represents class limits, and each of the following four columns represents a set of frequencies. We have really combined in Table 8.4 four frequency tables for easy comparison. To distinguish them, we have labeled the frequencies of the first distribution f_1, those of the second distribution f_2, those of the third f_3, and those of the fourth f_4. The table tells us, for example, that there are 16 cases in the 70–79 class in the first distribution, while there are 17 cases in the same class in the second distribution, 12 in the third, and 6 in the fourth.

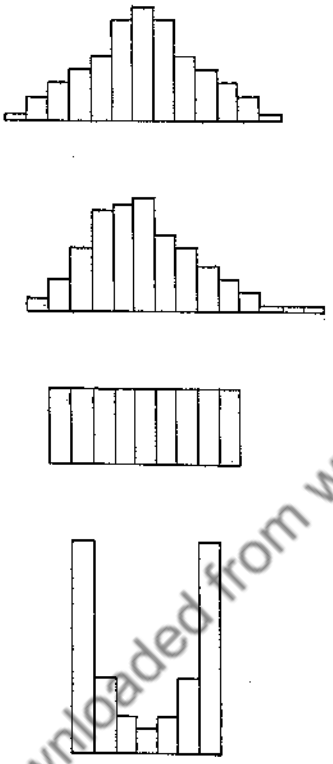


FIG. 8.3.—These four distributions all have exactly the same number of cases, the same arithmetic mean, and the same standard deviation.

If the student will take the trouble to compute the arithmetic means and standard deviations of these distributions, he will find that they are exactly the same. In each case the arithmetic mean is 85 and the standard deviation is 25.8. Moreover, in each distribution the total number of cases is 108. On the basis of these three statistics alone, we should be tempted to say that the four distributions were identical. Yet these distributions are shown graphically in Fig. 8.3. The student will see immediately that there is little similarity between the four cases in spite of the fact that they have exactly the same numbers of cases, exactly the same arithmetic mean and exactly the same standard deviation.

If, now, we compute the values of α_3 and α_4 we uncover the differences at once. The entire five frequency statistics for the four cases are

	Case 1	Case 2	Case 3	Case 4
N	108	108	108	108
$\bar{X}$	85	85	85	85
σ	25.8	25.8	25.8	25.8
α_3	0	+0.57	0	0
α_4	2.565	3.188	1.770	1.23

The values of α_3 and α_4 immediately serve to distinguish the distributions. We note that all distributions save number 2 are symmetrical, while that one has moderate positive skewness. Case 2 is more peaked than the normal curve (leptokurtic); case 1 is mound-shaped but flatter than the normal curve; case 3 (since there are 9 classes in that distribution) is exactly rectangular, as we see when we compare its α_4 value with that given in Table 8.3; and case 4 is U-shaped (since it was computed from data distributed in seven classes, and any value less than 1.75 signifies that such a distribution is U-shaped). A comparison of the values of the statistics of these four distributions with their histograms in Fig. 8.3 will help the student to understand the interpretation. In case 2 the skewness is so marked that the arithmetic mean is thrown about 0.28 standard deviations to the right of the mode. The student will find it informative to compute the frequency statistics from the data in Table 8.4 as a check.

In summarizing the description of any frequency distribution, then, we give five statistics. If we wish to summarize our description of the heights of Harvard students, we bring together the various statistics which we have computed as follows:

$$
\begin{aligned}
N &= 1000 \\
\bar{X} &= 175.335 \quad \text{(page 87)} \\
\sigma &= 6.582 \quad \text{(page 143)} \\
\alpha_3 &= -0.035 \quad \text{(page 197)} \\
\alpha_4 &= 2.93 \quad \text{(page 197)}
\end{aligned}
$$

The distribution is perhaps *very* slightly skewed in a negative direction, although it is practically symmetrical. It also is slightly flatter than the normal curve. The symmetrical distribution of Fig. 7.1 has the values $\alpha_3 = 0$ and $\alpha_4 = 2.86$. It is symmetrical and slightly platykurtic. The distribution in Fig.

8.1 has the values $\alpha_3 = 0.357$ and $\alpha_4 = 2.794$. It has positive skewness, with the arithmetic mean larger than the mode, and lying about 0.178 standard deviations to the right of the mode. It is also slightly flat-topped, but definitely mound-shaped rather than rectangular or U-shaped.

8.9. The Pearsonian System of Frequency Curves.—When α_3 differs very markedly from 0, or when α_4 differs very markedly from 3, we know that our curve is not normal. Then we have to turn to some other kind of frequency curve to describe our data. Many such non-normal curves have been described, but among the most useful are the families of curves described by Karl Pearson, the eminent English biometrician. Pearson's system includes frequency curves which are mound-shaped but asymmetrical, those which are J-shaped, those which are U-shaped, etc. A very large proportion of all actual frequency distributions can be described tolerably well by one or another of the 12 different classes of curves which Pearson describes. When we are fitting the normal curve to a distribution, we need know only the number of cases, the arithmetic mean, and the standard deviation (see Secs. 7.8 and 7.9). For most of Pearson's curves it is necessary, in addition, to know the values of α_3 and α_4, or, to use Pearson's symbolism, to know the values of β_1 and β_2 where

$$\beta_1 = \alpha_3^2$$
$$\beta_2 = \alpha_4$$

In deciding what type of curve to fit, it is also necessary to know the value of κ_2 which is defined as follows:

$$\kappa_2 = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)(2\beta_2 - 3\beta_1 - 6)}$$

Although there are a dozen classes of curves in the Pearson system, his Type III curve (of which the normal curve is a special case) is by far the most important. Fortunately it is also the easiest to fit. We shall confine our discussion of the Pearson curves to the Type III curve, and the student who wishes to investigate the other types is referred to the books suggested at the end of this chapter (see Sec. 8.13).

8.10. Fitting Pearson's Type III Curve.—Pearson's Type III curve may be fitted to any distribution in which

$$\frac{2\alpha_4 - 3\alpha_3^2 - 6}{\alpha_4 + 3} = 0$$

or, if we use Pearson's symbols,

$$\frac{2\beta_2 - 3\beta_1 - 6}{\beta_2 + 3} = 0$$

In the normal curve this expression will be equal to zero, and, in addition, α_3 will be equal to zero. The Type III curve covers all cases where the former expression is equal to zero, whether or not α_3 equals zero; hence we see that the Type III curve will cover asymmetrical as well as symmetrical curves.

The fitting of Type III curves is exactly parallel to the fitting of the normal curve which was explained and illustrated in Secs. 7.8 and 7.9. We can fit Type III curves either by areas or by ordinates, and we make use of tables of ordinates and areas of skewed curves which are somewhat similar to the tables of areas and ordinates of the normal curve in Appendixes I and II, except that we now have to have separate entries for each different degree of skewness. Detailed tables showing the areas and ordinates have been published by Salvosa,[1] and condensed extracts from these tables appear in the laboratory manual which accompanies this text.[2]

We illustrate the fitting of the skewed curve in Table 8.5. The data in the first column are class marks, while those in the second column are frequencies. If we go through the processes explained earlier in this chapter, we find the following values for this table:

$$N = 675$$
$$\bar{X} = 124.6$$
$$\sigma = 15.49$$
$$\alpha_3 = 0.408$$
$$\alpha_4 = 3.258$$

When we test the distribution to see whether or not a Type III curve should be fitted, we discover

[1] LUIS R. SALVOSA, Tables of Pearson's Type III Function, *Annals of Mathematical Statistics*, Vol. 1, May, 1930, pp. 191*ff*.

[2] ALBERT E. WAUGH, "Laboratory Manual and Problems for Elements of Statistical Method," Tables A3 and A4, McGraw-Hill Book Company, Inc., New York, 1938.

$$\frac{2\alpha_4 - 3\alpha_3^2 - 6}{\alpha_4 + 3} = \frac{2(3.258) - 3(0.408^2) - 6}{3.258 + 3} = 0.0027$$

This expression gives a value of zero in Type III curves, and here the value is so nearly equal to zero that it seems in advance fairly safe to try that type of curve. In fact, practice will show that this curve can well be fitted even when there is a considerable departure from zero.

Column 3 of Table 8.5 shows the distances of the class marks from the arithmetic mean, found by subtracting the value of the arithmetic mean (124.6) from each class mark. Column 4 is found by dividing each entry in column 3 by the standard deviation (15.49) to convert the distances into standard units. Column 5 is found from the printed tables, similar to those in Appendix II. The entries here are from Salvosa's detailed tables of ordinates of skewed curves, although approximately the same values could be found by interpolation in the tables of the author's manual. The student is warned that these entries in column 5 cannot be computed from or found in any material in this volume, but the process of finding them is parallel to the process of finding the figures in column 5 of Table 7.5, page 181, except that other tables are used. In the last column, we have

TABLE 8.5.—FITTING A SKEWED TYPE III CURVE BY THE METHOD OF ORDINATES

Class Mark (X)	Number of Cases (f)	x	x/σ	Tabular Value (y)	Computed Frequency (f')
184.5	2	59.9	3.87	0.0015	0.65
174.5	2	49.9	3.22	0.0062	2.7
164.5	6	39.9	2.58	0.0216	9.4
154.5	28	29.9	1.93	0.0647	28.2
144.5	76	19.9	1.28	0.1569	68.4
134.5	126	9.9	0.64	0.2918	127.2
124.5	169	− 0.1	−0.006	0.3980	173.5
114.5	159	−10.1	−0.65	0.3626	158.1
104.5	82	−20.1	−1.30	0.1923	83.8
94.5	24	−30.1	−1.94	0.0494	21.5
84.5	1	−40.1	−2.59	0.0041	1.8

the computed frequencies. These are found, as in Sec. 7.8, by multiplying the tabular values of column 5 by a constant

equal to $N(Ci)/\sigma$, which, in this problem, is 675(10)/15.49, or 436. In other words, each item in the last column is found by multiplying the corresponding item in column 5 by 436, and in any other problem the last column would always be found by multiplying the tabular values by a constant equal to $N(Ci)/\sigma$.

These computed frequencies in the last column are the frequencies for the corresponding Type III curve; that is, they are the frequencies which there would be in each class in a Type III curve when N was 675, $\bar{X}$ was 124.6, σ was 15.49, and α_3 was +0.408. We note, for example, that while there really were 169 cases in the class which centered at 124.5, we should have expected to find 173.5 cases in this class in a Type III curve. A comparison of the actual frequencies in column 2 with the computed frequencies in the last column will show that the actual distribution was very much like a Type III distribution; although the student is warned not to rely on such casual inspection to determine whether or not the correspondence between actual and computed frequencies is close. Even an experienced statistician would be unable to tell from superficial inspection how well the two sets of data corresponded, and we shall see at a later point in this chapter (Sec. 8.12) how one may test the "goodness of fit" quantitatively. We have now gone far enough, however, so that the student should see that it is a relatively easy matter to fit a Type III curve from the prepared tables, following the same general system which we have already used for the normal curve.[1]

8.11. The Poisson Series.—In addition to the normal curve and the Type III curves, there is another whole set of frequency curves often used in advanced statistics called the Gram-Charlier series. This series we cannot cover here, but we shall take time

[1] There are at least two respects in which the method of fitting Type III curves differs from the methods studied earlier for normal curves. In the first place, since the Type III curves are not symmetrical, we must keep track of the sign of our deviations from the arithmetic mean, since the height of the curve at a point 1.4 standard deviations above the mean will not equal the height at a point 1.4 standard deviations below the mean. And in the second place, the tables are made up for positive values of α_3. If we fit a Type III distribution to a set of data in which α_3 is negative, we must reverse the table by selecting points in the table which deviate on the opposite side of the mean from those in our problem, taking, for example, -1.7σ in the table when our class mark is $+1.7\sigma$.

to say a few words about another very useful type of frequency curve called the Poisson curve.[1] We have seen that the normal curve occurs in distributions that are subject to chance in which the result depends on a large number of causes, each of which is a 50–50 chance. If the causes are not 50–50 (that is, if $p \neq q \neq 1/2$), our distribution is asymmetrical or skewed, as we saw in Sec. 8.5. Sometimes this skewness is moderate, when p and q are almost the same size, but if either p or q becomes very small the distribution takes on a marked skewness and the normal curve cannot be used to describe it with even approximate accuracy. Under such circumstances the Poisson distribution is sometimes useful.

Let us start by looking back at the moments of probability distributions given in Sec. 8.5. Here we discover that, in such a distribution, we have the following values of the commoner statistics (see page 197):

$$\bar{X} = np$$
$$\sigma = \sqrt{npq}$$
$$\alpha_3 = \frac{q - p}{\sqrt{npq}}$$
$$\alpha_4 = \frac{1}{npq} - \frac{6}{n} + 3$$

Let us now assume that p has become very, very small, so that q is almost equal to 1. Yet let us suppose that we are dealing with a large enough number of cases so that np is a sensible quantity even though p is very small. Then we see right away that $\sqrt{npq}$ will be substantially the same as $\sqrt{np}$ (since q will be approximately equal to 1). Therefore under such circumstances the value of the standard deviation will be approximately $\sqrt{np}$ or $\sqrt{\bar{X}}$. Similarly the numerator of the value of α_3 will become approximately 1 (since q will be approximately 1, and p will be too small to have much influence); hence the value of α_3 will be approximately $1/\sqrt{\bar{X}}$, or $1/\sigma$. Similarly, if n is very large the value of $6/n$ will be negligible, and our formula for α_4

[1] Technical discussion of the basic assumptions of the Poisson series can be found in Lucy Whitaker, On the Poisson Law of Small Numbers, *Biometrika*, Vol. 10, 1914–1915, pp. 36*ff.*; and R. A. Fisher, "Statistical Methods for Research Workers," Oliver & Boyd, London, 1932.

will become approximately $3 + 1/\bar{X}$. In such a distribution, then, we can state our statistics in terms of the following adjusted formulas:

$$\bar{X} = np$$
$$\sigma = \sqrt{\bar{X}}$$
$$\alpha_3 = \frac{1}{\sqrt{\bar{X}}} = \frac{1}{\sigma}$$
$$\alpha_4 = 3 + \frac{1}{\bar{X}}$$

It is at once evident that if the value of the arithmetic mean is known we can compute the values of all the other frequency statistics. This is one of the peculiarities of the Poisson distribution—we need know only the arithmetic mean to fit the entire curve. It is a curve that is useful in describing data subject to chance, in which the chance of occurrence is very, very small, but in which there are enough total observations so that the phenomenon does actually occur sometimes. Suppose, for example, that we were considering the number of ministers murdered in Chicago each year, tabulating the number of years in which no minister was murdered, the number in which one was murdered, the number in which two were murdered, etc. It is probable that we should get a J-shaped distribution, with most of our years in the "no-murder" class, fewer in the "one-murder" class, and fewer and fewer years as the number of murders increased. Other hypothetical examples of this sort of distribution are given in Sec. 3.13, where J-shaped distributions are discussed. Poisson series are not always J-shaped, but they are usually either J-shaped or mound-shaped and badly skewed in a positive direction. In such distributions it will often pay to try the Poisson curve, especially since it is unusually easy to fit.

We shall illustrate the fitting of this distribution with figures showing the rate at which vacancies have occurred in the U. S. Supreme Court from 1837 to 1932.[1] During this period the numbers of years in which various numbers of vacancies occurred were as follows:

[1] The data are taken from W. Allen Wallis, The Poisson Distribution and the Supreme Court, *Journal of The American Statistical Association*, Vol. 31, June, 1936, pp. 376*ff*.

Number of Vacancies	Number of Years
0	59
1	27
2	9
3	1
Total	96

From this table we compute first the average number of vacancies, which is 0.500 per year. To fit the Poisson series we need no data other than the number of cases and the average.[1] The formula used is

$$y = N\frac{A^x e^{-A}}{x!}$$

where y is the estimated frequency, N the number of cases, A the average occurrence computed as directed in the footnote below, $e = 2.71828+$, and $x!$ is factorial x or the product of all positive integers from 1 to x. We substitute the values of N (96) and A (0.5) together with any value of x (say 1) and solve the equation to find the corresponding value of y.

In our problem we shall substitute for x the various numbers from 0 to 3, finding in each case the expected value of y. Our equation in this case becomes

$$y = 96\frac{0.5^x 2.71828^{-.5}}{x!}$$

It will be easier to solve this equation if we put it in logarithmic form. This gives us

$$\log y = \log 96 + x(\log 0.5) - 0.5(\log 2.71828) - \log x!$$

Substituting the logarithms in the places indicated, we have

$$\log y = 1.98227 - .30103x - .21715 - \log x!$$

[1] The average used in fitting this series must always be computed by numbering the classes from 0 up. In this case they were naturally so numbered, but if other class limits appeared the numbers 0, 1, 2, 3, . . . would be substituted for them and the average would be computed from these substituted values.

Let us first let $x = 0$ and substitute in the equation, solving to find the value of y.[1] This gives us

$$\log y = 1.98227 - .21715 = 1.76512$$
$$y = 58.23$$

Thus we have estimated that no vacancies would be expected to occur in 58.23 of the 96 years. If, now, we let $x = 1$ and make the necessary substitutions, we find

$$\log y = 1.98227 - .30103 - .21715 = 1.46409$$
$$y = 29.11$$

Now let $x = 2$:

$$\log y = 1.98227 - .60206 - .21715 - .30103 = .86203$$
$$y = 7.28$$

Letting $x = 3$, we find

$$\log y = 1.98227 - .90309 - .21715 - .77815 = .08388$$
$$y = 1.21$$

If we collect these computations and compare them with the actual occurrences, we have

Number of Vacancies	Number of Occurrences	
	Estimated	Actual
0	58.23	59
1	29.11	27
2	7.28	9
3	1.21	1
Totals.....	95.83	96

It will be seen that this method makes it possible to fit a mathematical curve which follows very closely the original data. We discover that two vacancies in a single year have occurred slightly more often than would have been expected, and that a single vacancy has occurred slightly less often; but by and large the

[1] It is necessary to remember here that the value of factorial zero is taken as 1. That is, when $x = 0$, $x! = 1$.

actual data follow very closely those which would be anticipated by chance.

If one is interested in the relative likelihood that various numbers of vacancies will occur rather than in the actual numbers of years out of 96 in which each will occur (that is, if one wants relative probabilities rather than actual frequencies), one will omit the N in the formula. The remainder of the formula solved just as we have solved it will yield these probabilities. In this case a trial will disclose that the probabilities are approximately the following:

Number of Vacancies	Probability of Occurrence
0	0.606
1	0.303
2	0.076
3	0.013

The probability that more than three vacancies will occur in a year is found by subtracting the total of the above probabilities from 1.00. In this case we find that this probability is 0.002. In other words, in two years out of 1000 we might expect to find more than three vacancies occurring in the Supreme Court if the general underlying causal factors continue to operate as they have operated in the period studied.

Although the work of fitting a Poisson distribution is very simple by means of logarithms, it can be speeded up somewhat if we use prepared tables. These tables sometimes show the proportion of the cases in each class for any value of the arithmetic mean, so that after we have found the arithmetic mean all we have to do is look in the table to find what percentage of the cases will be in class 0, what percentage in class 1, etc.[1] Other tables merely show the proportion of cases in class 0 of a Poisson distribution with any given arithmetic mean, leaving the worker to compute the proportion in other classes from the following simple relationship.[2]

[1] For example, Karl Pearson, "Tables for Statisticians and Biometricians," pp. 113–124, Cambridge University Press, London, 1914, gives the proportion of cases in each class of a Poisson series for various values of the arithmetic mean, by steps of one-tenth of a unit in the arithmetic mean. For example, it would give data for cases when the mean was 0.6 or 0.7, but not for 0.63.

[2] For example, in the Laboratory Manual, which is designed for use with

To find the proportion of cases in class 1 of a Poisson distribution, multiply the proportion in class 0 by the arithmetic mean. To find the proportion in class 2, multiply the proportion in class 1 by one-half the arithmetic mean. To find the proportion in class 3, multiply the proportion in class 2 by one-third the arithmetic mean. In general, to find the proportion of the cases in class n, multiply the proportion in class $n - 1$ by $1/n$th the arithmetic mean.[1] We see at once from these directions that if the arithmetic mean is smaller than 1, the number of cases in each class will be smaller than that in the preceding class. If the value of the arithmetic mean exceeds 1, class 1 will be larger than class 0; if the arithmetic mean exceeds 2, the second class will be larger than the first, etc. Thus the Poisson distribution may be mound-shaped, but will be J-shaped when A is less than 1. Also we see that the Poisson distribution approaches the normal distribution as a limit when the value of A gets very large. We have seen at the beginning of this section that in a Poisson distribution $\alpha_2 = 1/\sqrt{\bar{X}}$, and as the arithmetic mean gets very large this will make α_3 approach zero. This indicates that the curve is becoming more and more symmetrical. Likewise, we saw that in such a distribution $\alpha_4 = 3 + 1/\bar{X}$. As the arith-

this text, we find immediately that when A is ½ the proportion of cases in class 0 is 60.65 per cent. This table gives values of the arithmetic mean to the nearest hundredth rather than merely to the nearest tenth. See Waugh, *op. cit.*, pp. 64–65.

[1] The student who is interested in mathematics will see the reason for this rule at once. The formula for the proportion of cases in any class of a Poisson distribution is

$$\frac{A^x e^{-A}}{x!}$$

For class 0 this would give

$$\frac{A^0 e^{-A}}{0!} = \frac{e^{-A}}{1}$$

Similarly, for class 1 we get

$$\frac{Ae^{-A}}{1}$$

For class 2 we get

$$\frac{A^2 e^{-A}}{(1)(2)}$$

The proportion in each class is found by multiplying the proportion in the preceding class by A/n.

metic mean increases, this value will approach closer and closer to 3, so that when the arithmetic mean of a Poisson distribution is large, α_3 approaches 0 and α_4 approaches 3, which means that the distribution approaches the normal curve. The student must remember in fitting these distributions to compute the arithmetic mean in the special way indicated, with the first class given a value of zero, the next a value of 1, etc.

8.12. Goodness of Fit and the Chi-square Test.—During this and the preceding chapter, we have been computing various sorts of curves to represent frequency distributions—normal curves, Type III curves, Poisson curves, etc. We may well ask with regard to any one of these curves just how well or how poorly it describes the distribution to which we fit it. When we wish to compare actual data with hypothetical data, to see whether or not the hypothesis is reasonable in the light of the actual data, one of the most useful methods involves the application of what is called the chi-square test, or the χ^2 test. When we apply this test, we are always making a comparison between an actual occurrence and a hypothetical occurrence. For example, we might say that on the basis of theory we should expect a penny to come up "heads" 28 times in 56 throws. If it did actually come up "heads" 35 times, we should compare the actual occurrence (35 heads) with the expected or theoretical occurrence (28 heads), and we should try to decide whether the actual differed from the expected enough to force us to abandon the hypothesis that the penny could be expected to come up "heads" half the time.

In the case of frequency curves, our usual hypothesis is that the data are fundamentally normal or that they are fundamentally of the skewed Type III form, or that they are fundamentally of the Poisson type, etc. We realize that any finite number of cases, say 1000 cases, may not conform exactly to the hypothesis, just as in 1000 throws of a penny we might not get exactly 500 heads. But we are interested in comparing what we did get with what we would have got if the hypothesis had been correct.

Perhaps the idea will be simpler if we illustrate it with data that we have already studied. In the preceding chapter, in Sec. 7.8, we fitted a normal curve to data describing the distribution of students' heights. Evidently it was then our hypothesis that

the heights of students were normally distributed. When we got through, we found that the students' heights had not been distributed exactly normally, but that there were differences between the actual figures and those which would have been expected on the basis of our hypothesis that the data were normal. Our findings are given in the first three columns of Table 8.6.

TABLE 8.6.—APPLYING THE CHI-SQUARE TEST TO THE DISTRIBUTION OF STUDENT HEIGHTS

Class Mark (X)	Observed Frequency (f)	Expected Frequency (f')	$(f - f')$	$(f - f')^2$	$(f - f')^2/f'$
156	4	2.5	0.9	0.81	0.073
159	8	8.6			
162	26	23.6	2.4	5.76	0.244
165	53	53.7	− 0.7	0.49	0.009
168	89	97.9	− 8.9	79.21	0.799
171	146	146.8	− 0.8	0.64	0.004
174	188	177.7	10.3	106.09	0.597
177	181	175.3	5.7	32.49	0.185
180	125	141.0	−16.0	256.00	1.816
183	92	91.4	0.6	0.36	0.004
186	60	48.8	11.2	125.44	2.570
189	22	20.9			
192	4	7.4	− 2.9	8.41	0.272
195	1	2.1			
198	1	0.5			
Totals.....	1000	998.2			6.573

These are the first, second, and last columns of Table 7.5.

In applying the chi-square test, if we find some classes with very few cases (as we often do near the ends of a mound-shaped distribution) we lump two or three classes together. The reason for this is obvious. If I want to know whether a penny is well-balanced or not, I know better than to try to decide on the basis of only two or three tosses. After all, if I toss a penny only once it will come up heads or tails 100 per cent of the time, not 50 per cent of the time. Similarly, if some classes have less than 10 cases it is considered good practice to lump enough classes together so that there will be at least 10 cases in each group. Thus in Table 8.6 we have lumped together the top two classes, giving us 12 cases, and the bottom four classes, giving us 28

cases. We have then, in the fourth column, compared the accuracy of our estimates by subtracting the computed from the actual values. These figures in the fourth column show us the amounts of our errors. For example, if the distribution of

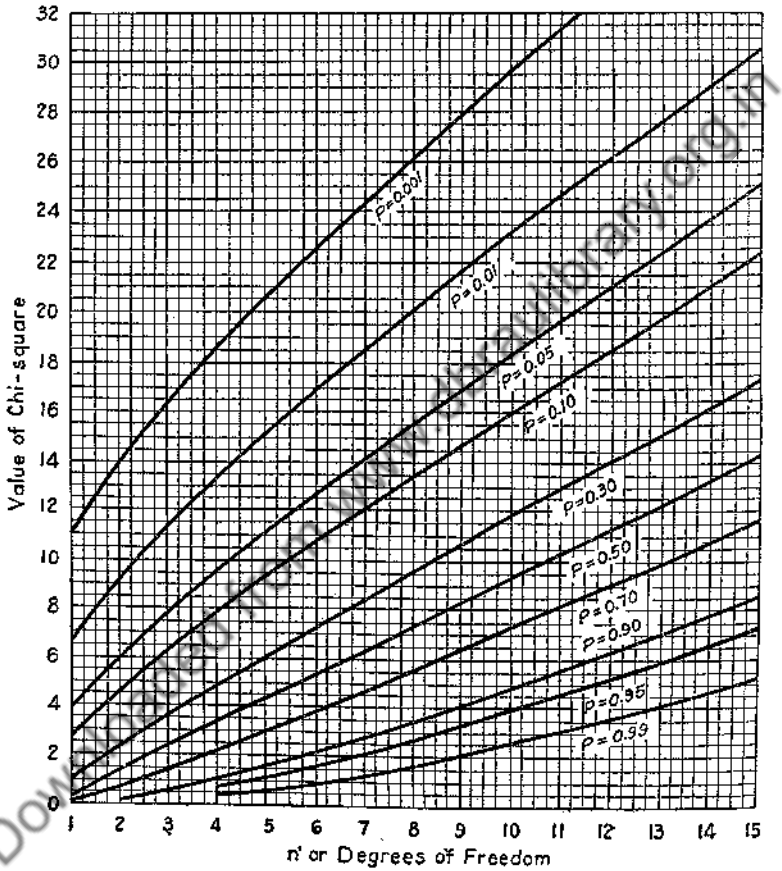


FIG. 8.4.—Values of P corresponding to various values of chi square and various degrees of freedom.

heights had actually been exactly normal, there would have been 141 cases in the 180 class, whereas there were but 125. Thus the error was 16 cases; −16 because the actual distribution was 16 cases short.

It would be unwise to judge the amount of our error by adding these errors of column 4 and using the sum, since every error

might be large, yet the positive and negative signs might balance, giving a small sum. Therefore we use the same procedure that we used under similar circumstances when computing the standard deviation; that is, we square the deviations. This gives us the fifth column. Yet even here we do not have a good measure, because surely we should feel that to come within 10 heads of the expected amount when we tossed a penny a million times was very close, while to miss it by 10 heads when we tossed the penny 22 times would be very far from expectation. We should naturally want to divide by the expected amount to get the result in percentage form. This is what we do in the last column of Table 8.5. Each figure in this column is found by dividing the corresponding figure in the preceding column by the corresponding figure in the third column. When we add these figures in the last column, we get the value of χ^2, which in this case is 6.573.

To interpret this answer, we have to know how large a value of χ^2 can be expected. Obviously this will depend on the number of items in the column. In our illustrative case there are 11 items in the column. But also in our illustrative case we were not content with any random normal curve. On the contrary, we insisted on fitting a normal curve that had the same total number of cases, the same arithmetic mean, and the same standard deviation as our original figures. Thus we forced our normal curve to agree with our original data in three particulars, as we say that we reduced the number of "degrees of freedom" by 3, from the original 11 items to 8. Therefore we let n' equal 8 for our particular example.[1] Let us now consult Fig. 8.4, finding

[1] In applying the chi-square test, we always compare a number of actual frequencies (here 11 of them) with the same number of theoretical frequencies. The number of degrees of freedom, n', is the number of classes the frequencies of which could be filled in at random without violating any of the totals, subtotals, etc. For example, if we take the case of Table 8.5, we want to make the total frequency equal 1000 to correspond with our original data. We could put numbers arbitrarily into any 10 of the 11 classes, but having filled up those 10 classes there is only one number that can be put in the 11 class to give us the right total. We do not have any "freedom" in making our last entry. There are but 10 "degrees of freedom." And when we stipulate that the mean must be 175.335 cm. and that the standard deviation must be 6.582 cm., we lose two more degrees of freedom, leaving us $n' = 8$. In general, when we fit normal curves, n' is smaller by 3 than the number of classes. When we fit a Type III curve, we must also use the

our value of chi square (6.573) on the vertical axis at the left and our value of n' (which is 8) on the base line. We discover that the point on this diagram at the right of $\chi^2 = 6.573$ and directly above $n' = 8$ lies between the line labeled $P = 0.50$ and the one labeled $P = 0.70$. In fact, if we interpolate roughly between these two lines, we might say that this point represents a value of 0.58, since it seems to be just barely closer to the 0.50 line than to the 0.70 line.[1] This is the probability that we should get a fit as bad as or worse than the fit of Table 8.6 by pure chance in drawing a sample of 1000 cases from a universe which was really normally distributed. To put this in other terms, since our value of P is 0.58 in this problem, we can say that 58 per cent of the time we should get fits as bad as this one or worse just by chance even if students' heights in the universe are really normally distributed. This, then, is a reasonably good fit; it is by no means unreasonable to assume that the distribution of heights is really normal, and that the departures from normality among our particular 1000 students are merely chance, haphazard variations.

If our value of P turns out to be very small, it means that our fit was very bad. Suppose, for example, that we had a case where χ^2 was 24 and n' was 3. We see from the chart that we should get a fit as bad as this far less than one time in a thousand by chance. If we look up the values in the tables, we find that such a poor fit would occur by chance only about 25 times in a million. In other words, if we had a very large number of students, and kept on trying again and again taking samples of

skewness; so we lose another degree of freedom, and n' is smaller by 4 than the number of classes. When we fit a Poisson curve, we use only the total frequency and the arithmetic mean; so n' is smaller by 2 than the number of classes. For a complete understanding of degrees of freedom, the student will find it necessary to consult some more advanced work, but these rules of thumb will be reasonably adequate for the types of work here described.

[1] The value of P corresponding to any values of n' and χ^2 can be computed, but the procedure is laborious. The results have fortunately been tabulated, as, for example, in Pearson, *op. cit.*, p. 26, and in Fisher and Yates, "Statistical Tables for Biological, Agricultural, and Medical Research," p. 27, Oliver & Boyd, London, 1938. A copy of the latter table appears in Waugh, *op. cit.* Linear interpolation in such tables gives 0.585 for the value of P, but one can make a visual interpolation on Figs. 8.4 and 8.5 with sufficient accuracy, and we shall continue here to use the rough value of 0.58.

1000 students each, the samples would not always be exactly normal even if the larger group from which they were selected was exactly normal—yet only 25 times in a million trials would any sample drawn at random differ so much from normal as this. Now if the chances are only 25 in a million that this sample came from a normal distribution, we are fairly safe in assuming that the actual distribution is not normal. When P is very small, we decide that our hypothesis is not tenable. Our hypothesis in this case was that the distribution was normal. We actually did get a value of P which led us to believe that the heights might well be distributed normally ($P = 0.58$).

It might be wise to point out here that we can also get a value of P which is so large that we look upon it with suspicion. We know that when we toss two good pennies the chances are that half of them will be heads and half tails. Yet if someone tosses two pennies time after time, and always gets one head and one tail (never getting two heads or two tails), we decide that there must be something wrong. It is "too good to be true." If you saw a man toss two pennies 500 times, and every single time he got one head and one tail, you would (or should) raise some question about it in your mind. Similarly, if every single class in a frequency table has exactly the expected number of cases, or almost exactly the expected number, we are suspicious. Thus if we got values of χ^2 and n' which yielded a value of $P = 0.999$, this would mean that we should get a better fit only once in a thousand by chance. Really what we look for in applying the chi-square test is values of P somewhere around 0.5. Some departure from this is common and raises no question in our minds. Usually if P is between 0.1 and 0.9, we accept it as meaning that our hypothesis is probably tenable. When we get outside these bounds, we begin, perhaps, to get the least bit suspicious of our hypothesis, and we prefer to try more cases to make sure. But we do not usually throw our hypothesis out altogether unless P is smaller than 0.01, and many statisticians would insist on using an even smaller value of P. Neither do we decide that the data are altogether too close to expectation to believe that it happened by chance unless P is greater than 0.99.

Let us apply the chi-square test to one other set of data and interpret the results. In the preceding section, we fitted a

Poisson distribution to data on Supreme Court vacancies (see page 219). The data appear again in Table 8.7. This time we have lumped together the last two classes to get at least 10 cases in each class. Again we find in the fourth column the differences between actual and expected, and in the fifth column

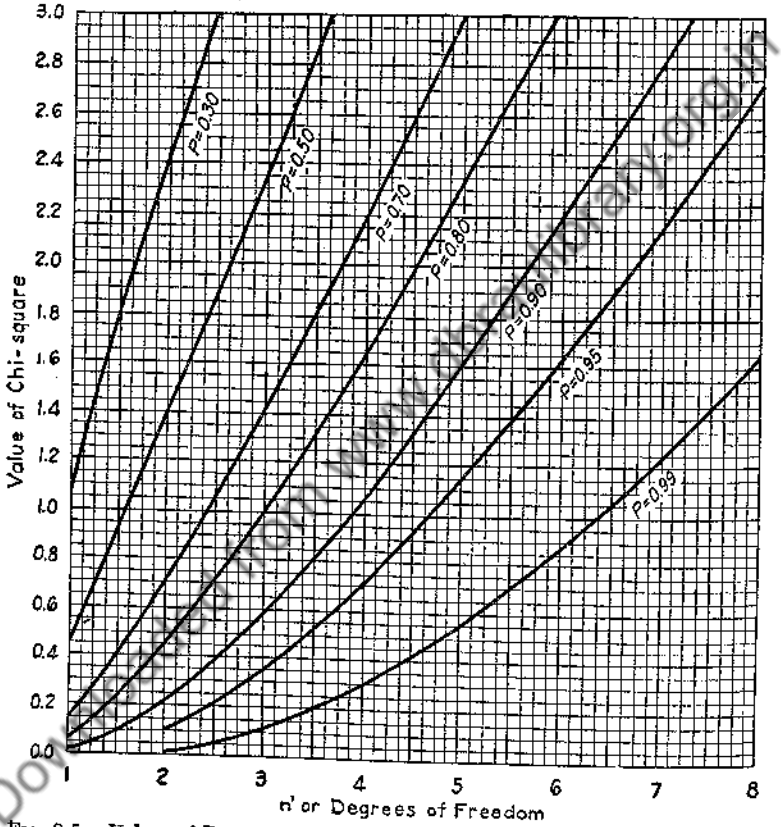


FIG. 8.5.—Values of P corresponding to various values of chi square and various degrees of freedom.

we find the squares of these differences. Then we divide each item of column 5 by the corresponding figure in column 3 to get column 6. The sum of this last column is chi square. In this Supreme Court problem, $\chi^2 = 0.432$. When we look at Fig. 8.4, we find it hard to tell what the value of P is with very much accuracy when n' is 1 and χ^2 is as small as 0.432, so we turn to

Fig. 8.5. This latter figure is merely the lower left-hand corner of Fig. 8.4 enlarged. In Fig. 8.5, we locate the point on the diagram which is directly at the right of the value 0.432 on the vertical scale and directly above the figure 1 on the horizontal scale. We see that this point falls almost exactly on the 0.50 line. Therefore we know that P has a value of approximately 0.50. This tells us that if Supreme Court vacancies were really distributed in a Poisson distribution, and if we selected cases at random, we should get results that fitted worse than these about 50 per cent of the time. The Poisson curve, then, gives a reasonable fit to the data of Table 8.7—just about what we might expect to get in samples from a distribution that was really an exact Poisson distribution.

TABLE 8.7.—APPLYING THE CHI-SQUARE TEST TO A POISSON DISTRIBUTION

X	f	f'	$(f-f')$	$(f-f')^2$	$(f-f')^2/f'$
0	59	58.23	0.77	0.5929	0.010
1	27	29.11	−2.11	4.4521	0.153
2	9	7.28	1.51	2.2801	0.269
3	1	1.21			
Totals.....	96	95.83			0.432

We can now summarize the rules for applying the chi-square test to a frequency distribution as follows:

1. Set down in a column the actual frequencies of the classes in the frequency table.

2. Set down beside them the corresponding frequencies that would be expected if the distribution were normal, skewed Type III, Poisson, or whatever your hypothesis calls for. You now have one column of actual frequencies and one column of computed or estimated frequencies.

3. If any of the classes of actual frequencies contain less than 10 cases, add them to adjacent classes until no class contains less than 10 cases.

4. Subtract each computed frequency from the corresponding actual frequency.

5. Square the differences just obtained.

6. Divide each of these squares by the corresponding computed frequency.

7. Add the quotients just obtained. The sum is χ^2.

8. Find the number of degrees of freedom by subtracting from the number of classes actually used (that is, from the number of entries in the last column of your table) the following number: 2 for a Poisson series, 3 for a normal curve, or 4 for a skewed Type III curve. The remainder is n', or the number of degrees of freedom.

9. In Fig. 8.4 or 8.5, find the point that lies vertically above your value of n' found in step 8 and at the right of the value of χ^2 found in step 7. Read from the lines in the figure the approximate value of P.

10. The value of P found in the preceding step is the probability that you would get by pure chance a worse fit than you did if your hypothesis had been correct. If the value of P is very small, it means that your hypothesis is probably incorrect. If the value of P is very large, it means that the data are suspiciously close to those expected, and that they have probably been computed rather than observed, or in some way adjusted to get them so close to expectation. The value of P can be as large as 1.00 and as small as 0.00. Values between about 0.10 and 0.90 should lead you to believe that there is no reason for abandoning your original hypothesis.

8.13. Suggestions for Further Reading.—The student who wishes to learn more about moments would do well to read Chap. IV of Part I, John F. Kenney, "Mathematics of Statistics," D. Van Nostrand Company, Inc., 1939; or Chap. 9 in G. Udny Yule and M. G. Kendall, "An Introduction to the Theory of Statistics," Charles Griffin & Company, Ltd., London, 1937. A discussion of Sheppard's corrections can be found in H. C. Carver's article on Frequency Curves which is Chap. VII of the "Handbook of Mathematical Statistics," edited by H. L. Rietz, Houghton Mifflin Company, Boston, 1924. Pearson's system of frequency curves, including both the Type III and other types, is discussed in Chap. IV of W. Palin Elderton, "Frequency Curves and Correlation," Layton, London, 1927. A much shorter and simpler treatment is found in Chap. III of C. B. Davenport and Merle P. Ekas, "Statistical Methods in Biology, Medicine and Psychology," John Wiley & Sons, Inc., New York, 1936. References to further discussions of the Poisson curve are given in the footnote on page 216. To these we might add Henry Lewis Rietz, "Mathematical Statistics," pp. 39–45, The Open Court Publishing Company, La Salle, Ill., 1927. The chi-square test was originated by Karl Pearson, and the student who wishes to investigate it further would do well to read his original article, On the Criterion That a Given System of Deviations from the Probable in the Case of Correlated Variables Is Such That It Can Reasonably Be Supposed to Have Arisen from Random Sampling, *Philosophical Magazine*, 5th series, Vol. 50, 1900, pp. 157*ff*. Also helpful is R. A. Fisher, "Statistical Methods for Research Workers," Chap. IV, 3d ed., Oliver & Boyd, London, 1930. For a discussion of the use of another kind of probability paper see F. C. Martin and D. H. Leavens, A New Grid for Fitting a Normal Probability Curve to a Given Frequency Distribution, *Journal of the American Statistical Association*, Vol. 26, new series No. 174, June, 1931, pp. 178*ff*.

EXERCISES

1. Compute the first four moments of the data of Table 5.9, page 124.

2. Table 8.8 shows the number of Kansas towns having various numbers of cream stations.[1] The distribution is obviously skewed. Determine by

[1] From THEODORE MACKLIN, "Efficient Marketing for Agriculture," p. 346, The Macmillan Company, New York, 1922.

inspection whether the skewness is positive or negative. Verify by computing each of the measures of skewness described in Sec. 8.6.

TABLE 8.8.—NUMBERS OF KANSAS TOWNS HAVING VARIOUS NUMBERS OF CREAM STATIONS

Number of Cream Stations	Number of Towns
1	282
2	240
3	151
4	101
5	50
6	11
7	8
8	2
9	1

3. Fit a Poisson curve to the data of Table 8.8.

4. Test the goodness of fit by chi square for your results in Exercise 3. Interpret your results. Does a Poisson curve give a good fit whenever one has a J-shaped distribution?

5. Fit a normal curve to the data of Table 8.5, page 214. Apply the chi-square test, and interpret your results.

6. Apply the chi-square test to the data of Table 8.5, page 214, using the Type III curve. Interpret your results.

7. Comparing your results on Exercises 5 and 6, which fits the data of Table 8.5 better, a normal curve or a skewed Type III curve?

8. Apply the Charlier check to your computations of Exercise 1.

9. Apply Sheppard's corrections to your computations of Exercise 1.

10. Fit a Type III curve to the data of Table 5.9, page 124. It will be necessary to use prepared tables of ordinates of the Type III curve.

11. Describe as well as you can without having seen it the characteristics of a frequency distribution which has the following frequency statistics:

$$\begin{aligned} \bar{X} &= 17.5 \\ \sigma &= 2.7 \\ \alpha_3 &= -0.9 \\ \alpha_4 &= 3.2 \end{aligned}$$

12. Explain why it is that we are suspicious of very large values of P in a chi-square problem; that is, what sort of things other than chance might account for a value of P of 0.9996?

13. In applying the chi-square test, does any value of P tell us that our original hypothesis was correct? What will be true of the value of P if our original hypothesis was correct?

14. On page 219 we computed various numbers of Supreme Court vacancies, using the formula for the Poisson curve. Later we found that, after computing the first of these, the others could have been computed easily by proportion. Check the results of page 219 by using the proportions of page 221.

15. Consider the following five values: 1, 2, 3, 4, 5. They are obviously distributed symmetrically. Compute the first, third, and fifth moments of this distribution and verify the statement made on page 190 that the odd moments of symmetrical distributions equal zero. Compute the second and fourth moments and discover why they do not equal zero. Use the formulas given on page 189.

16. Why is it true, as stated on page 193, that $\alpha_1 = 0$ and $\alpha_2 = 1$ in every distribution?

17. The statement is made on page 204 that Bowley's measure of skewness cannot be greater than +1 or less than −1. Under what circumstances would it equal +1? −1? 0?

18. Compute Sk. for the heights of Harvard students, using Kelley's method based on percentiles (page 204).

19. Find the mode of the data of Table 5.9, page 124, by the method described on page 205.

20. In Sec. 8.5, we evaluated the terms of the point binomial $(q + p)^8$ when p was equal to 0.3. The resulting skewed binomial appears at the lower left of Fig. 8.1. Evaluate the terms of the point binomial $(q + p)^8$ when $p = 0.2$. Plot your results, and compare them with the appropriate section of Fig. 8.1.

21. Suppose that you have fitted some sort of frequency curve to the data of a frequency table. The original table showed the data divided into 12 classes. You compare your computed frequencies with the actual frequencies, and apply the chi-square test. You discover that $\chi^2 = 15$. Interpret this result, using either a table of χ^2 or the charts of Figs. 8.4 and 8.5.

CHAPTER IX

MEASURES OF RELIABILITY

Occasionally a statistician works on a problem of such a nature that he can study all the existing facts—all the data are at his command. For example, if we wish to ascertain the average length of the terms of past Presidents of the United States, we can get data on each and every man who ever occupied the Presidential chair. Thus we can be sure that our average represents the facts (at least if the original data were accurate and if we made no mistakes in computation).

But suppose that we wish to discover the average yield per acre of potatoes in Maine, the average height of college students in the United States, the average weight of male babies at birth, or the average temperature on Aug. 7 in Duluth. The problem is then somewhat different. The chances are that we cannot get figures on every acre of potatoes in Maine or on every college student in the United States or on every male baby born or on every August temperature back to the beginning of time. Some figures we can usually obtain; but almost never can the statistician get figures on every occurrence of the event. A complete enumeration may be too expensive, or it may be entirely impossible regardless of expense, as it would be in the case of Duluth temperatures.

9.1. Sample and Universe.—When it is impossible to get complete data (and it almost never is possible), the statistician finds it necessary to fall back on a *sample*. The entire body of data which describe every occurrence of the event which ever existed is called the *universe*. For example, if we take the problem of determining the average birth weight of male babies, the universe would consist of the weights of all babies who were ever born (that is, of all male babies, or of all babies of the kind in which we were interested). The sample would consist only of those weights of which we actually had records. In such a study it is probable that the sample would be relatively small as compared with the size of the universe; that is, the figures usually

given on average weights of babies are based on observations of a number of births which is small when compared with the total number of births.

What is actually done, of course, is to weigh a relatively small number of babies (a few thousand at the most) and to call their average weight the average weight of babies. Similarly we take temperature records in Duluth for a relatively few years (surely less than a century) and call the average temperature for these years the average Duluth temperature. We find the yield of potatoes on each of a few hundred acres in Maine and call the average of those figures the average yield per acre in Maine. Thus the statistician studies the characteristics of a sample, and then imputes the same characteristics to the universe. We study the heights of 1000 college students. We find the average height of these students, the dispersion in their heights, the skewness in their heights, etc. We then ascribe the same average height, the same dispersion of heights, and the same skewness of heights to students in general.

This habit of studying the peculiarities of a sample and attributing the same peculiarities to the universe seems all the more peculiar when we stop to realize that, if we were to take another group of 1000 students selected at random, they would almost certainly not have exactly the same average height as did the first group measured. If we were to take 1000 such groups of students (1000 groups of 1000, each selected at random) we should find variations in the averages of the groups. Some groups would have higher average heights; some would have lower. Likewise in some groups the standard deviation of heights would be larger than in others, and each of the other measures by which we describe the group of heights would vary from sample to sample.

We are usually interested in the characteristics of the universe. We study past potato prices mainly because we are interested in future potato prices. We study past scholastic records primarily because we want to know what to expect in the future. As Professor Frank Knight has pointed out, the scientist who dissects a dead dog does it because he is interested in live dogs—he cuts open a dog not because he is interested in that dog but because he is interested in the universe of dogs (or the universe of mammals or in life itself). Now if we must expect to find changes in our answers whenever we change our sample, what faith can be put

in our conclusions? How can we ascribe to the universe the characteristics of a sample when we are certain that other samples would have yielded other characteristics? How do we dare say that the average male baby weighs 7.6 lb. at birth when we have weighed but a few thousand babies and when we realize that the average weight of a few thousand other babies would probably be somewhat different?

If the variations in our answers are not due to chance we are not justified in drawing conclusions with regard to the universe. That is to say, if the variations are the result of faulty calculation or if they are the result of the fact that our samples are not really drawn from the same universe, then we cannot speak with any certainty about the universe. But the variations in which we are interested in this chapter—the "errors" which we are studying—are not mistakes in arithmetic. Neither do they arise from the fact that one of our samples is composed entirely of Caucasian babies while our other is a group of Chinese babies. In this case, where the samples do not represent the same thing, one would have to expect variation. But even if all of our samples were of red-headed male babies of native white parentage, we realize that we should get variations from sample to sample. The average weight of one group of 1000 would differ from the average weight of another such group.

If, however, the variations are not due to arithmetical error or to changes in the universe from which the cases are drawn—if, that is, the variations are due to chance—we can say something rather definite about the characteristics of the universe.

9.2. Standard Error of the Arithmetic Mean.—Although it is true that we cannot be sure that the average weight of any particular group of 1000 babies is equal to the average weight of all babies, it can be demonstrated[1] that, if we took an infinite number of samples of 1000 babies each and calculated the mean of each of these samples, the average of these means of samples would be equal to the average weight of all babies, and the standard deviation of the means of these samples would be equal to

$$\sigma_M = \sigma_x \sqrt{\frac{U - S}{S(U - 1)}}$$

[1] C. H. Richardson, "An Introduction to Statistical Analysis," pp. 259–260, Harcourt, Brace and Company, New York.

where σ_M is the standard deviation of the means of the samples, σ_x is the standard deviation of the weights of all the babies in the universe, U is the number of cases in the universe, and S is the number of cases in the sample. Since it is usually true that U is tremendously large when compared with S, we shall not be led far astray in assuming that U is infinite in size. If we do this we simplify the formula greatly, obtaining the following:

$$\sigma_M = \frac{\sigma_x}{\sqrt{S}}$$

Or, since S is the number of cases studied, and since this is usually represented by the letter N, we have

$$\sigma_M = \frac{(\sigma_x)}{\sqrt{N}}$$

It will be noted that the standard deviation given in the numerator of this fraction is the standard deviation of the weights of all babies in the universe. This is, in any actual problem, unknown. We can discover the standard deviation of the weights of the babies in the sample, but there is no way of knowing the facts relative to the universe. As a matter of practice, in the absence of the necessary data describing the universe, we do assume that the standard deviation of the universe is equal to that of the sample. It has been shown empirically that the error made in assuming this is not great. It does, however, make our conclusions approximate rather than exact.

If we can make this assumption that the standard deviation of the weights in the universe is equal to the standard deviation of weights in the sample, we can then make a definite statement relative to the distribution of the means of samples. Let us take the case of students' heights which we have been discussing. We found that the standard deviation of the heights in the sample was 6.58 cm. (page 143). The number of cases studied was 1000. Now if we can assume that the standard deviation of the heights of all college students is 6.58 cm. (which is probably better than guessing at the standard deviation in the universe, but which is nevertheless probably not exactly accurate), we can make a statement about the distribution of means of samples of 1000 students. If we substitute in the last formula, we have

$$\sigma_M = \frac{(\sigma_x)}{\sqrt{N}} = \frac{6.58}{\sqrt{1000}} = 0.208$$

Let us see what this means. If we took many, many samples of 1000 students each, the average heights would not be the same from sample to sample. Some averages would be larger than others. But the standard deviation of these averages would be 0.208 cm.; that is, about two-thirds of all the averages (actually 68.27 per cent of them) would be within 0.208 cm. of the average height of all students in the universe. About 95 per cent of all the samples would have means within twice this distance, or within 0.416 cm. of the mean of all student heights in the universe. And practically never should we get a sample whose mean differed from the mean in the universe by more than 3 (0.208) = 0.624 cm.

In our sample of 1000 students we found an average height of 175.335 cm. (page 87). We do not know that this is the average height of all students; other samples would give other means. But on our assumptions we know that two-thirds of all these other means will be within 0.208 cm. of the mean of the universe. It is therefore true that the chances are 2 out of 3 that this mean is within 0.208 cm. of the mean of the universe. Conversely, it is true that the chances are 2 out of 3 that the mean of the universe is within 0.208 cm. of the mean of this sample. Since the mean of this sample is 175.335 cm., the chances are 2 out of 3 that the mean of the universe is within 0.208 cm. of 175.335 cm.; or, that the mean of the universe is between 175.127 and 175.543 cm. Likewise we can now deduce the facts that the chances are 95 out of 100 (19 to 1) that the mean of the universe is within 0.416 of 175.335, that is, between 175.751 and 174.919 cm. Likewise it is almost certain that the mean of the universe is within 3(0.208) of the mean of the sample, or, that the mean of the universe is between 174.711 and 175.959 cm.[1]

Note, then, that we can make definite statements about the universe *if* we accept the assumptions we have been forced to make, that the standard deviation of the universe is equal to the standard deviation of the sample. We have studied 1000 heights and found an average of 175.335 cm. True, it may be that the average of the universe differs somewhat from this figure. But we are practically certain that the mean of the universe lies between 174.711 and 175.959 cm.; and we can compute the chances that the mean lies at any particular distance from that found in the sample.

[1] The instructor may wish to have the student read Sec. 9.7 at this point.

Let us illustrate again. We wish to know how many shaves a man can get on the average with a given brand of razor blade. Obviously we cannot experiment on every man in the world, nor can we get records on every razor blade of the particular brand sold. But if we can get records on a representative sample of, say, 500 blades, and if we find that these blades gave an average of 14 shaves each, with a standard deviation of 4 shaves, we can go ahead with the computations just illustrated. Assuming that the standard deviation in this sample of 500 blades is equal to the standard deviation of all blades of this brand, we say

$$\sigma_M = \frac{(\sigma_x)}{\sqrt{N}} = \frac{4}{\sqrt{500}} = 0.179$$

We now say that our best estimate of the average number of shaves is 14, but that the chances are 2 out of 3 (2 to 1) that the real average for all blades of this make is between $14 + 0.179$ and $14 - 0.179$, that is, between 13.821 and 14.179. Similarly, we are practically certain that the average for the universe is between $14 + 3(0.179)$ and $14 - 3(0.179)$, or, between 13.463 and 14.537.

It will be noted that we have been making estimates of the probable position of the mean of the universe, basing our estimates on the characteristics of the sample. As has been pointed out, it is impossible to make an exact computation of the location of the mean of the universe, because we have to assume that the standard deviation of the universe is equal to the standard deviation of the sample. If we knew the standard deviation of the universe, we could compute the standard deviation of the distribution of the means of all samples. As it is, we make an estimate of this standard deviation and call this estimate the *standard error* of the mean. The standard error of the mean is, in brief, our estimate of the standard deviation of the distribution of the means of an infinite number of similar samples. It is, of course, the best estimate that we can make; but it is nevertheless only an estimate. We have seen how it is estimated and how the results are interpreted.

If you were asked to state the average wage of barbers in Peru, and if you had data on the wages of but five barbers, you would not feel that you could put much reliance on your average. If you had data on 500 barbers, you would feel that your conclusions

were much more significant. The larger the sample which you studied, the more faith you would have that the average was "correct" (not in the sense that there had been no errors of computation, but in the sense that it was really the average wage of all Peruvian barbers). In other words, you would feel that your computed average became more and more trustworthy as you increased the number of cases from which it was computed. The formula for the standard error of the mean indicates that you would be justified in such a feeling. The larger the number of cases, the smaller (other things being equal) will be the standard error of the mean; that is, if we increase the number of cases on which we base the mean, we decrease the amount by which the means of such samples will fluctuate by chance. It will be noted from the formula, however, that doubling the number of cases will not double the reliability of the mean, since the formula is based on the square root of the number of cases rather than on the number of cases itself. Thus we see that it would be necessary to quadruple the number of cases in order to double the reliability. If we need 5 times the reliability, we must take 25 times as many cases, etc. Thus in our illustrative problem of students' heights we have discovered that the standard error of the mean is 0.208 cm., and that one should not expect to find the mean of any sample of 1000 students farther from 175.335 cm. than 3(0.208) or 0.624 cm. (page 237). But suppose that this still gives too much leeway. You want to be sure of the location of the mean within 0.25 cm. That would mean that you wanted to cut the error down to $\frac{4}{10}$ of its present amount, or that you wanted to increase the accuracy to $1\frac{1}{4}$ of its present amount ($0.624/0.25 = 1\frac{1}{4}$). But $(1\frac{1}{4})^2 = 6.25$, so we know that we should have to take 6.25 times as many cases in order to get the desired accuracy (if the standard deviation still stayed the same). Since we took 1000 cases before, this would mean that we should have to study a total of 6250 cases.

Suppose we apply this latest conclusion and test it out. If the standard deviation did stay the same (6.58 cm.) and if we did have 6250 cases, what would be the standard error of the mean? We substitute in the formula

$$\sigma_M = \frac{(\sigma_x)}{\sqrt{N}} = \frac{6.58}{\sqrt{6250}} = 0.0833$$

The greatest amount by which we should expect the mean of any other group to differ from this would be three times the standard error, or 3(0.0833) or 0.25 cm. This is the accuracy we set out to attain.

To return to our estimate of wages of Peruvian barbers, we have seen that, as one should expect, the accuracy of the mean depends on the number of cases studied (varying as the square root of the number of cases). But if you had learned the wages of 25 barbers and noted that their wages were practically identical, varying by but 2 or 3 centavos, you would rely much more on the average than you would if you found great variation. If there were great differences among the few cases studied, you would fear that there might be considerable variation in the averages of such variable groups. And again you would be right. We note from the formula for the standard error of the mean that great variation in the original figures (as shown by a large standard deviation) will give great variation in the means of samples. Just as might have been expected on purely a priori grounds, then, we find from our formula for the standard error of the mean that we can increase the realiability of the mean by studying more cases, and that the reliability is greater also when the variation among the original figures is small.

9.3. The Probable Error.—The standard error of the mean tells us a range within which two out of three means of samples will lie. Similarly, the standard error of any other measure tells us the range within which two out of three similar measures will lie in other samples. For some unknown reason many people are interested in the range within which the chances are even that the mean (or any other measure) will lie. Thus, if we say that the average number of shaves per razor blade is 14 in a given sample, but that the mean of the universe may differ from this somewhat, many people want to know within what range the chances are even that the true mean lies.

It has been seen that the standard error of the distribution of the means of an infinite number of samples would be 0.179 (page 238). Thus the chances are 2 to 1 that the mean of the universe is within 0.179 of 14 shaves. That is to say, if we kept trying over and over again until we had found the means of an infinite number of samples of 500 blades, we should find variations in the means of the samples. We estimate that the standard

deviation of the means of samples would be 0.179 shaves; two samples out of three would, by our estimate, lie between 14 + 0.179 and 14 − 0.179 shaves. Within what range can we expect the means of half the samples to lie?

This question is easy to answer if we remember our study of the relationship that exists between measures of dispersion (page 146). We discovered that the semi-interquartile range (which includes half the cases) and the standard deviation were related in a normal distribution in this way:

$$Q = 0.6745\sigma$$

The distance that will include half of the cases is just over two-thirds of the standard deviation.

If we remember this relationship we can compute from any standard error the distance within which the chances are even. We know that 0.6745 times the standard error will give us the desired result. This value is known as the *probable error,* and is symbolized by the letters *PE*. Thus the probable error of the mean is 0.6745 times the standard error of the mean, and our formula would be

$$PE_M = 0.6745\sigma_M = (0.6745)\left(\frac{\sigma_x}{\sqrt{N}}\right)$$

If we apply this to the case of students' heights, we get

$$PE_M = 0.6745\sigma_M = (0.6745)(0.208) = 0.14$$

We can now say that the mean height of the students in our sample is 175.335 cm., and that the chances are even that the mean of all students' heights in this universe is between 175.335 + 0.14 and 175.335 − 0.14; that is, the chances are even that the true mean lies between 175.195 and 175.475 cm. It is important to note that the chances are also even that the mean of the universe will lie outside this range. Students occasionally acquire the idea that the probable error sets the limits of error—that it is the same as "possible" error. This is by no means true. Statisticians usually assume as rough limits that chance phenomena will not vary from the mean by more than three standard deviations, which would be 3/0.6745, or almost 4.5 probable errors.

Illustrating once more with the case of razor blades, we have discovered in our hypothetical example that the average blade will last for 14 shaves and that the standard error of this mean is 0.179 shaves. The probable error, then, will be $0.6745(0.179) = 0.121$ shaves. Thus we can say that while the average service of the 500 blades studied was 14 shaves, the chances are even that the mean of the universe (of all blades used) would lie between $14 + 0.121$ and $14 - 0.121$ shaves; that is, the chances are even that the true mean is between 13.879 and 14.121 shaves.

It is common in scientific work to state the probable error of a mean (or of any other measure) immediately after the statement of the mean itself but preceded by a $\pm$ sign. Take, for example, the case just studied. We have said that the mean is 14 shaves and that the probable error of the mean is 0.121 shaves. Commonly this would be written

$$\bar{X} = 14 \pm 0.121$$

Statisticians reading this would understand it to mean that the mean of the sample studied is 14, and that the probable error of this mean is 0.121. It is becoming increasingly common for men to use the standard error rather than the probable error, and there are decided advantages in so doing. When the standard error is given the fact should always be pointed out, however, because it is understood when one sees two figures separated by a $\pm$ sign that the second figure is a probable error. One might, in giving standard errors, make a statement similar to this:

"The average number of shaves per blade and the standard error of the average are 14 ± 0.179."

One could, of course, invert the $\pm$ sign in giving standard errors to distinguish them from probable errors, thus:

$$\bar{X} = 14 \mp 0.179$$

If such a convention could be universally adopted, it would save much explanation. At present, however, such usage would not be understood.

9.4. Other Standard Errors and Probable Errors.—Just as we have discovered that means of various samples differ, so is there variation in the standard deviations of different samples. The values of the median will not be the same for all samples; there

will be variation in the values of the quartiles and of α_3 and α_4. Just as we wish to discover the amount of such variation which can be expected in the cases of means of samples, so we wish to know the reliability of other measures. The concept is the same as that already explained in the case of the mean; hence it will not be necessary to go through most of the explanation again. We shall give the formulas for the standard errors of several of the measures which we have discussed in earlier chapters, and in most cases give an example of the computation.

Standard Error of the Standard Deviation:[1]

$$\sigma_\sigma = \frac{\sigma_x}{\sqrt{2N}} = 0.707107\sigma_M$$

In the case of students' heights we have found the standard deviation of heights to be 6.58 cm. with 1000 cases (page 143). Substituting in the formula, we get

$$\sigma_\sigma = \frac{6.58}{\sqrt{2000}} = 0.147$$

We had already computed the σ_M as 0.208 (page 236). Thus we can use the second formula above to get

$$\sigma_\sigma = 0.707107(0.208) = 0.147$$

This means that the chances are 2 out of 3 that the true standard deviation of the universe is within 0.147 cm. of the standard devi-

[1] This formula for the standard error of the standard deviation is strictly correct only in a normal distribution. F. C. Mills states (in his "Statistical Methods," p. 556, Henry Holt and Company, Inc., New York, 1924) that one can determine the standard error of the standard deviation of any distribution, normal or otherwise, by the formula

$$\sigma_\sigma = \sqrt{\frac{v_4 - v_2^2}{4v_2(N)}}$$

The v's in this formula are the higher moments about the mean as found in the preceding chapter (pp. 192*ff.*). If Sheppard's corrections are used, the corrected moments should be substituted. If we substitute the values of the moments of the heights of Harvard students from p. 192, this formula becomes

$$\sigma_\sigma = \sqrt{\frac{5508.567 - 43.353^2}{(4)(43.353)(1000)}} = 0.145$$

In this case, since the distribution of heights is practically normal, the result obtained by this method is almost identical with the result obtained by the method more commonly used.

ation of the sample, that is between 6.58 + 0.147 and 6.58 − 0.147, or between 6.433 and 6.727 cm. It is practically certain that the true standard deviation lies between 6.58 + 3(0.147) and 6.58 − 3(0.147), or between 6.139 cm. and 7.021 cm. We have seen that the probable error of any measure is 0.6745 times the standard error. Thus we can say that

$$PE_\sigma = 0.6745\left(\frac{\sigma_x}{\sqrt{2N}}\right) = \frac{(0.6745)(6.58)}{\sqrt{2000}} = 0.099$$

Hence we can say the chances are even that the true standard deviation of the universe lies between 6.58 + 0.099 and 6.58 − 0.099, or that it lies between 6.679 and 6.481 cm. The standard deviation would usually be written in conjunction with its probable error, thus:

$$\sigma = 6.58 \pm 0.099 \text{ cm.}$$

Standard Error of the Median:

$$\sigma_{\text{Med.}} = \sqrt{\frac{\pi}{2N}}\sigma_x = 1.25331\sigma_M$$

We have found that the median height was 175.28 cm. (page 94). We have also found that σ_M is 0.208 cm. (page 236). Substituting in the formula above, we have

$$\sigma_{\text{Med.}} = 1.25331(0.208) = 0.261$$

The chances are 2 out of 3 that the true median lies between 175.28 + 0.261 and 175.28 − 0.261, or between 175.019 and 175.541 cm. It is almost certain that the median of the universe lies between 175.28 + 3(0.261) and 175.28 − 3(0.261); that is, between 174.497 and 176.063. The probable error is, of course, 0.6745 times the standard error, so the chances are even that the true median lies within (0.6745)(0.261) of the median of the sample. The chances are even that it lies within 0.176 of the median of the sample, or between 175.104 and 175.456.

Standard Errors of the Alphas:

$$\sigma_{\alpha_3} = \sqrt{\frac{6}{N}}$$

$$\sigma_{\alpha_4} = \sqrt{\frac{24}{N}} = 2\sigma_{\alpha_3}$$

We discovered the following values:

$$\alpha_3 = -0.034 \text{ (page 193)}$$
$$\alpha_4 = 2.926 \quad \text{(page 193)}$$

The standard errors of these values are

$$\sigma_{\alpha_3} = \sqrt{\frac{6}{1000}} = 0.0775$$
$$\sigma_{\alpha_4} = 2\sigma_{\alpha_3} = 2(0.0775) = 0.1550$$

The probable errors of these measures are

$$PE_{\alpha_3} = 0.6745(0.0775) = 0.0523$$
$$PE_{\alpha_4} = 0.6745(0.1550) = 0.1045$$

Thus we can write the measures

$$\alpha_3 = -0.034 \pm 0.0523$$
$$\alpha_4 = 2.926 \pm 0.1045$$

It will now be recalled that we used these values to determine whether or not the distribution of students' heights was normal. In a normal distribution these values would have been $\alpha_3 = 0$ and $\alpha_4 = 3$. We now note that in half the cases the values of α_3 will fall within 0.0523 of -0.034; that is, between -0.0863 and $+0.0183$. The value of the normal distribution would be 0. The value of our sample is -0.034, which differs from normal by 0.034, or by $(0.034/0.0775)\sigma$. We have earlier discovered how to compute the chances of such an occurrence (page 170). We have here a deviation of $(0.034/0.0775)\sigma$ or of 0.44σ. In 50 per cent of the cases our deviations would be in the other direction, and the tables show us (see page 509, Appendix I) that between the mean and 0.44σ from the mean lie 17 per cent more of the cases. Altogether, then, there are 67 per cent of the cases with less deviation than this. Hence 33 per cent of the cases would deviate more, or in 33 per cent of the cases the value of α_3 would have been either 0 or positive. In other words, it is quite possible that the distribution from which this sample was drawn was not really skewed. If the value of α_3 differs from 0 by more than three times its own standard error, we should say that such a skew could not be expected to arise by chance in a sample drawn at random from a symmetrical universe. In other words, when the value of α_3 differs from 0 by more than three

times its standard error, we conclude that there is good evidence that the universe is itself skewed.[1] When, as in the present case, the value of α_3 differs from 0 by less than three times its standard error, we are not certain that the universe was itself skewed. It is quite possible that one would by chance draw a sample with as much skewness as the present one from an unskewed universe. In fact we should get as much skewness as that of our present sample in about 33 per cent of all chance samples from unskewed universes.[2] We conclude, then, that there is no certain indication of skewness in students' heights in the data of our sample.

Similarly we find that the value of α_4 is 2.926, although the value for a normal distribution is 3.0. The value in our sample differs from the normal by $2.926 - 3 = -0.074$. The standard error itself has a value of 0.1550. Thus the difference is

$$-\frac{0.074}{0.1550} = -0.48\sigma$$

By pure chance we should get an absolute difference less than this in 50 + 18.4 per cent = 68.4 per cent of the cases. This means that we should get samples with values of α_4 as small as this, or smaller from distributions in which there was actually no kurtosis, in 31.6 per cent of the cases merely through the operation of chance. Since this is true, we see that the value of α_4 may well be as high as 3 in the universe even though it is but 2.926 in the sample; that is, there is no evidence that kurtosis exists in the universe unless the value of α_4 differs from 3 by an amount which is more than three times the standard error of α_4. Here the value of α_4 differs from 3 by an amount equal to but 0.48 times the standard error of α_4.

Standard Error of a Relative Frequency (Percentage):

$$\sigma_{\%} = \sqrt{\frac{pq}{N}} = \frac{\sqrt{pq}}{\sqrt{N}}$$

[1] In Fig. 8.2, p. 202, are two skewed distributions. In the upper distribution the value of α_3 is +0.307 and its standard error is 0.0142. The fact that α_3 is 21 times its standard error leads us to believe that the wage distribution from which this sample was drawn was almost certainly skewed positively. In the lower distribution the value of α_3 is −0.518 and its standard error is 0.0438. Since the value of α_3 is over 11 times the value of its standard error, we conclude that the egg-production data from which the sample was drawn were almost certainly negatively skewed.

[2] See Table 9.1, p. 259.

Dr. Charles V. Chapin of Providence, Rhode Island, states[1] that 25.45 per cent of 53,280 people who were exposed to diphtheria between 1889 and 1915 caught the disease. What is the standard error of this figure, 25.45 per cent? We find it from the formula. We know that p is the probability that the event will happen and q is the probability that it will fail to happen (see page 156). Here our sample shows that $p = 0.2545$ and q must equal 0.7455. (Since 25 per cent of the exposed persons were afflicted, we know that $^{25}/_{100}$ of them were afflicted, or 0.25 of the total number. Thus $p = 0.25$—or, to be exact, 0.2545.) The number of cases studied is given as 53,280. Substituting in the formula, we have

$$\sigma_{\%} = \sqrt{\frac{(0.2545)(0.7455)}{53{,}280}}$$
$$= 0.0019$$

Thus 0.2545 did fall ill, and the standard error is 0.0019. Putting both figures back in percentage terms by multiplying by 100, we say that the attack rate was 25.45 per cent, with a standard error of 0.19 per cent. Practically never should we expect to get an attack rate higher than 25.45 per cent + 3(0.19 per cent), or 26.02 per cent. Practically never should we expect to get an attack rate lower than 24.88; that is, if we continued to take samples from the same universe (samples of people of the same age getting the same kind of medical care, leading the same kinds of lives, etc.) we should expect always to find that between 24.88 per cent and 26.02 per cent of the people exposed would come down with diphtheria. The probable error is, of course, 0.6745 times the standard error. This gives us

$$PE_{\%} = (0.6745)(0.0019) = 0.00128$$

We could, then, state the attack rate thus:

Attack rate = 25.45 per cent ± 0.128 per cent

Standard Error of the Semi-interquartile Range:

$$\sigma_Q = 0.7867\sigma_M$$

We have found that the standard error of the mean is 0.208 cm. in the case of heights of students. The semi-interquartile range

[1] Quoted in WHIPPLE, "Vital Statistics," p. 376, John Wiley & Sons, Inc., New York, 1933.

is 4.44 cm. (page 129). We now see that the standard error of this semi-interquartile range is

$$\sigma_Q = 0.7867(0.208) = 0.164$$

The probable error of the semi-interquartile range is

$$PE_Q = (0.6745)(0.164) = 0.11$$

Thus the semi-interquartile range can be written as

$$Q = 4.44 \text{ cm.} \pm 0.11 \text{ cm.}$$

Standard Error of the Average Deviation:

$$\sigma_{AD} = 0.605\sigma_M$$

We discovered in our illustrative example that the average deviation of the students' heights in our sample is 5.28 cm. (page 134). We have already seen that the standard error of the mean is equal to 0.208 cm. (page 236). Substituting in the formula, we find

$$\sigma_{AD} = 0.605(0.208) = 0.126 \text{ cm.}$$

As before, the probable error of the AD is

$$PE_{AD} = (0.6745)(0.126) = 0.085$$

Hence we might write the average deviation thus:

$$AD = 5.28 \text{ cm.} \pm 0.085 \text{ cm.}$$

Standard Error of Either Quartile:

$$\sigma_{Q_1} = \sigma_{Q_3} = 1.36263\sigma_M$$

This means for our problem:

$$\sigma_{Q_1} = \sigma_{Q_3} = (1.36263)(0.208) = 0.284$$

Similarly the probable error of either quartile is

$$PE_{Q_1} = PE_{Q_3} = (0.284)(0.6745) = 0.191$$

We can then write the quartiles (taking the values of the quartiles of the sample from page 129) as follows:

$$Q_1 = 170.95 \pm 0.191$$
$$Q_3 = 179.84 \pm 0.191$$

Standard Error of β_2.—We have seen on page 206 that β_2 is identical with α_4. Hence the formula given on page 244 for the standard error of α_4 will apply also to β_2.

Standard Error of Measures of Skewness.—On page 206 we used as a measure of skewness the value

$$\text{Sk.} = \frac{\sqrt{\beta_1}\,(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

The standard error of this value is

$$\sigma_{\text{Sk.}} = \frac{1.225}{\sqrt{N}}$$

This formula does not, of course, hold for other measures of skewness. If we apply this formula to our illustrative case, the value of skewness is found to be 0.0179 (page 206). It is based on 1000 cases. Hence the standard error is

$$\sigma_{\text{Sk.}} = \frac{1.225}{\sqrt{1000}} = 0.0386$$

The probable error is $0.6745(0.0386) = 0.026$. Thus we might well write the result of measuring skewness by this method in this fashion:

$$\text{Sk.} = 0.0179 \pm 0.026$$

When one notes that the probable error in this case is greater than the measure of skewness itself, it is evident that we cannot be sure that there was skewness in the universe. Unless this measure of skewness differs from 0 by an amount greater than three times its standard error, we must say that we are not justified in assuming that skewness existed in the universe from which our sample was drawn.

This same formula for the standard error of the measure of skewness can be applied when skewness is measured on the basis of the difference between the mean and the mode, according to the formula on page 201.

Standard Error of the Coefficient of Variation:

$$\sigma_V = \frac{V}{\sqrt{2N}}\sqrt{1 + 2\left(\frac{V}{100}\right)^2}$$

We discovered that the coefficient of variation of the students' heights in the sample is 3.75 per cent (page 152). The standard error of this figure would, then, be

$$\sigma_V = \frac{3.75}{\sqrt{2000}}\sqrt{1 + 2\left(\frac{3.75}{100}\right)^2}$$
$$= 0.084(\sqrt{1.0028}) = 0.084$$

The probable error would, then, be

$$PE_V = (0.6745)(0.084) = 0.0566$$

The coefficient of variation would be written

$$V = 3.75 \text{ per cent} \pm 0.0566 \text{ per cent}$$

If the coefficient of variation is less than 10 per cent, we can approximate its standard error closely enough by the formula

$$\sigma_V = \frac{V}{\sqrt{2N}}$$

Standard Error of the Difference between Two Measures.—In statistical work we are very often interested in differences and in their significance. Suppose, for example, that we wish to discover whether or not there is a significant difference between the number of spears borne on male and on female asparagus plants. Haber tells us[1] that the average number of spears on the male plants which he studied was 15.37 and the average number of spears per female plant was 9.39. We know that if he had taken another sample of male plants the average number of spears might have differed somewhat from 15.37, and in another sample of female plants the average might have differed from 9.39. We are interested in the difference between the two means, 15.37 and 9.39. The difference is $15.37 - 9.39 = 5.98$ spears. With other samples yielding other means, could it be expected that there would continue to be a difference of this kind? Could we expect that the means would still show the male plants bearing more spears on the average than the females? We discover the answer to this question by computing the standard error of the difference by means of the formula

$$\sigma_{\text{Diff.}} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

[1] E. S. HABER, *Journal of Agricultural Research*, Vol. 45, July, 1932, p. 103.

in which $\sigma_{\text{Diff.}}$ represents the standard error of the difference between two measures, σ_1 represents the standard error of the first measure, and σ_2 the standard error of the second measure.[1] In our case this means that we must have the standard errors of the two measures whose difference we are studying. These are given by Haber as 0.88 for males and 0.05 for females. (Note that these are *not* the standard deviations of the numbers of spears, but the standard errors of the two means found by the methods described on page 236.) Now that we know the two means and their standard errors, we can proceed to discover the standard error of the difference between the means.

	Average Number of Stalks	Standard Error of $\bar{X}$
Male plants.....	15.37	0.88
Female plants.....	9.39	0.05

The difference itself is $15.37 - 9.39 = 5.98$. Its standard error is

$$\sigma_{\text{Diff.}} = \sqrt{0.88^2 + 0.05^2} = \sqrt{0.7744 + 0.0025}$$
$$= \sqrt{0.7769} = 0.881$$

The difference is 5.98, and its standard error is 0.881. Its probable error is $(0.6745)(0.881) = 0.595$. Thus the difference can be written

$$\text{Diff.} = 5.98 \pm 0.595$$

It will be noted that the difference is equal to 10 times its probable error and 6.8 times its standard error. We should almost never get, by pure chance, a difference equal to more than three

[1] This formula is accurate in the form given here only if the two measures whose difference is being studied are uncorrelated. If correlation exists between them the formula becomes

$$\sigma_{\text{Diff.}} = \sqrt{\sigma_1^2 + \sigma_2^2 - 2r_{12}\sigma_1\sigma_2}$$

When there is no correlation, this formula reduces to the one above. The nature of correlation is discussed in Chap. XIII; the formula given in this footnote can be understood after that chapter has been mastered. In the meantime, this is one of the cases in which the simpler formula may correctly be used; we shall later see why.

times its standard error (only about three times in a thousand).[1] A difference 6.8 times its standard error would not arise once in a billion times. In other words, it is inconceivable that this difference between the numbers of stalks on male and on female plants should arise by chance from data whose averages are really the same. It must be true that in the entire universe of male asparagus plants the average number of stalks per plant is greater than is the average in the entire universe of female asparagus plants.

If, then, we find a difference between two measures which is greater than three times the standard error of the difference, we say that such a difference would not be expected to arise from pure chance. It must have arisen because the samples were drawn from two universes whose means were different.

Just as we have found the standard error of the difference between two means, we can find the standard error of the difference between any other two measures whose standard errors are themselves known. If we have the two measures and their standard errors, we can compute the standard error of their difference by the formula given on page 250. We shall show one more example. A group of 1150 Wellesley freshmen had an average height of 64.13 in., with a standard deviation in heights of 2.24 in. A group of 1017 Hollins College freshmen had an average height of 63.86 in. with a standard deviation in heights of 2.09 in.[2] From these data, by the formula given on page 243, we can compute the standard errors of the two standard deviations. They are

$$\sigma_{\sigma_1} = \frac{2.24}{\sqrt{2300}} = 0.0466$$

$$\sigma_{\sigma_2} = \frac{2.09}{\sqrt{2034}} = 0.0464$$

The difference between the two standard deviations is $2\,24 - 2.09 = 0.15$ in. The standard error of the difference is

$$\begin{aligned} \sigma_{\text{Diff.}} &= \sqrt{0.0466^2 + 0.0464^2} \\ &= \sqrt{0.00218 + 0.00215} \\ &= \sqrt{0.00433} = 0.0658 \end{aligned}$$

[1] See pp. 255*ff*.

[2] G. L. PALMER, *Journal of the American Statistical Association*, Vol. 24, March, 1929, p. 42.

We now compare the difference with its standard error. The difference is 0.15 in. Its standard error is 0.0658 in. The difference between these two standard deviations is 2.28 times its standard error. If the universe from which the Wellesley girls and the universe from which the Hollins girls were drawn really had the same variability (that is, if the standard deviations of the two universes were the same), then half the time when we drew samples we should find the standard deviation of the Hollins sample larger than that of the Wellesley sample. Moreover, reference to Appendix I, page 509, reveals that another 48.9 per cent of the samples would have Wellesley standard deviations above but within 2.28σ of the Hollins standard deviation. Hence in 98.9 per cent of the cases differences would be found smaller than this. By pure chance the Wellesley standard deviation would exceed the Hollins standard deviation by this amount or more in 1.1 per cent of the cases, or 11 cases out of 1000.[1] The statistician would say that 11 cases in 1000 is a significant proportion, and that it is not certain that the Wellesley girls were really more variable in height than the Hollins girls. It is quite possible that this case is one of the 11 cases in which such a difference would arise by chance. We could tell only by taking more cases and seeing whether the difference persisted. Had the difference between the standard deviations exceeded three times its standard error, we should have said it was evident that the Wellesley girls came from a different universe than did the Hollins girls—from a universe in which the standard deviation in heights was certainly larger than in the Hollins universe.

Since the standard error of this difference is 0.0658 in., the probable error is $(0.0658)(0.6745) = 0.0444$ in. We should write the difference thus:

$$\text{Diff.} = 0.15 \text{ in.} \pm 0.0444 \text{ in.}$$

Standard Error of the Sum of Two Measures.—If two measures (such as two averages or two standard deviations) of uncorrelated data are added, the standard error of their sum is given by the formula

$$\sigma_{\text{Sum}} = \sqrt{\sigma_1^2 + \sigma_2^2}$$

[1] This result can be approximated directly from Table IV of the Appendix, where we see that one would exceed 2.3 standard deviations 107 times out of 10,000.

This is seen to be the same as the formula for the standard error of the difference when the measures are for uncorrelated data. If the data are correlated, the formula for the standard error is the same as that given in the footnote on page 251, except that the minus sign under the radical is changed to a plus sign.

9.5. Modifications for Small Samples, Etc.—We have been computing standard errors on the assumption that the standard deviation in the sample can be substituted without modification for the standard deviation of the universe. In other words, those formulas for standard errors which are based on standard deviations are based on the standard deviation in the universe, yet we use the standard deviation of the sample. To be strictly correct we should not do this the way we have done it. For example, we have given the formula for the standard error of the mean (see page 236) as

$$\sigma_M = \frac{\sigma_x}{\sqrt{N}}$$

If we are to use the standard deviation of the sample instead of the standard deviation of the universe, we should really divide by $\sqrt{N-1}$ instead of by $\sqrt{N}$. When N is large it makes little difference in our answer. Thus in the case of students' heights we have computed the standard error of the mean (page 236) to be

$$\sigma_M = \frac{\sigma_x}{\sqrt{N}} = 0.208$$

If we had substituted $N - 1$ for N, we should have had

$$\sigma_M = \frac{\sigma_x}{\sqrt{N-1}} = \frac{6.58}{\sqrt{999}} = 0.208$$

Unless we carried our answers to more significant figures this would have no effect. But when the number of cases is smaller, there will obviously be more effect from the subtraction of one. Subtracting one from 1000 makes a reduction of $\frac{1}{10}$ of 1 per cent. Subtracting one from 10 makes a reduction of 10 per cent. Hence when one is working with small samples and is using the standard deviation of the sample in place of the standard deviation of the universe, one should use $N - 1$ instead of N. This

would change the formula for the standard error of the standard deviation from

$$\sigma_\sigma = \frac{\sigma_x}{\sqrt{2N}}$$

to

$$\sigma_\sigma = \frac{\sigma_x}{\sqrt{2(N-1)}}$$

and similarly for other formulas.

McCall suggests[1] that one may neglect this refinement when dealing with over 30 cases. He suggests that we use $N - 1$ in place of N when the number of cases is between 20 and 30, that we use $N - 2$ instead of N when the number of cases is between 10 and 20, and that we use $N - 3$ in place of N when N is less than 10. As a matter of fact the student should understand very clearly that the reliability of all measures is exaggerated when they are based on a small number of cases. The usual formulas for measures of reliability (standard errors) should be used only when 30 or more cases are involved, and preferably only when 50 or more are involved. Certainly when one runs over 100 cases, the change that is made in the answers by the refinements here mentioned is too small to be worth attention. The inaccuracies of the original data are too great to warrant such minor adjustments.

9.6. The Significance of Differences.—There is no difference so large that it could not occur by chance in two samples drawn from the same universe. Conceivably two such samples might differ by any amount in their means, in their standard deviations, or in any other way. Yet some happenings are so unlikely that their occurrence can hardly be looked on as a chance phenomenon. If someone throws two dice 15 times and gets a total of 7 spots on each throw, one wonders if chance is the only force that is operating. It is possible that an honest man should throw one hundred 7's in succession with honest dice, but it is so unlikely that most opponents would decide long before the 100th throw that either the thrower or the dice were dishonest. The fact that an event can happen by chance does not mean that we are willing to ascribe such a happening to chance when it occurs;

[1] W. A. McCall, "How to Experiment in Education," The Macmillan Company, New York, 1923.

If its happening as a result of chance is extremely unlikely, we usually decide that some factor other than chance has played a part.

This is true not only with dice but with all events in which chance operates. We realize, for example, that even if there were no difference between the body weight of male and female rats, were we to take a sample of males and a sample of females and compute the mean weight for each sample the two means would probably differ somewhat. And actually there is no limit to the amount of the difference that might arise from chance. But some differences would arise so seldom by chance that, if they arose in our samples, we should be led to believe that some factor other than chance was responsible. We should decide that the rats had been drawn from different universes—that the universe of male rats differed significantly in this respect from the universe of female rats. Hatai weighed 45 male rats and 37 female rats and found the following:[1]

	Male Rats	Female Rats
Number of cases	45	37
Mean weight (grams)	214.9	167.3
σ of weights (grams)	52.89	20.47

From these data we compute the following standard errors of the means by the formula given on page 254:

Males	7.9
Females	3.37

The difference between the two means is 214.9 − 167.3 = 47.6 grams. The standard error of the difference, computed according to the formula on page 250, is 8.59. Since the difference is 47.6 and its standard error is 8.59, the difference is 5.55 times its standard error. Could such a difference arise by chance? Yes, any difference could arise by chance. But if the male and female rat universes were the same, in half the cases the mean of the male sample would be less than the mean of the female sample. And the table in Appendix I, page 509, tells us that

[1] Quoted in H. H. DONALDSON, The Rat, *Wistar Institute of Anatomy and Biology Memoir No.* 6, Philadelphia, 1924, p. 50.

between the mean and 5σ from the mean there are included another 49.99997133 per cent of the cases. Thus even if we go up only to 5σ (and this problem goes beyond this to 5.55σ), we have included 99.99997133 per cent of the cases. We should get a difference over five times its standard error from pure chance but once in over 1,700,000 times (see Appendix IV). To believe that this difference between mean weights of male and female rats arose from chance is somewhat like believing that an honest man can throw honest dice and get 25 successive 7's. It could happen, but would happen so seldom that we are prone to ascribe its occurrence to factors other than chance. Thus here we should say that, although male and female rats could yield such different means by chance, we are forced to decide that such a difference did arise from some other source—namely, from the fact that the universes were different.

How much of a difference shall we allow to exist before we say that chance did not account for it? This is like asking how many times you will allow a man to throw 7's with dice before you will look for non-chance explanations. Men differ in their credulity. In a gambling game their credulity depends somewhat, perhaps, on whether they have something at stake. Some men would be suspicious of dishonesty in the throwing of 25 successive 7's; others would look upon the affair as unusual but still due to chance. The same is true in statistical work. People differ in their credulity. It is possible even that some people might believe that a difference such as the one we have just discovered between rats' weights arose by chance. If someone tosses a penny and it falls with "heads" uppermost, no one would rule out the possibility that it happened by chance, even though there was an even chance that it would come up "tails." Most people would not rule out chance if something happened against which the odds were 2 to 1. The statistician does not rule out the possibility of chance even when things happen against which the odds are 100 to 1.[1] He takes an

[1] Statisticians, like other people, differ in credulity. Some call a statistical result "significant" if it would arise by chance only once in 100 times. Others set other points. Any such point is arbitrary—as are the points which we have set here. Our point, three standard errors, errs if at all in requiring too much before the possibility of chance is ruled out. It is, however, the most commonly accepted point among American statisticians.

arbitrary point, in order that others may understand what he is doing, and says that any difference is "significant" (that is, significant of differences in the universes and not the result of chance differences in the samples) if it exceeds three times its standard error. If chance alone were operating, we should get differences smaller than this in 99.74 per cent of the cases, or in 9974 cases out of 10,000. We should get differences this large or larger but 26 times out of 10,000 (see Appendix III). The chances against such an occurrence are greater than 300 to 1.

We see, then, that, while any difference may arise by chance, when a difference is over three times its standard error the likelihood that it did arise from chance is so small that the statistician feels justified in neglecting it and in assuming that the difference is significant of the operation of forces other than chance. Likewise if we say that the value of α_3 is "significantly" different from zero we mean that, although such a value could have arisen in a sample drawn from a universe in which the value of α_3 was zero, nevertheless the likelihood of such an occurrence is so small that we neglect it and assume that some forces other than chance operated. Whenever the difference between the value of α_3 and zero is greater than three times the standard error of α_3, we come to this conclusion that the difference is significant. One might have chosen the arbitrary point of 2.98 times the standard error or of 3.7 times the standard error. Actually it is more convenient to take an even number of standard errors and base our test of significance on it.

The reader should remember that, when a difference is less than three times its standard error, there is no guarantee that it did arise from chance. On page 253 we found that the difference between two standard deviations was equal to but 2.28 times its standard error. The odds are better than 40 to 1 against the occurrence of a difference as great as this. Yet in spite of the fact that there may well have been a difference between the standard deviations of the two universes from which these samples were drawn, the statistician does not feel sure of it when the odds against it are but 40 to 1. Odds of 40 to 1 are not certainty to a statistician. He would say in this case that the proper thing to do would be to measure more Wellesley and Hollins freshmen to see whether the difference in standard deviations continued to exist. If it did not, then it

would signify that the difference had arisen from chance; if it did continue to persist as more and more cases were added (that is, as N increased in each sample), the standard errors of the standard deviations would fall, the standard error of the difference would fall, and ultimately the difference would be three times its standard error. Then the statistician would decide that he had gone far enough, and that he would be safe in ascribing the difference to forces other than chance.

In this connection it may be helpful for the student to see just what the chances are of getting a difference by chance which is various numbers of times its standard error. The chances can be easily computed from tables such as that on page 509 of Appendix I. In Table 9.1 the first column is the difference divided by its standard error, and in the other column are the approximate chances *against* the occurrence of such a difference by pure chance.

TABLE 9.1.—CHANCES AGAINST THE OCCURRENCE OF A DEVIATION FARTHER FROM THE MEAN THAN THE DISTANCE STATED

Difference / ($\sigma_{Diff.}$)	Chances
0.6745	1 to 1
1.0	2.15 to 1
1.5	6.48 to 1
2.0	21.0 to 1
2.5	79.5 to 1
3.0	369 to 1
3.5	2,150 to 1
4.0	15,800 to 1
4.5	147,000 to 1
5.0	1,740,000 to 1

Perhaps it is as well to point out here that a *large* difference, an *important* difference, and a *significant* difference are not at all the same concept. When we say that a difference is large, we are referring to the actual size of the remainder left after subtraction. Thus a difference of 5 lb. is 16 times as large as a difference of 5 oz. Yet the 5 lb. difference may not be significant because its standard error is large, while the 5-oz. difference may be significant because its standard error is small. When we say that a difference is significant, we mean that we are con-

vinced it did not arise by chance, but reflects a difference which actually exists in the universes from which the samples were drawn. And finally, a difference might be both large and significant, yet unimportant. Whether a difference is important or not depends on what it can contribute toward an explanation of the problem being studied. A difference which does not concern the statistician in any way—which raises no problems for him and does not help him to explain the phenomena that he is considering—is unimportant. The student must learn that statistical manipulations never make anything important, that they are a means and not an end. To be sure, if a difference is important to us we are likely to wish to test its significance, and if it is large it is more likely to be significant than if it is small. The three ideas are related—but they are by no means the same.

9.7. Fiducial Probability and the Confidence Interval.—The use of terms in the field of probability is by no means standardized. Authorities disagree even as to the definition of the term "probability" itself, finding it one of the hardest of concepts to define without circularity. The student should be warned, therefore, that the followers of some schools of thought would object to certain of the expressions used in this chapter, preferring to state the conclusions of reliability analysis in other terms.

For example, in Sec. 9.2, page 236, we found that the arithmetic mean height of a group of students was 175.335 cm. and that the standard error of this mean was 0.208 cm. We interpreted this by saying that, although we could not know for certain the exact size of the average height in the universe, nevertheless the chances were 2 out of 3 that this mean height lay in the interval from 175.127 to 175.543 cm.; the chances were 95.5 out of 100 that it lay in the interval from 174.919 to 175.751 cm.; and the chances were 99.7 out of 100 that it lay in the interval from 174.711 to 175.959 cm. We could summarize those conclusions as follows:

Interval of Height (cm.)	Probability That Mean of Universe Lies in This Interval
175.127–175.543	0.6827
174.919–175.751	0.9545
174.711–175.959	0.9973

Now as we have just said, some authorities would say that these statements are incorrect. If you were to ask them, "What

is the probability that the arithmetic mean in the universe lies between 175.127 and 175.543 cm.?" they would answer that it is not a question of probability at all, but a matter of fact. Either the mean of the universe does lie in this range, in which case the probability is 1, or it does not lie in this range, in which case the probability is 0. They would argue that there is no probability of 0.6827, but a probability which is either zero or unity—or that it is not really a case of probability in the strict sense at all.

We could argue, of course, that the same stand can be taken on any probability problem. What is the probability that if I toss a coin it will come up "heads"? I can argue that if I do toss a coin it will either come up heads (in which case the probability is 1), or it will come up tails (in which case the probability is 0). So I could maintain that there is no probability of ½ in this case, but a probability of either 0 or 1. Yet we do know that the idea of a probability of ½ is useful in the case of the coin, and that it describes something which is relevant to the problem. While it is true that on any particular toss the coin will fall either heads or tails, it is also true that if I toss it over and over again, many, many times, I will be right approximately half the time if I predict each time that it will come up heads.

In the case of the students' heights, if I say that the mean of the universe lies in the interval from 175.127 to 175.543 cm., I shall be either right or wrong. But if, on many statistical problems, I draw similar inferences, all based on this same sort of reasoning, I shall be right approximately 68.27 per cent of the time. We notice, then, that the probability about which we are talking is not the probability that the arithmetic mean of the universe has some given size in a particular problem, but the probability that our statements about statistical results are correct. Actually, this is the same thing that we are doing if we say that the chances are 0.5 that I will get a head on the toss of a coin. If I make many such statements I shall be right about half the time.

This sort of probability, which refers really to the likelihood that statements about statistical results are correct, is called *fiducial probability*, and where we have said throughout this chapter that the chances are 0.6827 that the mean of the universe lies within the interval from 175.127 to 175.543 cm., some statis-

ticians would prefer to say that the fiducial probability of the *confidence interval* 175.127 to 175.543 cm. is 0.6827.

We have used three different confidence intervals, one covering two standard errors, one covering four standard errors, and the other covering six standard errors. The wider the confidence interval, the larger the proportion of our statements which will be correct. If we make many such statements as that the arithmetic mean height in the universe lies between 175.127 and 175.543 cm., we shall find that about 68.27 per cent of our statements are correct. If we make many such statements as that the arithmetic mean height in the universe lies between 174.919 and 175.751 cm., we shall find that about 95.45 per cent of our statements are correct. A scientist who draws correct inferences 95 per cent of the time, and is led off on a false scent but 5 per cent of the time, is doing pretty well. And if we draw many such conclusions as that the arithmetic mean height in the universe lies between 174.711 and 175.959 cm., we shall find that about 99.73 per cent of our statements are correct; we would be led astray only about three times in a thousand, which is pretty good for a statistician. The wider our confidence interval, the more confidence we have in our results. On any particular single problem our statement that the mean lies within a given range is either right or wrong. But if we draw many, many statistical inferences by the methods here described, we can know in advance about how often we shall be right and how often wrong.

9.8. The Analysis of Variance.—The method of testing the significance of differences which we have just described is one of the most useful of elementary statistics. Every day the imaginative student will run across cases where it should be applied. Yet the method has the obvious handicap that it can be applied only where we are comparing two phenomena. If we compare two breeds of cattle, we can find if they differ significantly in milk production, but the method will not help us if we wish to compare three or four or more breeds. Two different feeding practices, two different spraying programs, or two different methods of teaching shorthand can be compared, and the importance of any difference between them evaluated. But life is not made up exclusively of dichotomies. There are probably more cases in scientific work where one wishes to compare three or four or more sets of data than there are where the comparison is limited to two.

Fortunately a rather simple extension of the ideas that we have just treated makes it possible to handle these more complex cases. In the work of Sec. 9.6, we computed two arithmetic means (or any other two comparable

statistics) and compared them, noting the amount of the difference between them, and finding whether the difference was or was not too great to have arisen by pure chance from data that were fundamentally similar. Suppose, now, that we have computed the average milk production of four breeds of cows—Jerseys, Guernseys, Holsteins, and Ayrshires. We get four different averages. Instead of talking about the difference between two of the breeds, we now ask ourselves whether or not the four averages vary more than could be expected on the basis of chance alone. We know that if we had chosen four different groups of Jersey cows and computed the average milk production for each group, there would have been some variation in the averages. When we find the average productions of the four different breeds, do the averages vary more than they ordinarily would if we had no breed differences to contend with?

This is evidently a question of dispersion—of variability—such as we discussed in Chap. VI. Here, however, it is the dispersion of a set of averages in which we are interested. We found in Sec. 6.10 that if we know the dispersion of several small groups, we can tell what dispersion there will be in the large group of which they are component parts. In our present problem, if we know the variation in production among the Jerseys, among the Guernseys, among the Holsteins, and among the Ayrshires, we can tell what variation there will be among the whole group thrown together. Or, to look at it another way, we can break down the variability of the whole group into its component parts—the variation among cows of the same breed, and the variation among the breeds themselves. We can then discover whether or not there is more variation among the breeds than we would get by chance.

This sort of investigation is called *analysis of variance*,[1] and the development of the method is one of the most important advances in recent statistical methodology. The subject matter is properly a part of advanced statistics rather than of an elementary course. Consequently the detail of computation and interpretation will not be covered here. Yet even the elementary student can see the kind of problem to which the method applies, and he should realize that the method is there for use if needed. If we were to apply the methods of analysis of variance to our hypothetical milk-production example, we should arrive at a solution something like that which we have found in the preceding section when we were dealing with differences between two measures—for example, we might discover either that there was more variation among the averages of the four breeds than would be likely to arise by chance; or that there was only as much variation between the means of the four breeds as would be found one time in six if we had tested different cattle of the same breed, so that there was not any real reason to conclude that the breeds were significantly different in production; or that we had not studied enough cases to be quite sure whether the variation among the breeds was significant.

9.9. Suggestions for Further Reading.—A number of interesting examples of the application of measures of reliability are found in William V. Lovitt

[1] The student will recall that the variance of a distribution is a measure of dispersion found by squaring the standard deviation (see sec. 6.10).

and Henry F. Hotzclaw, "Statistics," Chap. XV, Prentice-Hall, Inc., New York, 1929. A good elementary discussion of the concepts involved in measuring reliability is found in Frederick C. Mills, "Statistical Methods Applied to Economics and Business," Chaps. XIV and XVIII, Henry Holt and Company, Inc., New York, 1938. A very able but more advanced discussion of this problem appears in Burton H. Camp, "The Mathematical Part of Elementary Statistics," D.C. Heath and Company, Boston, 1931. For the best description of the new methods of treating these problems, the student is referred to R. A. Fisher, "Statistical Methods for Research Workers," 3d ed., Oliver & Boyd, London, 1930, especially Chaps. IV and V. The student who wishes to learn something of analysis of variance will find its most authoritative treatment in this same book by Fisher, but far simpler and more understandable treatments may be found in the book by Mills cited earlier in this paragraph, at Chap. XV; in George R. Davies and Dale Yoder, "Business Statistics," Chap. XIX, John Wiley & Sons, Inc., New York, 1941; or in George W. Snedecor, "Statistical Methods Applied to Experiments in Agriculture and Biology," Chaps. 10 and 11, Collegiate Press, Inc., of Iowa State College, Ames, Iowa, 1937. The same author and publishers issue a helpful manual entitled "Calculation and Interpretation of Analysis of Variance and Covariance," published in 1934. One of the simplest, most lucid discussions of the general problem of reliability will be found in C. H. Richardson, "An Introduction to Statistical Analysis," Chap. 11, Harcourt, Brace and Company, New York, 1935. A very fine advanced treatment is in John H. Smith, "Tests of Significance: What They Mean and How to Use Them," University of Chicago Press, Chicago, 1939.

EXERCISES

1. List a few cases in which it would be possible for the statistician to study all the cases in the universe, so that he would not have to estimate the characteristics of the universe from a sample.

2. From Exercise 2, page 154, compute the standard errors of the means and of the standard deviations, and likewise of the two coefficients of variation.

3. Compute the standard error of the coefficient of variation in the heights of Smith College girls from Exercise 3, page 154.

4. Compute the standard error of the mean, of the standard deviation, and of the coefficient of variation of the mothers' ages given in Exercise 4, page 154.

5. The average number of offspring in 55 completed families is given in Exercise 5, page 154, as 3.55. The standard deviation is 1.79. Hence the standard error of the mean is 0.244. (Check this computation.) How many cases will it be necessary to take if we are to reduce the standard error of the mean to 0.1? Assume that the standard deviation in the number of offspring remains the same as we increase the number of families studied.

6. Exercise 6, page 154, gives figures to show that the 22,498 divorces in Wisconsin from 1887 to 1906 were preceded by an average married period of 10.37 years, with a standard deviation of 8.39 years. The average and the standard deviation for the year 1929 were 9.83 and 8.26 years,

respectively. Had there been a significant decrease in the length of marriages preceding divorces? Was the decrease in variability as shown by the smaller standard deviation significant, or might it arise from chance? If the latter, how likely is it that such a difference would arise by chance alone?

7. Suppose that a group of people are tested with respect to strength of grip in their right hands. The average turns out to be 40 kg., with a standard deviation of 1 kg. There are 40 people in the group. Is it reasonable to assume that this group of 40 people is drawn from the same universe as the people for whom figures are given in Exercise 7, page 155? (There were 12 men in this latter group.)

8. In Exercise 3, page 124, we computed the mean and median hourly wage and the quartiles. In Exercise 1, page 154, we computed the standard deviation of these figures. From these figures already computed find the standard error of the mean, of the standard deviation, of the median, of the quartiles, and of the semi-interquartile range.

9. In Exercise 2, page 230, we computed the value of α_3 for a given distribution. Compute the standard error of α_3. Was there significant skewness in the distribution? (Did α_3 differ from zero by an amount equal to over three times the standard error of α_3?)

10. In Exercise 13, page 186, we found that 53 of the 625 diphtheria cases studied turned out fatally. The fatality rate is thus 8.5 per cent. What is the standard error of this percentage? If in a new epidemic there were 200 cases and 8 of them resulted in fatalities, would the difference in fatality rates be significant? If so, what would you conclude? Suppose there were 25 fatalities in this new epidemic. What would you then conclude? Give your reasoning.

11. Is there a significant difference between the heights of the Smith College students mentioned in Exercise 3, page 154, and the heights of the Harvard students mentioned in the illustrative example in the text (on pages 141*ff.*, for example)? Is either group significantly more variable in height than the other group?

12. In Exercise 8 above you have found several standard errors. Compute the probable errors of the same values. Write the values followed by their probable errors as they would commonly appear.

13. Hatai measured the lengths of the craniums of 53 male rats and says that the average length was 43.3 ± 0.17 mm.[1] Explain this combination of figures.

14. In Providence, Rhode Island, in 1915 there were 43 cases of diphtheria in children between the ages of one and two years. Of these, 13.95 per cent resulted in deaths (6 deaths out of 43 cases). In the same city in the same year there were 62 cases of diphtheria in people who were 20 years old or over. Of these, 3.23 per cent resulted in deaths (2 deaths out of 62 cases). Was there a significant difference between these percentages?[2]

15. In the "World Almanac" can be found the monthly mean temperatures in New York City over a considerable period of years. Compute

[1] Quoted in DONALDSON, *op. cit.*, p. 50.

[2] Based on figures in Whipple, "Vital Statistics," John Wiley & Sons, Inc., New York, 1923, p. 377.

for January and for July the average and the standard deviation of temperatures. Is there a significant difference between January and July temperatures? Is there a significant difference between the variability of temperatures in January and July? Is there a significant difference between the coefficients of variation for the two months?

16. On page 256 are given certain figures for the weights of male and female rats. Is there a significant difference in variability of weights between the sexes?

17. Are female rats significantly different from Hollins College girls in variability of weight? Compute the standard error of the difference between the two coefficients of variation, getting the basic figures from pages 252 and 256.

18. A study of milk consumption in metropolitan Boston in December, 1930, showed that the average per-capita consumption of milk was 0.391 qt. $\pm$ 0.00262 qt.[1] Explain the meaning of these figures when taken in combination. What was the standard error of per-capita milk consumption?

19. In the period 1925–1927 the average operator's income on 105 Connecticut tobacco farms growing Havana seed tobacco was \$905. The average operator's income on 97 Connecticut tobacco farms growing broadleaf tobacco was −\$450 (that is, a loss of \$450). The standard deviations of the operators' incomes were \$1409 on the farms raising Havana Seed tobacco and \$2305 on the farms raising broadleaf tobacco.[2] Was there a significant difference between the operators' incomes on these two groups of farms?

[1] Based on figures in F. V. Waugh, Consumption of Milk and Dairy Products in Metropolitan Boston in December, 1930, *New England Council on Marketing and Food Supply*, September, 1931, pp. 4 and 11.

[2] Based on data on C. I. Hendrickson, An Economic Study of the Agriculture of the Connecticut Valleys, *Storrs Agricultural Experiment Station Bulletin* 165, pp. 123 and 142.

CHAPTER X

HISTORICAL DATA—SECULAR TREND

Up to this point we have been describing methods of dealing with data which exist at a point of time. We have not been describing the rate of growth of Harvard students, but have been depicting conditions as they existed without regard to passage of time. In fact, we have not considered changes in data at all. Yet the statistician is often deeply interested in time changes. The biologist studies rates of growth both of individuals and of populations, the psychologist studies rates of learning, the economist interests himself in the sequence of price changes, etc. We shall now develop the simpler methods for dealing with data which are spread over time.

10.1. The Use of Two Variables.—As soon as we do this we are faced with the fact that we are treating two variables at once rather than one. If we study the history of milk prices, our two variables are milk price and time. If we collect data showing the size of the population of the United States at each census, our two variables are size of population and time. In our previous examples but one factor has been changing. If we take the case which has been used so often for purposes of illustration in the earlier chapters of this book, the only variable was the height of the Harvard students. But if our data had been segregated by years, so that we could discover the change in the character of the heights from year to year, then both height and time would have been varying.

When we have two or more variables in a problem, it becomes necessary to set up some system for distinguishing them. When but one thing varies, we can talk about the standard deviation or the mean and everyone knows that we mean the standard deviation or the mean of the only thing that varies. But if two things vary, we must state which is referred to.

We have been referring to the variable in each of our problems (since there was but one variable) as X. In our formulas ΣX has meant, "Add the values of the thing which varies."

Obviously if two or more things vary, this direction would not be sufficient. Hence the statistician adopts one or the other of two conventional modes of expression. Suppose his problem is one of studying the change through time of wheat acreage and wheat prices. There are three variables: (1) wheat acreage, (2) wheat price, (3) time. He may distinguish them by assigning different letters to the different variables, thus:

Let X represent time.
Let Y represent wheat acreage.
Let Z represent wheat price.

Or he may distinguish them by using numerical subscripts, thus:

Let X_1 represent wheat price.
Let X_2 represent wheat acreage.
Let X_3 represent time.

If he follows the first of these plans, he will refer to the averages, standard deviations, medians, quartiles, etc., as follows:

$\bar{X}$ = average of the X's; that is, average time
$\bar{Y}$ = average of the acreages
$\bar{Z}$ = average price
y = deviation of any acreage from the average acreage
σ_z = standard deviation of the prices
Med._y = median acreage
Q_{1z} = first quartile of the prices
Σx^2 = sum of the squares of the deviations of the times from the average time

etc.

If he follows the second plan, he will distinguish the variables by using numerical subscripts:

$\bar{X}_1$ = average price
$\bar{X}_2$ = average acreage
x_1 = deviation of a price from the average price
σ_2 = standard deviation of the acreages
Med._3 = median time (the median year, the median week, or the median minute, depending on the periods into which time is divided in the problem)
Σx_2^2 = sum of the squares of the deviations of the acreages from the average acreage.

etc.

10.2. Calendar Variation.—In his dealings with time series it is important for the student to remember that data given on a

monthly basis are seldom comparable from month to month. We can illustrate with the figures in Table 10.1, which show the number of deaths from tuberculosis in the United States registration area in 1914, distributed by months.[1] In studying this

TABLE 10.1.—DEATHS FROM TUBERCULOSIS BY MONTHS, UNITED STATES REGISTRATION AREA, 1914

Month	Number of Deaths
January	7522
February	7524
March	8537
April	8238
May	7782
June	6901
July	6528
August	6209
September	6031
October	6009
November	6212
December	6873

table one may be misled unless one remembers that the months differ in length. It will be noted that the number of deaths in February is almost identical with that in January, but February is (except in leap years) only a little over 90 per cent of the length of January. To make this table strictly comparable from month to month it would be necessary to reduce the figures in the table to deaths per day by dividing the number of deaths in each month by the number of days in that month. This computation would give us Table 10.2, which shows for the same area and the same period the deaths per day from tuberculosis by months.

In economic problems the situation is often far more complicated than the one just illustrated. It is necessary to adjust not only for differences in length of months but differences in the number of business days in the different months and in the same month in different years. In one year January may have five Sundays, and in the next but four. In one year Easter falls

[1] From WHIPPLE, "Vital Statistics," p. 368, John Wiley & Sons, Inc., New York, 1923.

in March, and in another year it falls in April. If our data are given by weeks, we have the additional difficulty that the months are not made up of a whole number of weeks (save in the case of February), and that some holidays occur in some years during one week of the month and in other years during other weeks. These calendar difficulties make for considerable confusion in statistical work, and when neglected may lead to foolish conclusions. Presumably the situation could be improved somewhat

TABLE 10.2.—DEATHS PER DAY FROM TUBERCULOSIS, UNITED STATES REGISTRATION AREA, 1914, BY MONTHS

Month	Deaths per Day
January	242
February	269
March	275
April	274
May	251
June	230
July	210
August	200
September	200
October	194
November	207
December	222

by some kind of reform of the calendar. Until and unless such reform takes place, the statistician must be on his guard whenever he deals with historical data. At present, even one year may vary from the next in length by one day, or roughly $\frac{3}{10}$ of 1 per cent.[1]

10.3. Types of Movements in Historical Data.[2]—Changes of several kinds take place in historical data. The types of change can best be differentiated by illustrations. Let us start with

[1] D. J. Cowden has published a very useful "Flexible Calendar of Working Days" showing the number of calendar days, Sundays, Saturdays, and holidays by months from 1900 to 1940. See F. E. Croxton and D. J. Cowden, "Practical Business Statistics," p. 515, Prentice-Hall, Inc., New York, 1934.

[2] For an unusually good discussion of the kinds of movements in historical data see Edmund E. Day, "Statistical Analysis," Chap. XV, The Macmillan Company, New York, 1925.

figures showing the production of crude petroleum in the United States. Table 10.3 gives annual figures on crude-petroleum output in the United States from 1906 to 1929.

TABLE 10.3.—PRODUCTION OF CRUDE PETROLEUM IN UNITED STATES, 1906–1929[1]

Year	Output (millions of barrels)
1906	126
1907	166
1908	179
1909	183
1910	210
1911	220
1912	223
1913	248
1914	266
1915	281
1916	301
1917	335
1918	356
1919	378
1920	443
1921	472
1922	558
1923	732
1924	714
1925	764
1926	771
1927	901
1928	901
1929	1007

[1] Figures taken from *Statistical Abstract*, Table 706, p. 682, 1933.

Even a cursory inspection of these data makes it evident that there has been a tendency for the output of crude petroleum to increase from year to year. Sometimes the increase in output has been greater than at other times, but there is a very noticeable general tendency throughout the period for output to rise. This becomes even more evident when the figures are presented in graphical form (see Fig. 10.1).

Whenever there is a long-time tendency for data to increase or decrease, we say that there is a *secular trend* or a *secular*

movement in the data. It is not necessary that the rise or fall continue each and every year throughout the period. If we have a quarter of a century during which prices tend generally to fall, we should say that there was a secular decline in prices during the period even though there might be an occasional isolated year in which there was a small rise in price. Just so long as we can truthfully say that the period was one which was generally characterized by an upward movement or by a

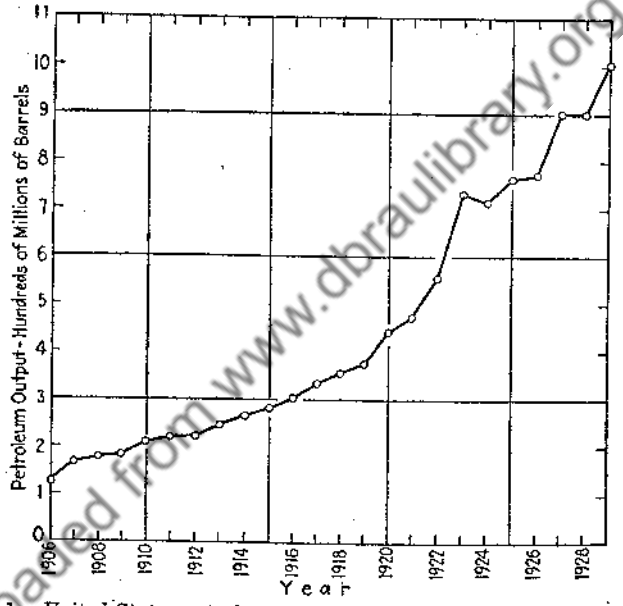


FIG. 10.1.—United States petroleum production, 1906–1929. (*Data taken from Table* 10.3.)

downward movement, we say that there was a secular movement present.

In contrast with the secular movement just pictured let us note the data of Table 10.4, showing egg prices in New York City by months for a period of five years. The eggs were of a grade known as "near-by-hennery whites," and the prices are average monthly prices.[1] The figures in the body of the table are prices in cents per dozen taken to the nearest cent.

[1] Data quoted from *New York Price Current* by I. G. Davis in Connecticut's Poultry Industry, *Connecticut Agricultural College Extension Bulletin* 79, 1924, pp. 35–36.

If one is to follow the historical changes in these data, it will be necessary to look down each column from the top to the bottom, returning then to the top of the next column and repeating. If this is done, one sees immediately that the movement in the data is neither a regular rise nor a regular fall, but rather an alternation of short rises and short falls. This becomes very

TABLE 10.4.—PRICES OF NEAR-BY-HENNERY WHITE EGGS, NEW YORK CITY, BY MONTHS, 1919–1923

Month	1919	1920	1921	1922	1923
January	72	84	76	56	57
February	56	70	52	49	47
March	48	59	43	38	44
April	52	54	38	38	39
May	53	53	33	38	40
June	56	56	39	43	41
July	63	65	50	45	45
August	68	71	57	55	53
September	75	82	71	66	62
October	88	100	86	82	77
November	98	102	95	89	83
December	82	95	78	70	64

evident when the data are presented graphically (Fig. 10.2). The movement turns out to be wavy. When we have data in which there are regular "ups and downs," as there are here, giving this wavy appearance to a graph of the data, we say that there is a *cyclical movement* in the data, or that there are *cycles* in the data. Usually if these cycles tend to be just a year long, varying regularly with the seasons of the year (as is true in this case), we say that there is a *seasonal movement* in the data. It should be evident that a seasonal movement is one type of cyclical movement—one in which the cycle is 12 months long.

It would be quite possible to find data in which cyclical and secular movements were combined, both types of movement appearing coincidentally in the same data. A hypothetical case of this kind is represented by Fig. 10.3, which purports to show the monthly sales of corporation *X* for a five-year period. Inspection of the diagram will make it evident at once that there are regular seasonal swings in sales. Obviously this product moves on to the market largely in the summer; the winters are

slack. Also it is obvious that the output of this corporation is becoming larger from year to year; that is, there is a secular

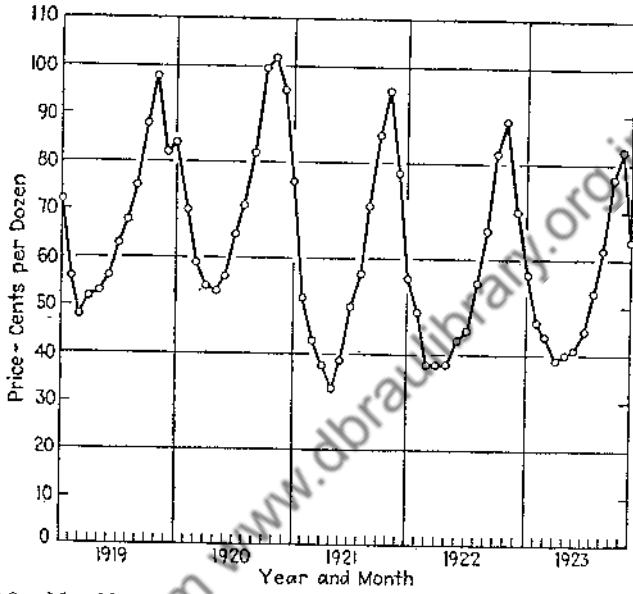


FIG. 10.2.—Monthly prices of "near-by-hennery white" eggs on the New York City market, 1919–1923. (*Data from Table* 10.4.)

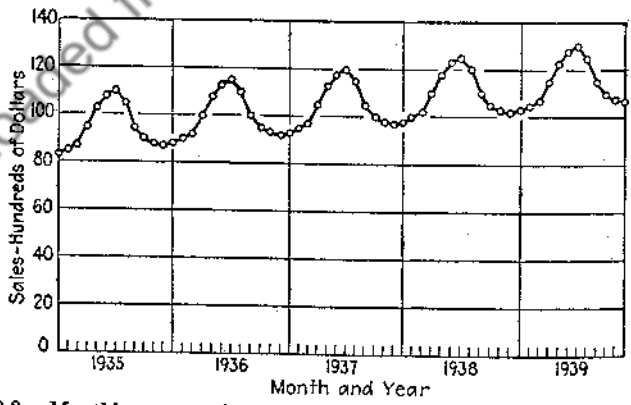


FIG. 10.3.—Monthly gross sales of corporation *X*, 1935–1939. An idealized combination of seasonal and secular movements.

trend underlying the seasonal movements, and the seasonal movements are fluctuating around this secular trend.

In addition to these two types of change in which the historical movements seem to follow some plan, there are movements in historical data that seem entirely erratic—planless. They are of a type which could not be said to have any regularity at all. Such movements in data are called *random movements* or *erratic movements* or *residual movements*. If we take another hypothetical case, merely changing the previous case a little, we can show a combination in which secular, seasonal, and random movements all appear in the same data. It is not uncommon for this to

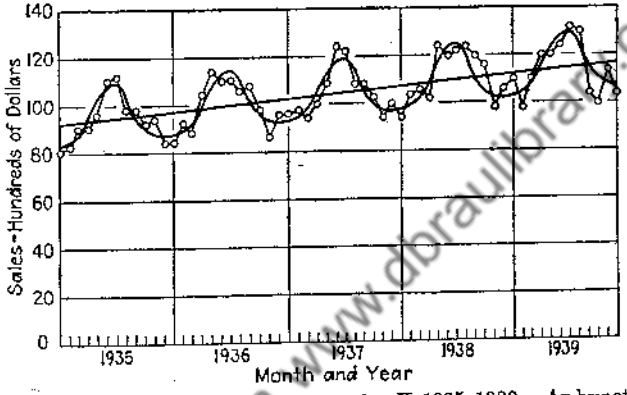


FIG. 10.4.—Monthly gross sales of corporation *X*, 1935–1939. An hypothetical case combining secular trend (the straight line), seasonal movement (the regular wavy line), and random movements (shown by the departure of the actual data, shown in circles, from the smooth curve).

occur, although it seldom occurs with such evident regularity as in the idealized case given in Fig. 10.4.

A statistician who is studying corporation *X* may be interested in the secular movement. This long-run increase in output may be the thing he is trying to explain. On the other hand, he may be interested in the seasonal changes, and in order to study them he may wish to eliminate the secular movement entirely so that he can better understand the seasonal movements. Often he will be interested in studying the random movements in order to discover why they occur. In this case he may wish to eliminate the secular and seasonal (or other cyclical) movements so that the pure random movements will remain for study. The student should remember, then, that the statistician may wish to describe any regular type of historical movement either because he is interested in the movement itself or because he

wants to get rid of it so that it will not obscure movements of other kinds. We shall explain in this chapter the simpler ways of dealing with secular changes.

10.4. The Secular Trend.—We have discovered that any movement constitutes a secular trend if it continues in general in the same direction for a considerable period of time. If we are counting the bacterial population of a culture every 5 min., and the population continues to increase (or decrease) fairly regularly for several days, we should say that this was a secular change. It lasted but a few days, but these comprised a great many 5-min. periods. If, on the other hand, we were studying changes in infant mortality in the United States and if we had figures by years, the mere fact that there were decreases in infant mortality for two or three successive years would not indicate the existence of a secular trend. In the one case a trend which lasts for a few days is called a secular trend, while in the other case a trend which lasts two or three years is not so called. But the reason should be obvious. There is no specific time during which a movement must continue if it is to be classed as secular. The word is comparative, not absolute. Just as 5 hr. may seem short to one who is about to be electrocuted but very long to one who is seated in a dentist's chair, so a given length of time may be secular under some conditions and not under others. When one says that a given movement was secular, he means that it lasted for a period that one would call a long time—long for such data to continue to change uniformly.

There are many ways of describing and dealing with secular trends, and sometimes the simplest of these ways is the best. We shall start with the simplest way and proceed to some of the others.

10.5. Freehand Trend.—The simplest and most informal method of describing the secular trend in data is to plot them and draw on the graph a line that seems to the eye to follow the general long-run movement of the data without following the minor short-run fluctuations. For example, if we plot the data showing petroleum output in the United States from 1917 to 1929 which are tabulated on page 271, we get a graph such as Fig. 10.5. We now draw on this graph a line, straight or curved as the data may determine, which follows the general direction of the data. In this diagram the actual production figures are

represented by the solid line and the freehand trend by the broken line.

If the secular change appears to be approximately along a straight line, the freehand trend may be drawn along a ruler, and then it is common to determine the average of the values which are varying through time and to locate this average on the graph at the central period of time. In this case, for example, the average petroleum output for the period was 641 million barrels.

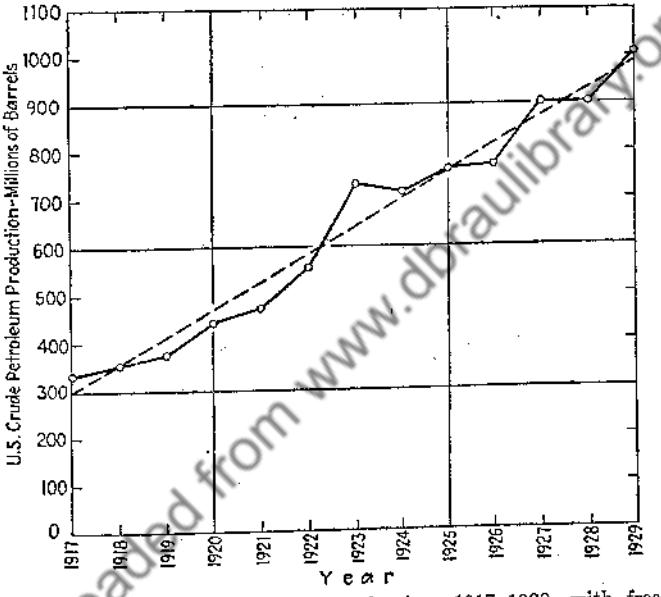


FIG. 10.5.—United States petroleum production, 1917–1929, with freehand straight-line trend.

There are 13 years in the period covered, so the mean of the years (or the median of the years, since this will coincide with the mean in such a distribution) is the seventh year, or 1923. We could locate on the graph opposite 1923 a point representing 641 million barrels and draw our freehand trend through this point. This is but a rough guide, however; the advantage of the freehand method is that it allows the statistician to do whatever looks right without worrying about rules. The method has the obvious disadvantage that two equally competent workers may draw quite different trend lines, and the same worker may well draw different trend lines through the same data at different

times. But the method is simple, quick, easy, and fairly satisfactory for rough work. If accuracy is needed it may be worth while to attempt some of the other methods. Many a statistician uses the freehand method to get a first approximation of the trend before using the other methods, since it takes little time and gives him some idea of the nature of the trend.

10.6. Method of Selected Points.—This method is merely an addition to the freehand method, and can well be used if one decides to use the freehand method. Having drawn the freehand trend as in the preceding section, we select two points on the trend line, one near each extreme. Let X represent the year and let Y represent the petroleum output.[1] Suppose we select the two years 1918 and 1928; that is, first we shall let $X = 1918$ and next we shall let $X = 1928$. We determine in each case the value of Y by reading the position of the freehand trend at the selected year. Thus in 1918 the trend value seems to be 350, and in 1928 it appears to be 930. We have, then, two points which can be described as follows:

First point:

$$X = 1918; \qquad Y = 350$$

Second point:

$$X = 1928; \qquad Y = 930$$

If we are to fit a straight line to these data, as we did before, we must use an equation of the general type $Y = a + bX$. This is the equation of the straight line. Every straight line can be described by an equation of this type, and every equation of this type describes some straight line.[2] It is necessary for us to find the values of a and b in order that we may know exactly which straight line we are dealing with here.

We have two observations of concurrent values of X and Y. In the first case $X = 1918$ and $Y = 350$. Let us substitute

[1] In dealing with historical data statisticians always have the letter X represent the time variable and the letter Y represent the other variable. There is some further discussion of this at a later point in this book in connection with the description of regression lines (see p. 384).

[2] Lines and their equations are discussed in greater detail on pp. 439*ff*. The student who is interested may read these pages now. Others may take the form of the straight-line equation on faith for the present. Any text on elementary algebra covers the subject under the heading "linear equations."

these values for X and Y in our type equation $Y = a + bX$. This computation gives our first *observation equation:*

$$350 = a + 1918b$$

In the case of the other point, the values were $X = 1928$ and $Y = 930$. Substituting these, we get our second observation equation:

$$930 = a + 1928b$$

This gives us our two observation equations, and if we solve them simultaneously we find the values of a and b which we need to describe our line. Solving the two equations, we find that $a = -110{,}894$ and $b = +58$. Our equation then becomes (by substituting these values of a and b in the equation $Y = a + bX$)

$$Y = -110{,}894 + 58X$$

If we recall that X = the year and Y = the petroleum production, we can estimate the production for any year from the equation just given. For example, what was the petroleum production in 1920? Substitute 1920 for X in the equation and you have

$$Y = -110{,}894 + 58(1920) = 466$$

We estimate that the 1920 production was 466. Since our original production figures were in millions of barrels, this means 466 million barrels. Reference to the original data (page 271) will show that the 1920 production was actually 443 million barrels. Inspection of Fig. 10.5, page 277, will reveal that the trend line gave a value for 1920 of 466 million barrels, however. Experiment will show that any production estimated by this equation will give a point on the trend line. It is for this reason that we say that this equation is a description of the line.[1]

[1] Although this equation is extremely helpful in estimating the value of Y within the time period to which it was fitted, the student must be cautioned against a too free use of the equation in estimating the value of Y beyond that range. Here we computed the line for data which ran from 1917 to 1929. Within this period the errors of estimation will be reasonably small. But if the student wishes to see the danger of estimating for times beyond the limits of this period, let him use this equation to estimate petroleum production in 1906 and compare his result with the actual figure for 1906 on page 271.

This equation is very useful because it gives a concise, exact description of the trend. Before we computed the equation we could have described the trend to others only by drawing it on a graph and sending it to them. Even in that case there would have been the difficulty that the scale is too small on most graphs to permit accurate reading. Now that we have the equation,

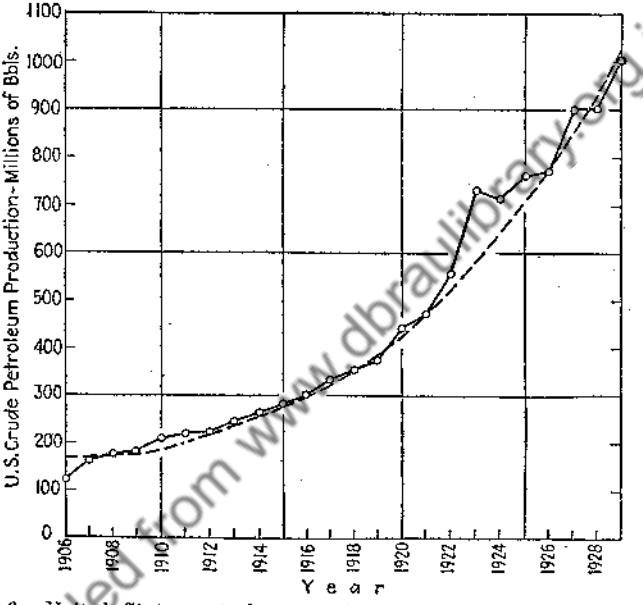


FIG. 10.6.—United States petroleum production, 1906–1929, with freehand parabolic trend.

however, we can tell anyone that the trend of petroleum production from 1917 to 1929 was

$$Y = -110{,}894 + 58X$$

This tells all one needs to know. In fact, it tells all that we know ourselves.

10.7. Curvilinear Trends by Selected Points.—Of course, the trend may be a curve rather than a straight line. If, for example, we plot the entire data on petroleum production from the table on page 271, we get Fig. 10.6. First we draw a freehand curve through the data showing the trend; this is shown by the broken line on the chart. Next we select three points on this curve, one near each extreme and one near the center.

Suppose that we select the positions of the curve for the years 1908, 1920, and 1928; that is, the values of X will be 1908, 1920, and 1928. This time we shall, however, follow a plan that reduces the amount of arithmetic considerably. We shall shift the *origin* of our time series. Time is usually reckoned from the birth of Christ, but this custom gives us numbers in the thousands which are hard to handle, especially when it becomes necessary to square them or to multiply them by large numbers. Of course, one could figure time from any other convenient point. Thus the year which we commonly call 1945 is the year 7453–7454 of the Byzantine era, the year 5705–5706 according to the Jewish calendar, the year 2698 since the founding of Rome, etc. The more recent the starting point which we select, the smaller are the numbers with which we shall have to deal. Hence it is common in statistical problems to take a basis for reckoning at some very recent date, and it is most common to take the center of the time series being studied and call it the year 0.

Suppose that, in the present case, we decided to reckon time from what is usually known as the year 1920. We say that we shift the origin to 1920. This year, then, becomes the year 0. The year 1921 becomes the year 1, the year 1922 the year 2, the year 1919 the year -1, the year 1915 the year -5, etc. Under our new plan, therefore, the years we have selected (1908, 1920, and 1928, as above) become the years -12, 0, and $+8$. We note that the trend at these years has values of 170, 425, and 930. Our three pairs of values are, then,

$$\begin{aligned} X &= -12; \quad Y = 170 \\ X &= 0; \quad Y = 425 \\ X &= 8; \quad Y = 930 \end{aligned}$$

To fit a curvilinear trend of this type, we need an equation of the form[1]

$$Y = a + bX + cX^2$$

If we substitute the values of X and Y given above, we get the following three observation equations:

[1] This is the equation of a second-degree parabola. Again the student whose memory of mathematical curves is rusty will have to take this equation on faith or else read now pp. 440*ff*.

$$170 = a - 12b + (-12^2)c$$
$$425 = a + 0 + 0$$
$$930 = a + 8b + 8^2c$$

It is at once evident that the value of a is 425. (It also becomes evident immediately that it saves computation to take as the origin the year of our central observation!) Solving, we find that the values of b and c are

$$b = 46.33$$
$$c = 2.09$$

Substituting these values of a, b, and c in the type equation ($Y = a + bX + cX^2$), we get the equation of this particular curve, which is

$$Y = 425 + 46.33X + 2.09X^2$$
Origin at 1920

When such an equation is given, it is important that the origin be stated with the equation. Otherwise the results are meaningless.

From our type equation let us estimate the trend value for 1910. When 1920 is the origin, the year 1910 becomes the year -10; that is, X becomes -10 when we wish to estimate the 1910 production. Substitute -10 for X in the type equation and it becomes

$$Y = 425 + 46.33(-10) + 2.09(-10^2) = 170.7$$

Thus our estimate of petroleum production in 1910 is 170,700,000 barrels (since our production figures are in millions of barrels).

This equation now describes our curvilinear trend, and we can easily and accurately tell others our conclusions.

It must be remembered, however, that two investigators fitting trends by the method of selected points may well obtain somewhat different results. The freehand trends which they draw originally to guide them may well differ somewhat, and the points they select from which to get values for their observation equations may differ. Thus this combination of methods (freehand trend plus selected points) has the disadvantage that two equally competent workers may differ in their conclusions. The methods are, however, quick and easy to apply; and every statistician finds them useful at times. One might by the same methods, of course, fit more complicated curves. When we fitted a straight

line (one with no bends), we had to find values of a type equation with two unknowns, a and b. When we fitted a second-degree parabola with its one curve, we had to find three unknowns, a, b, and c. For each bend in the curve we should have to add one unknown, and the added accuracy of results (if any) is seldom sufficient to offset the added arithmetic. Surely when the method employed is that of freehand trends and selected points, it will seldom pay an investigator to fit a curve more complicated than the third-degree parabola, the type equation of which is

$$Y = a + bX + cX^2 + dX^3$$

Here there are four unknowns (a, b, c, and d) to find, and it is necessary to select four points on the freehand curve and solve four observation equations. This curve will have two bends.

10.8. Moving Average.—We may also find the trend value of data by the method known as the *method of moving averages*. This method is based on the assumption that minor variations in the Y variable are to be considered as unusual, and that they can be removed by the process of averaging. Suppose I wish to know what petroleum production was "normal" for 1920. Would it not be fair to tell me the average production for 1920 and the two or three years before and after? This is exactly what constitutes the process of computing the moving average.

Let us go back to our figures of petroleum production. The original dates and production figures are repeated as the first two columns of Table 10.5. In the third column of this table we have the moving average itself. Opposite each year is the average production of that year and of the two years preceding and the two years following; that is, each figure in the last column is an average of five years' production centered at the given year. The moving average of production for 1917 is 330.2. This is the average production for 1915, 1916, 1917, 1918, and 1919. Similarly with each other year. In this case we have a five-year moving average. We might, of course, use the average of three years or seven years or some other number of years. Obviously if our average includes the same number of years before as after the given year, the total number of years in the period will be odd. Thus here we include the year itself, two years before, and two years after, or five years altogether. Usually we use an odd

number of periods for our moving average and center the average at the middle year, as in the example given here.

The moving average is most commonly applied to data which are characterized by cyclical movements. It is employed to

TABLE 10.5.—COMPUTATION OF MOVING AVERAGE OF PETROLEUM PRODUCTION

Year	Output (millions of barrels)	Moving-Average Output
1906	126	?
1907	166	?
1908	179	172.8
1909	183	191.6
1910	210	203.0
1911	220	216.8
1912	223	233.4
1913	248	247.6
1914	266	263.8
1915	281	286.2
1916	301	307.8
1917	335	330.2
1918	356	362.6
1919	378	396.8
1920	443	441.4
1921	472	516.6
1922	558	583.8
1923	732	648.0
1924	714	707.8
1925	764	776.4
1926	771	810.2
1927	901	868.8
1928	901	?
1929	1007	?

eliminate the cycles and leave the general trend of the data. We literally "average out" the seasonal or other cyclical variations. In such a case it is necessary to select a period for the moving average which coincides with the length of the cycle; otherwise the cycle will not be entirely removed. When the period of the moving average and the period of the cycle in the data differ, the moving average will display a cycle which has

the same period as the cycle in the data, but which has less amplitude than the cycle in the data. Often the statistician finds that the cycles in the data are not of uniform length. In such a case he usually takes a moving-average period equal to or somewhat greater than the average period of the cycle in the data.

TABLE 10.6.—FOUR-YEAR MOVING AVERAGE OF PETROLEUM PRODUCTION

Year	Output (millions of barrels)	Four-year Moving Average	Moving Average Centered
1906	126		?
1907	166		?
1908	179	163.5	174.0
1909	183	184.5	191.2
1910	210	198.0	203.5
1911	220	209.0	217.1
1912	223	225.25	232.2
1913	248	239.25	246.9
1914	266	254.5	264.2
1915	281	274.0	284.9
1916	301	295.75	307.0
1917	335	318.25	430.0
1918	356	342.5	360.2
1919	378	378.0	395.2
1920	443	412.25	437.5
1921	472	462.75	507.0
1922	558	551.25	585.1
1923	732	619.0	655.5
1924	714	692.0	718.6
1925	764	745.25	766.4
1926	771	787.5	810.9
1927	901	834.25	864.7
1928	901	895.00	?
1929	1007		?

We have noted that one usually selects an odd number of periods for his moving average, since the process of centering is then simplified. But if data have a marked cycle which extends over an even number of periods (as, for example, a 12-month cycle, which is very common) it is necessary to take an even number of periods for the moving average. If we take, for example, a four-year moving average of the data in Table

10.5, showing the results in Table 10.6, we discover that the moving averages appear between the years rather than at the years. The first figure in the third column, 163.5, is the average petroleum production for the years 1906, 1907, 1908, and 1909. The center of this series of years is halfway between 1907 and 1908, and we therefore enter our moving average halfway between them. But ultimately we want the value for each year, and not the value at points between. Therefore when our moving average is based on an even number of periods we add a fourth column, which is a centered two-period moving average of the third column.

The first figure in the last column of Table 10.6 is 174.0, the average of the first two figures in the third column. It is entered halfway between the first two figures of the first column, which sets it opposite the year 1908. The other figures in the last column are found similarly.

10.9. The Progressive Mean.—When we are told that the five-year moving average of petroleum production was 648 million barrels in 1923 (figures from Table 10.5), we understand that this is not necessarily the actual output (which was 732 million barrels), but it is a "normal" output for the five-year period centered at 1923. It is the simple arithmetic average of the outputs of the five-year period centered at 1923. It has been suggested by some statisticians that in a case of this kind the center year should be given more weight in computing the average than the other years, and the farther we go from the center of the period being averaged the less weight we should give. It would be a simple matter, of course, to compute a weighted moving average, always weighting the center year 10, the year each side of the center 6, and the second year from the center in either direction 1, or any other such set of weights, diminishing as we draw farther from the center of the period. When such weighting is used, it has been common to weight the years with the coefficients of the binomial expansion which has the requisite number of terms. If we turn to Table 7.1, page 163, we discover that when the binomial has five terms the coefficients are 1, 4, 6, 4, and 1. If we were to find the weighted arithmetic average of the first five years of Table 10.5, using these weights, we should perform the computations shown in Table 10.7.

TABLE 10.7.—COMPUTATION OF THE PROGRESSIVE MEAN

Year	Output (X)	Weight (W)	(XW)
1906	126	1	126
1907	166	4	664
1908	179	6	1074
1909	183	4	732
1910	210	1	210
Totals.....		16	2806

Using now our formula for the weighted arithmetic mean (see page 62), we have

$$\bar{X} = \frac{\Sigma(XW)}{\Sigma W} = \frac{2806}{16} = 175.4$$

This average would be set opposite the central year, 1908. Similarly we would compute a weighted arithmetic mean for each other set of five consecutive years, using each time these same weights, and always setting the weighted mean opposite the central year of the period. When the weights used in computing the weighted moving average are the coefficients of the expanded binomial, as in the example just given, the result is known as a *progressive mean;* that is, a progressive mean is a weighted moving average with the binomial coefficients as weights. It is obvious that the work entailed in computing a progressive mean is far greater than that required for the computation of the ordinary unweighted moving average, and as a result the latter is far more common in practice.

10.10. Moving Average with Curvilinear Trends.—The moving average gives a very good picture of the general, long-run movement in data if the data contain rather uniform cycles and if the trend in the data, if any, is linear or approximately so. But if there is a long-run secular curvilinear trend in the data, the ordinary moving average will contain a biassed error. If the trend line is concave upward (like the side of a bowl), the value of the moving average will always be too high; if the trend line is concave downward (like the side of a derby hat), the value of the moving average will always be too low. We can illustrate this with the data of Table 10.8. The moving average appears

in the third column, and the original data with the moving average appear in Fig. 10.7. In the figure, the original data are shown as points in small circles, while the values of the moving average are shown as small crosses. It will be seen at once

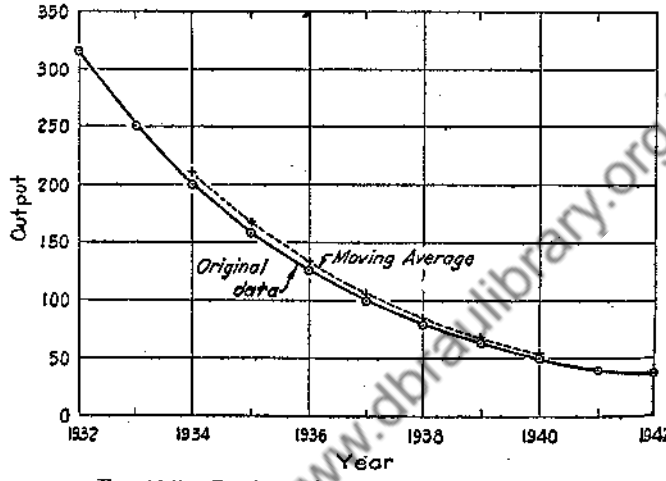


FIG. 10.7.—Persistent bias in the moving average.

that the moving average is consistently too high, since the original data fall along a curve which is concave upward.

TABLE 10.8.—MOVING AVERAGE OF DATA WHICH ARE CONCAVE UPWARD

Year	Output	Moving Average
1932	316.2	?
1933	251.2	?
1934	199.5	210.3
1935	158.5	167.0
1936	125.9	132.7
1937	100.0	105.4
1938	79.4	83.7
1939	63.1	66.5
1940	50.1	52.8
1941	39.8	?
1942	31.6	?

We might try to overcome this tendency toward error by weighting, giving the large weights to the small values and the

small weights to the large values if the curve is concave upward; but giving the large weights to large values and small weights to small values if the curve is concave downward. Such a procedure would be entirely arbitrary, however, and there would be no reason to believe that the results gave the true long-run trend. In some cases, however, our long-run trend seems to follow some particular law, and in such cases we can sometimes alter our moving average to correct for the error. When the

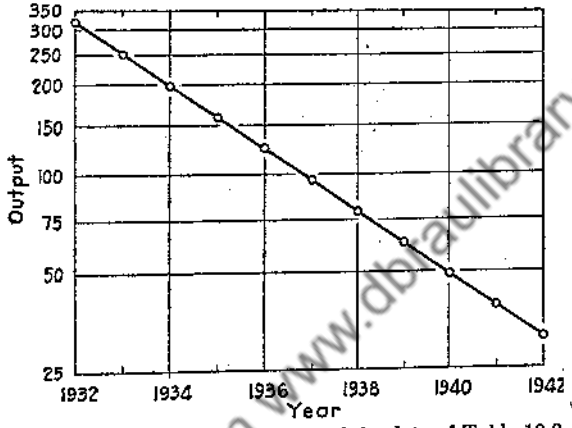


FIG. 10.8.—Semilogarithmic chart of the data of Table 10.8.

data of Table 10.8 are plotted on semilogarithmic paper, we discover that the curvilinear trend of Fig. 10.7 has become a straight line (see Fig. 10.8). This tells us that the trend in this particular example is a geometric trend, and we can correct for the error by using the moving geometric average rather than the moving arithmetic average.[1] The process of computation is shown in Table 10.9.

In Table 10.9 the first two columns are repeated from Table 10.8. The third column consists of the logarithms of the corresponding figures in the second columns. In the fourth column are figures for a five-year moving average of the logarithms of the third column. Each figure in the last column is the anti-

[1] The student is referred to Sec. 5.19, where this use of the geometric average is explained. Whenever one contemplates the use of a moving average, and the original data seem to fall approximately along a straight line when plotted on semilogarithmic paper, the moving geometric mean should be used in preference to the moving arithmetic mean.

logarithm of the corresponding figure in the fourth column. Comparison of the second and the last columns in the table will demonstrate at once that the moving average is neither too high nor too low, but exactly right for this particular problem.

It must not be thought that the moving geometric mean will solve our difficulties whenever there is a curvilinear trend. The moving geometric mean gives the correct answer only when the values of the trend form a geometric progression, as is shown by the data falling approximately in a straight line on semi-logarithmic paper. The fact that the method works so well with the data of Tables 10.8 and 10.9 is merely because these

TABLE 10.9.—COMPUTATION OF MOVING GEOMETRIC MEAN

Year	Output	Logarithm of Output	Moving Average	Anti-logs
1932	316.2	2.5000	?	?
1933	251.2	2.4000	?	?
1934	199.5	2.2999	2.3000	199.5
1935	158.5	2.2000	2.2000	158.5
1936	125.9	2.1000	2.1000	125.9
1937	100.0	2.0000	2.0000	100.0
1938	79.4	1.8998	1.8999	79.4
1939	63.1	1.8000	1.7999	63.1
1940	50.1	1.6998	1.6998	50.1
1941	39.8	1.5999	?	?
1942	31.6	1.4997	?	?

data were computed purposely to illustrate the point. It is quite possible for the moving geometric mean to give trend values which are further in error than those of the moving arithmetic mean.

The moving average has the advantages that all workers get the same result when they compute it, that it eliminates the short-run changes in the data and yet does not lose all its flexibility. In Table 10.5, for example, the moving average rises slowly at first and later more rapidly. This is exactly what happened to petroleum production. The increased steepness of the curve of the moving average merely reflects the facts. No straight line fitted to these data could show this fact, since the straight line increases by constant amounts.

Yet the moving average is subject to a serious disadvantage. In Table 10.5 there are question marks in the moving-average column opposite the years 1906, 1907, 1928, and 1929. The reason is obvious when one tries to compute the moving average for these years. What was the average production for 1907 and the two years before and after? We can find it for 1907 and the two years after and one of the years before; but we have no figure for 1905. Hence we have to omit the two years at each end of the series because the data before or the data after them are absent. If we had computed a three-year moving average, we should have had to omit one year at each end. Had we computed a four-year moving average, we should have had to omit two years at each end. Counting the omissions at both ends, we always have to omit one less year than the length of the period of the moving average if the number of years in the period is odd, or just as many years as there are in the period if the number is even.

Yet these extreme years are often the very ones in which we are most interested. What is "normal" production at present? It does us little good to be told that the average production around 1917 was 330 million barrels if we wish to know the 1929 average and cannot find it. The moving average suffers from the disadvantage that it cannot be extended to the extremes of the period studied, and of course the extension of the moving average to times outside the period is out of the question.

10.11. The Method of Least Squares.—In Secs. 10.5 and 10.6 we found that it was possible to describe some long-time trends by straight lines. In those sections we determined the location of the straight line subjectively, picking out the one that looked best. Yet we know that two equally competent statisticians might now draw exactly the same line. This raises the question as to whether or not some one line is better than another —whether or not we can pick out some one "best" line to describe the trend.

When the problem is stated in this way, we realize at once that there is a "best" trend line. It is the line along which the values would actually have moved if they had been subjected to the long-run forces only—if all cyclical and random forces had been eliminated. Petroleum production, for example, is subject to many temporary forces, such as strikes, transportation

holdups, seasonal fluctuations, and wartime demands. Actual production figures such as those of Table 10.5 reflect all these short-run movements as well as the basic, long-run changes which accompany the growth of population, the development of good roads, etc. The "best" trend line would be the one that would eliminate all the temporary forces, and show what would actually have happened to petroleum production if the long-run forces had been the only effective forces.

But unless one has faith in the crystal ball or the Ouija board, he can never know what would have been true if some forces had been different. We are therefore forced to guess what would have happened.[1] Yet some guesses are better than others. If I am asked to guess the height of someone, knowing nothing save that he is a Harvard student, I can make a better guess on the basis of the facts in Table 5.1, page 82, or on the basis of the statistical summary of these facts on page 211, than I can unassisted. If you are to toss a penny fifty times, and I am asked to guess how often it will fall with "heads" uppermost, I am wiser to base my guess on reasoning than I am to select a number at random. And in selecting trend lines it is also true that some guesses are better than others. An infinite number of straight lines can be drawn upon a chart, and all of them may slant approximately in the direction of the secular movement, but just as 25 "heads" are more likely in 50 throws of a coin than 28 or 30 heads, so one of these straight lines is more likely to be correct than any of the others.

Let us look at the chart in Fig. 10.9. Suppose we wish to represent the long-run movement of this chart by a straight line, as seems reasonable from casual inspection of the data. It is immediately apparent that no straight line will describe what happened in the sense that it will pass through the various points on the diagram. The points do not lie along any straight

[1] Some people may prefer to dignify the processes involved here by calling them "estimates" rather than "guesses." The name is really not important if the student understands that the process is one based on reasoning. In practice it seems to be true that students more often put too much faith in the results of least squares than too little. They think that somehow the mathematical processes of the least-squares method give them an answer that is "correct," rather than an estimate or guess of what is correct. It is to offset this tendency toward blind and innocent acceptance that I prefer to speak of the processes involved as guesswork rather than as estimation.

line. Therefore any straight line which we draw will have errors. We might draw a line so high on the diagram that all the actual points would lie below it; on the other hand, we might place

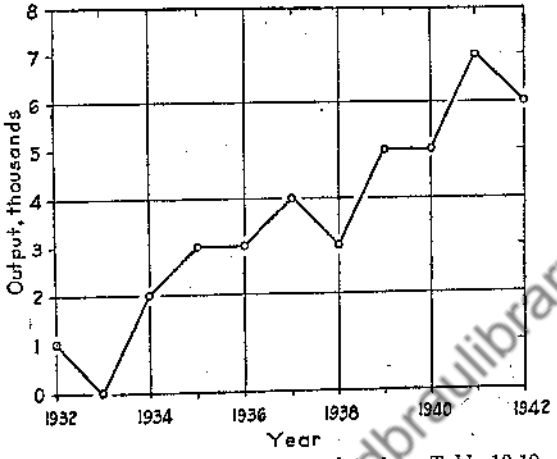


FIG. 10.9.—Hypothetical output data from Table 10.10.

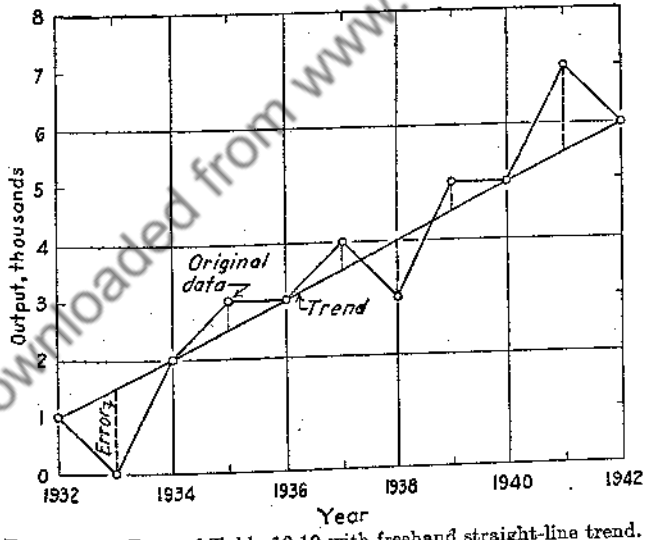


FIG. 10.10.—Data of Table 10.10 with freehand straight-line trend.

our line so low that all the actual points would lie above it. In practice we would be more likely, however, to draw a line something like the one in Fig. 10.10, such that some of the actual

points lie above and some below the line. In this case we can show our errors by means of the light dotted vertical lines connecting each of the original points with the straight trend line.

If we assume, now, that our trend is to show the ordinary or expected course of events (the ordinary course of petroleum production when undisturbed by temporary variations, for example), and that the variations around the trend are due to more or less chance occurrences, we can determine the chances that any particular set of errors or deviations would occur. Under these assumptions it can be shown that that trend line is most likely from which the sum of the squared deviations is a minimum. Perhaps we can illustrate the meaning of these terms best by using the data of Fig. 10.10 as an example. The original data of this figure are shown in the first two columns

TABLE 10.10.—ILLUSTRATION OF SQUARED ERRORS FROM DATA OF FIG. 10.10

Year	Output (000)	Trend Value	Error	Squared Error
1932	1	1.0	0.0	0.00
1933	0	1.5	−1.5	2.25
1934	2	2.0	0.0	0.00
1935	3	2.5	0.5	0.25
1936	3	3.0	0.0	0.00
1937	4	3.5	0.5	0.25
1938	3	4.0	−1.0	1.00
1939	5	4.5	0.5	0.25
1940	5	5.0	0.0	0.00
1941	7	5.5	1.5	2.25
1942	6	6.0	0.0	0.00
Sum				6.25

of Table 10.10. The third column shows for each year the height of the straight line trend which appears on the chart. The fourth column shows the amount of the error or the residual, found by subtracting the trend value from the actual value. The last column shows the squares of these residuals, and at the bottom of this last column is the sum of the squared residuals, or the sum of the squared errors. In this particular case the sum amounts to 6.25.

If we were to draw other straight lines on the chart, we should get other sets of errors, other squared errors, and different sums

of the squared errors. For example, the student might try using as trend values the numbers 0.5, 1.1, 1.7, 2.3, 2.9, 3.5, 4.1, 4.7, 5.3, 5.9, and 6.5. The difference between each successive pair of numbers in this series is 0.6, and if they are plotted on Fig. 10.10, they will give a straight line which is slightly steeper than the straight line already shown in that figure and crosses it at the middle of the chart. If the student will make up a table similar to Table 10.10, using these new trend values, computing his new errors and new squared errors, and adding his column of squared errors, he will find a sum of 5.15 instead of the sum 6.25 discovered in Table 10.10. The fact that the sum of the squared errors is smaller in the new case than in the old one means that the new line is, on the basis of our assumptions, more likely to be right than the old line. It does not show that it *is* the correct line, and we never do know what is the correct line. But the smaller the sum of the squared errors, the more likely the line is to be correct.[1]

If we wanted to find the one straight line which fitted the data best of all—which gave a smaller sum of squared errors than any other straight line—it would evidently take too long to go at it by trial and error. We cannot try 20 or 60 or 150 different lines, in each case computing the sum of the squared errors, and finally select the line with the smallest sum of the squared errors. This would take too long. But fortunately we can find the line we want by a very simple method. This method is called the *method of least squares* because it gives us the one line from which the sum of the squares of the errors is the smallest possible for any line of the type being fitted. We shall see how to find the "best fitting" straight line, the "best fitting" second-degree parabola, and the "best fitting" reciprocal curve by the method of least squares. The student must remember that the least-squares line is not necessarily the best one. In the first place, the straight line fitted by least squares is merely more likely to be right *than any other straight line.* Perhaps the basic trend was not a straight line at all. In that case the straight line fitted by least squares will not be the correct line. Similarly a second-degree parabola fitted by least squares is more likely to be correct

[1] For a simple proof of this statement, under our assumptions that the errors are pure chance affairs, see F. L. Griffin, "Introduction to Mathematical Analysis," pp. 456–457, Houghton Mifflin Company, Boston, 1921.

than any other second-degree parabola. We can generalize by saying that when we fit any line or curve by the method of least squares we get the line or curve that is more likely to be correct than any other line of the particular "family" which we could fit. Even so, it is more likely to be correct than other lines of the same family only if we are correct in our assumption that the errors or the residuals around the line are the results of random chance forces. In such a case the residuals will tend to be normally distributed, most of them clustered close to the line, with points getting less and less common as we get farther and farther from the line, and with points above and below the line being approximately evenly balanced.

10.12. Fitting a Straight Line by Least Squares.—We learned in Sec. 10.6 that every straight line will have the general form

$$Y = a + bX$$

We wish to find the values of a and b that will give the one straight line which fits best (the "best fit" being defined as that which minimizes the sum of the squared residuals or deviations). It is easy to show that these values of a and b can be determined from the following two *normal equations:*[1]

$$Na + b\Sigma X = \Sigma Y$$
$$a\Sigma X + b\Sigma X^2 = \Sigma XY$$

In order to solve these two equations we need the following values:

$$\begin{matrix} N & \Sigma X & \Sigma Y \\ \Sigma X^2 & & \Sigma XY \end{matrix}$$

[1] Each year (or other period of time) is designated by X.

Each value of the other variable (as petroleum production) is designated by Y.

The equation of the straight line at any point X is $Y = a + bX$.

If the point does not fall on the line, the distance from the point to the line (that is, the deviation) is represented by d. Thus

$$\begin{aligned} Y + d &= a + bX \\ d &= a + bX - Y \\ d^2 &= (a + bX - Y)^2 \end{aligned}$$

This is true for each deviation. If we sum all such terms, to get the sum of the deviations, we have

$$\Sigma d^2 = \Sigma(a + bX - Y)^2$$

This is the value we wish to minimize. Let us represent it by f. We mini-

If we list the values of petroleum production from 1917 to 1929 from Table 10.3, page 271 (and thus use the same years that

TABLE 10.11.—COMPUTATION OF STRAIGHT-LINE TREND OF PETROLEUM PRODUCTION, 1917–1929

Year	Production (Y)	Year (origin 1923) (X)	XY	X^2
1917	335	−6	−2,010	36
1918	356	−5	−1,780	25
1919	378	−4	−1,512	16
1920	443	−3	−1,329	9
1921	472	−2	− 944	4
1922	558	−1	− 558	1
1923	732	0	0	0
1924	714	1	714	1
1925	764	2	1,528	4
1926	771	3	2,313	9
1927	901	4	3,604	16
1928	901	5	4,505	25
1929	1,007	6	6,042	36
Totals.....	8,332	0	10,573	182

we used in fitting the straight line by freehand methods on page 277), we get Table 10.11. It will be noticed that in this table

mize by setting the partial derivatives with respect to a and b equal to zero. That is,

$$\frac{\partial f}{\partial a} = 2\Sigma(a + bX - Y) = 0$$

$$\frac{\partial f}{\partial b} = 2\Sigma(a + bX - Y)X = 0$$

Dividing by 2 and then summing as directed, we get

$$\Sigma a + b\Sigma X - \Sigma Y = 0$$

$$a\Sigma X + b\Sigma X^2 - \Sigma XY = 0$$

Since Σa (when a is a constant) $= Na$, we have, by transposing the last terms,

$$Na + b\Sigma X = \Sigma Y$$

$$a\Sigma X + b\Sigma X^2 = \Sigma XY$$

If we solve these equations for a and b after substituting the proper values of N, ΣX, ΣY, ΣX^2, and ΣXY, we get the values of a and b which minimize Σd^2. These are what we want.

we have let X represent the time variable (years) and Y the other variable (petroleum production), as we always do. Also it will be noted that we have taken as our origin of time the year in the middle, 1923, calling it the year 0, and reckoning the other years plus and minus from 1923. We almost always find it advantageous to use the center of our period as the time origin, as we discovered in Sec. 10.7, since it reduces by a good deal the amount of arithmetical computation.

We now have the necessary values to substitute in the normal equations, as follows:

$$N = 13 \qquad \Sigma X = 0 \qquad \Sigma Y = 8{,}332$$
$$\Sigma X^2 = 182 \qquad \Sigma XY = 10{,}573$$

Substituting these values in the normal equations, we get

$$13a + 0b = 8{,}332$$
$$0a + 182b = 10{,}573$$

The shift of the time origin to a recent date gave us small numbers to work with. The fact that we shifted to the central year made $\Sigma X = 0$ and reduced the labor of computation still more. If we solve these two equations, we have the following values:

$$a = 641$$
$$b = 58.1$$

Hence our equation for the straight line, which we get by substituting these values for a and b in the type equation

$$Y = a + bX$$

is

$$Y = 641 + 58.1X$$
Origin 1923

We can now estimate the values for other years. If we wish to draw the trend on a graph of the data, we first plot the original data. We then estimate the trend value for the first and last years (since these two points, like any other two points, will determine the location of a straight line). The first year, 1917, is year -6 when the origin is 1923. If we substitute -6 for X in our equation and solve for Y, we get

$$Y = 641 + 58.1(-6) = 292.4$$

We thus know that the trend value for 1917 is 292.4, and we locate this point on the graph opposite the year 1917. Similarly we find the trend value for the year 1929 (the year +6) to be

$$Y = 641 + 58.1(6) = 989.6$$

This value we locate opposite the year 1929. Thus we have two points, and we connect them with a straight line. This is the trend line (see Fig. 10.13, page 310).

It is not necessary to change the time origin, and if it is changed it is not necessary to locate it at the central year. Such changes, however, minimize arithmetical computation.

10.13. Meaning of Constants in Regression Equation.—Perhaps one more word should be added regarding the formula for the straight line which we fitted by the least-squares method on page 298. That formula, it will be recalled, was

$$Y = 641 + 58.1X$$
(Origin 1923)

First suppose that we want to estimate the 1923 petroleum production. This year is the year of origin, so that the deviation from the origin is 0. Substituting 0 for X in the equation, we get

$$Y = 641$$

In other words, when we determine the value of a in the trend equation we are really finding the trend value at the time origin. We were in actuality estimating the 1923 petroleum production. This is always true of the formula for the straight line and also for the other formulas, such as that which we shall shortly find for the second-degree parabola. That is, a = the value of Y at the time origin.

What will be the value of Y one year after the origin (that is, in 1924)? Obviously it will be 641 + (1)(58.1). What will it be two years after the origin (in 1925)? Obviously 641 + (2)(58.1). And the third year it will be 641 + (3)(58.1); etc. In other words, the value of b (58.1) is evidently the amount which is added to production every year as estimated by our trend. The straight-line trend must, on account of the nature of a straight line, rise the same amount each year. The value of b tells us the amount by which it tends to rise each year. In some years petroleum production increased more than in other years, but our formula tells us that the tendency was for it to

rise 58.1 million barrels a year throughout the period (during the period to which the trend was fitted, which was 1917–1929). The 1929 production was 1007 million barrels, and the 1917 production was 335 million barrels. The actual increase in production was, then, 672 million barrels in 12 years, or 56 million barrels per year. The value of b is not the actual average increase, which could be computed easily by merely dividing the difference between the output of the first and the last year by the number of years. But neither the first nor the last year is a "normal" year; neither of them falls exactly on the trend. We are likely to be led astray if we base our estimate of the rate of increase in production for the entire period on the output in these two years alone. The figure 58.1 million barrels (the slope of the trend line) is in some ways a much more useful figure. Insofar as the data tended to increase by a fixed amount each year, we can best state that amount as 58.1 million barrels.

We see, then, that the value of a tells us the value of Y at the time of origin; and the value of b tells us the amount of the increase or decrease along the trend line per unit of time. In our problem the value of a and the value of b are both positive; either or both may be negative. If a is negative, the value of Y is negative at the time of origin (just as temperatures may be below zero, or as profits may become losses and be expressed negatively). If b is negative, it is evident that we are subtracting more and more each year, and that the trend line is falling. Rising trends have positive values of b, and falling trends have negative values of b.

10.14. Fitting a Second-degree Parabola by Least Squares.— By this method one can also fit curves of various kinds to data. If one were to fit a second-degree parabola, the normal equations would be[1]

$$Na + b\Sigma X + c\Sigma X^2 = \Sigma Y$$
$$a\Sigma X + b\Sigma X^2 + c\Sigma X^3 = \Sigma XY$$
$$a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 = \Sigma X^2Y$$

[1] The student who is interested and qualified can derive these equations for himself. The process parallels that involved in computing the regression equations for the straight line, except that the residuals are defined thus:

$$d = a + bX + cX^2 - Y$$

The other steps are the same, and one differentiates partially with respect to a, b, and c. The solution yields the normal equations given above.

If we wish to fit such a parabola to the data of petroleum production throughout the entire period 1906–1929, as we did by the method of selected points on page 282, the computation would be that of Table 10.12, in which the origin is taken at year 1917½, and the deviations are in half years. Thus, 1917 is

TABLE 10.12.—COMPUTATION OF SECOND-DEGREE PARABOLIC TREND OF PETROLEUM PRODUCTION, 1906–1929

Year	Output (Y)	Year (origin, 1917.5) (X)	X^2	XY	X^2Y	X^3	X^4
1906	126	−23	529	−2,898	66,654	−12,167	279,841
1907	166	−21	441	−3,486	73,206	− 9,261	194,481
1908	179	−19	361	−3,401	64,619	− 6,859	130,321
1909	183	−17	289	−3,111	52,887	− 4,913	83,521
1910	210	−15	225	−3,150	47,250	− 3,375	50,625
1911	220	−13	169	−2,860	37,180	− 2,197	28,561
1912	223	−11	121	−2,453	26,983	− 1,331	14,641
1913	248	− 9	81	−2,232	20,088	−729	6,561
1914	266	− 7	49	−1,862	13,034	−343	2,401
1915	281	− 5	25	−1,405	7,025	−125	625
1916	301	− 3	9	−903	2,709	−27	81
1917	335	− 1	1	−335	335	−1	1
1918	356	+ 1	1	356	356	1	1
1919	378	3	9	1,134	3,402	27	81
1920	443	5	25	2,215	11,075	125	625
1921	472	7	49	3,304	23,128	343	2,401
1922	558	9	81	5,022	45,198	729	6,561
1923	732	11	121	8,052	88,572	1,331	14,641
1924	714	13	169	9,282	120,666	2,197	28,561
1925	764	15	225	11,460	171,900	3,375	50,625
1926	771	17	289	13,107	222,819	4,913	83,521
1927	901	19	361	17,119	325,261	6,859	130,321
1928	901	21	441	18,921	397,341	9,261	194,481
1929	1,007	23	529	23,161	532,703	12,167	279,841
Totals.....	10,735	0	4,600	85,037	2,334,391	0	1,583,320

½ year before the origin, 1916 is ³⁄₂ years before the origin, 1920 is ⁵⁄₂ years after the origin, etc. The reason for this adjustment will soon be evident.

Here the arithmetical work is tedious enough at best, but it is clear that the symmetry which results from taking a time origin

at the center of the period has reduced it tremendously. This has made it unnecessary to add two of the columns at all and has made it possible to get the sum of two others by adding half the column and multiplying the sum by 2, since the other half of the column merely duplicates the first half.

Let us now substitute the totals at the proper points in the normal equations. We get

$$\begin{aligned} 24a + 0b + 4600c &= 10{,}735 \\ 0a + 4600b + 0c &= 85{,}037 \\ 4600a + 0b + 1583320c &= 2{,}334{,}391 \end{aligned}$$

From the second of the normal equations we see at once that the value of b is 18.5. Solving for the others, we find the three values to be

$$\begin{aligned} a &= 371.8 \\ b &= 18.5 \\ c &= 0.394 \end{aligned}$$

If we substitute these values in the type equation

$$(Y = a + bX + cX^2)$$

we get

$$Y = 371.8 + 18.5X + 0.394X^2$$

Origin 1917.5

Deviations in half years

It is necessary in this case to state the origin and also the fact that the deviations are in half years. Let us now estimate the petroleum output for the year 1927 by this formula. The year 1927 is 9.5 years after the origin, but since we are measuring in half-year units we must convert the 9.5 years to half years, getting 19 half years for our value of X. This we substitute for X in the formula to get

$$Y = 371.8 + 18.5(19) + 0.394(19^2) = 865.5$$

Thus this formula gives an estimate of 865.5 million barrels for 1927. If we estimate the trend value for each of the 24 years by substituting the various values of X in our formula, and if we locate the estimates on a graph of the data, we can

easily draw our curvilinear trend through the points so estimated (see Fig. 10.14, page 313). This curve will give better estimates than any other second-degree parabola, just as the straight line fitted to these data by least squares gives better estimates than any other straight line.

10.15. Fitting a Reciprocal Curve by Least Squares.—Fitting other types of curves by the method of least squares should now be easy. If we once have the requisite normal equations, all we have to do is to find the values to be substituted, and solve our equations. Reciprocal curves are of the general type

$$\frac{1}{Y} = a + bX$$

The normal equations are

$$Na + b\Sigma X = \sum\left(\frac{1}{Y}\right)$$

$$a\Sigma X + b\Sigma(X^2) = \sum\left(\frac{X}{Y}\right)$$

If we take our origin at the center of the period to which the trend is being fitted, the terms involving ΣX will equal zero, and can be dropped. In this particular case we can find the values of a and b from these formulas

$$a = \frac{\Sigma(1/Y)}{N}$$

$$b = \frac{\Sigma(X/Y)}{\Sigma(X^2)}$$

Taking a purely hypothetical example, suppose that the first two columns of Table 10.13 represent the average cost of producing one unit of product in a given factory over a period of years. It is obvious that the cost has been falling and also obvious (see Fig. 10.11) that the trend line is curvilinear, falling more and more slowly. This is the type of trend which can often be described by a reciprocal curve, as we shall point out later (see Sec. 10.17). We need values of N, $\Sigma(1/Y)$, $\Sigma(X/Y)$, and $\Sigma(X^2)$. These are found in the later columns of the table. Since we have taken the origin at the center of the period, we can substitute directly in the simpler of the sets of equations just given, getting the following values of a and b for this problem:

TABLE 10.13.—ILLUSTRATION OF FITTING RECIPROCAL CURVE BY LEAST SQUARES

Year	Cost per unit (Y)	Year (origin 1936) (X)	$(1/Y)$	(X/Y)	(X^2)
1929	$4.10	−7	0.2439	−1.707	49
1930	3.25	−6	0.3077	−1.846	36
1931	2.90	−5	0.3448	−1.724	25
1932	2.50	−4	0.4000	−1.600	16
1933	2.25	−3	0.4444	−1.333	9
1934	2.00	−2	0.5000	−1.000	4
1935	1.80	−1	0.5556	−0.556	1
1936	1.65	0	0.6061	0.000	0
1937	1.55	1	0.6452	0.645	1
1938	1.40	2	0.7143	1.429	4
1939	1.33	3	0.7519	2.256	9
1940	1.25	4	0.8000	3.200	16
1941	1.18	5	0.8475	4.237	25
1942	1.10	6	0.9091	5.455	36
1943	1.05	7	0.9524	6.667	49
Totals			9.0229	14.123	280

$$a = \frac{9.0229}{15} = 0.602$$

$$b = \frac{14.123}{280} = 0.050$$

Substituting these in our type equation $(1/Y) = a + bX$, we get

$$\frac{1}{Y} = 0.602 + 0.050X$$

Origin 1936

If we wish to estimate the cost per unit of output by this equation for the year 1943, we note that this is the year 7 under the terms of our problem. We substitute 7 for X in the equation to get

$$\frac{1}{Y} = 0.602 + (0.050)(7)$$

$$= 0.602 + 0.350 = 0.952$$

$$Y = 1.05$$

We can compute our estimates for the other years by substituting other values for X in the equation, and if we plot the estimates

for all the years on the diagram of Fig. 10.11, we can easily connect them with a smooth curve which shows the trend. Since it has been fitted by least squares it is more likely to be right than any other reciprocal curve that could be fitted (if we assume that the deviations from the curve are the result of chance forces).

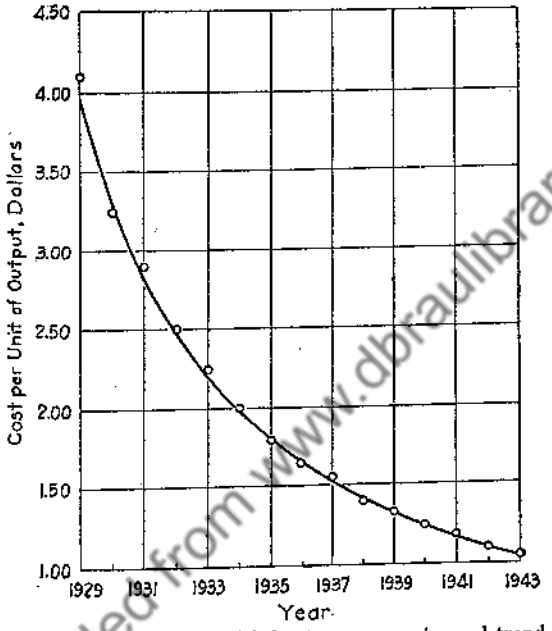


FIG. 10.11.—Data of Table 10.13 with least-squares reciprocal trend. Original data shown by small circles, and trend values shown by smooth curve.

10.16. Fitting a Semilogarithmic Curve by Least Squares.—By now the method of least squares should be second nature to the student, and a very sketchy description of the method of fitting a geometric trend, or semilogarithmic curve, should suffice. The general formula for such trends is[1]

[1] This formula is also sometimes given as

$$Y = cd^x$$

If the student will take logs of both sides of this equation, he will find that it becomes

$$\log Y = \log c + (\log d)X$$

But since c and d are constants, we can let a stand for $\log c$ and b stand for

$$\log Y = a + bX$$

The normal equations are

$$Na + b\Sigma X = \Sigma(\log Y)$$
$$a\Sigma X + b\Sigma(X^2) = \Sigma(X \log Y)$$

If we take our origin at the center, our simplified formulas become

$$a = \frac{\Sigma(\log Y)}{N}$$
$$b = \frac{\Sigma(X \log Y)}{\Sigma(X^2)}$$

We illustrate the use of the method by fitting a semilogarithmic trend to the population of the United States at the first seven censuses. The data appear in Table 10.14.

TABLE 10.14.—ILLUSTRATION OF FITTING SEMILOGARITHMIC TREND BY LEAST SQUARES TO U.S. POPULATION FIGURES

Year	Population (millions) (Y)	Decade (origin 1820) (X)	$\log Y$	$X \log Y$	X^2
1790	3.9	−3	0.5911	−1.7733	9
1800	5.3	−2	0.7243	−1.4486	4
1810	7.3	−1	0.8633	−0.8633	1
1820	9.6	0	0.9823	0.0000	0
1830	12.9	1	1.1106	1.1106	1
1840	17.1	2	1.2330	2.4660	4
1850	23.2	3	1.3655	4.0965	9
Totals			6.8701	3.5879	28

Solving for a and b we get

$$a = \frac{6.8701}{7} = 0.981$$
$$b = \frac{3.5879}{28} = 0.128$$

---

log d. This gives us

$$\log Y = a + bX$$

as above.

Our equation for this particular problem becomes[1]

$$\log Y = 0.981 + 0.128X$$
$$\text{Origin } 1820$$

We can estimate the population for any time from the equation. Suppose we want the estimated population for 1807. This is 13 years before the origin, but since our formula reckons in decades we must count it as -1.3 decades, and substitute -1.3 for the value of X. This gives us

$$\log Y = 0.981 + (0.128)(-1.3)$$
$$= 0.981 - 0.166 = 0.815$$
$$Y = 6.53$$

Our original values of Y were in millions, so our answer is also in millions. Our estimate is that the 1807 population was 6,530,000. The original figures and the semilogarithmic trend appear in Fig. 10.12.

10.17. How to Decide What Trend to Use.—We have just illustrated methods of fitting four different trends: straight line, parabolic, reciprocal, and semilogarithmic. One could take any given set of data (say the population figures of Table 10.14) and fit any one or all of these types of trends to the figures. How is one to know what sort of trend to fit? This problem is discussed in some detail in a later chapter (see Sec. 14.3), but we can give these brief comments here.

Before fitting any trend at all, the statistician plots his data on a chart. If the data fall along a straight line, or, even though they are not on a straight line, if it looks as though a straight line would give a fair picture of the general, long-run

[1] If we prefer the equation in the nonlogarithmic form mentioned in the preceding footnote, we need merely remember that we let a stand for $\log c$ and b stand for $\log d$. Taking antilogs we have $c = 9.57$ and $d = 1.34$. Thus our equation becomes

$$Y = (9.57)(1.34)^X$$

This statement of the equation has the advantage that the value of c tells us the trend at the time of origin (here the trend was 9.57 million people in 1820), and d tells us the rate of increase per unit of time. In this problem the unit of time is the decade, and our equation tells us that we are to multiply the population of any decade by 1.34 to get that of the following decade. Thus we see that population was increasing at the rate of 34 per cent each decade.

tendency,[1] the statistician will fit a straight line of the general form $Y = a + bX$ as described in Sec. 10.12. If the data seem to fall along some curve, there are a number of tests that can be applied to find what sort of curve to fit (see pages 443*ff.*). But we can start by making a chart showing the reciprocals of the data rather than the data themselves. If this chart seems to show a straight line (when we plot values of $1/Y$ instead of

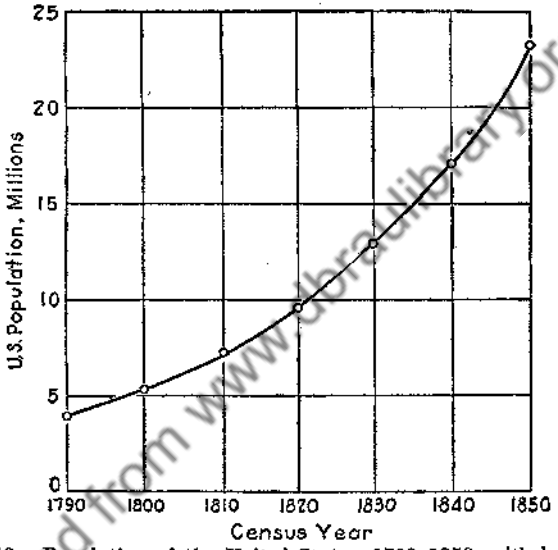


FIG. 10.12.—Population of the United States, 1790–1850, with least-squares logarithmic trend. The dots represent actual census population figures, and the smooth curve represents the trend.

values of Y), we should fit a reciprocal curve of the general form $1/Y = a + bX$ as described in Sec. 10.15. If the chart does not show a straight line, we can try plotting values of log Y instead of values of Y (or, what amounts to the same thing, we can make our original chart on semilogarithmic paper). If this chart shows a straight line, we should fit a geometric trend (a semilogarithmic trend) of the general form $\log Y = a + bX$ as described in Sec. 10.16. This is the sort of case referred to in Sec. 5.19 as one of the times when the geometric mean should be used in preference to the arithmetic mean. Finally, if neither

[1] The student will note from Fig. 10.13, page 310, that the points do not actually fall along the trend line, but that the straight trend line does seem to give a reasonable picture of the long-run movement.

of these methods "straightens out" the curve, we can try taking second differences as described in Sec. 14.3 page 445. If these second differences are constant, we should fit a second-degree parabola of the general form $Y = a + bX + cX^2$ as described in Sec. 10.14.

There are many other forms of curves that can be fitted to data—third-degree parabolas and parabolas of higher degrees, double logarithmic curves in which both variables must be converted to logarithmic form, Pearl-Reed curves, Gompertz curves, and many other more complicated sorts with specialized uses which are described in the more advanced literature. But the simple forms of curvilinear trends described here will cover most of the cases which the statistician meets, and they all have the advantage that they can be fitted by simple, straightforward methods.

The choice of the proper curve to use is a matter that requires a good deal of experience and judgement. It is one of the most important of decisions to make in the whole problem of studying trends, yet it is subjective and nonmathematical—more nearly art than science. There are no sure-fire, hard and fast rules to follow, and the student is required to rely on his own good sense and his knowledge of the data to lead him in the right direction.

While we are considering normal equations, we may note here the normal equations for the third-degree parabola fitted by least-squares methods. The type equation is, of course,

$$Y = a + bX + cX^2 + dX^3$$

We must obtain values of a, b, c, and d. The normal equations are

$$\begin{aligned} Na + b\Sigma X + c\Sigma X^2 + d\Sigma X^3 &= \Sigma Y \\ a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4 &= \Sigma XY \\ a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5 &= \Sigma X^2Y \\ a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6 &= \Sigma X^3Y \end{aligned}$$

From this point the student should be able to carry out the work himself, and also to see how other and more complex normal equations would be formed.

Among the principal advantages of the least-squares method of fitting the secular trend are the following:

1. There is but one possible answer, and (aside from mistakes in arithmetic) all workers get the same result.

2. The trend can be easily and succinctly described by a mathematical formula.

3. If the residuals are normally distributed around the trend in a chance distribution, it can be shown that the trend fitted by this method is more likely to be the "true" trend than any other line of the same general form.

10.18. Residuals from the Trend.—We have seen that the line fitted by least squares minimizes the sum of the squared

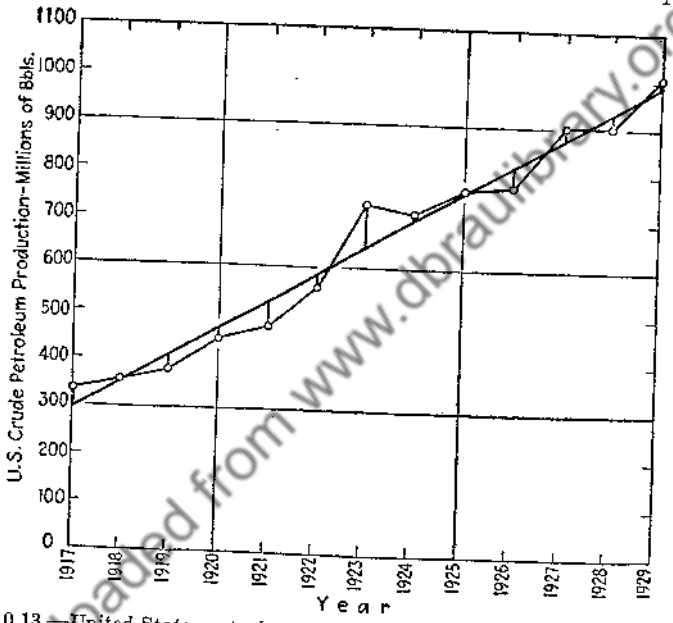


FIG. 10.13.—United States petroleum production, 1917–1929, with least-squares straight-line trend, and with deviations from trend indicated.

residuals. Just what are these *residuals* to which we refer? An illustration will make the matter clear. Figure 10.13 shows petroleum production from 1917 to 1929. We have added the secular trend computed by the formula we discovered on page 298. It will be noted that the trend does not coincide with the actual petroleum production in each year. In fact, the purpose of the trend is to avoid the minor fluctuations in the data and give a picture of the general tendency.

If, then, we estimate the production for any year by the trend formula, we shall usually be somewhat in error. Our estimates

will fall along the trend line, although the actual amounts of production were not on the line. The amount of our error can be seen by inspection of the chart. In 1917 the actual production was above the trend. Our estimate was too low by an amount equal to the small vertical line joining the actual production for 1917 with the trend value in 1917. The 1918 estimate was almost correct, but in 1919 the trend gives a figure that is too high. The amount of the error (or residual) is shown by the short vertical line connecting the actual 1919 production with the trend. In each year we have some error. If we let Y = the actual production and Y' = the estimate made from the trend, then the amount of the error (the residual) can be defined as

$$d = Y - Y'$$

Sometimes the residual will be positive, sometimes negative, and, in those cases where the trend line passes exactly through one of the points on the diagram, the residual will equal zero.[1]

It is not necessary to read the errors from the chart. Such a method would not be particularly accurate. Our trend equation makes it possible for us to estimate the trend values for each of the years (as we did for two of the years on page 299). If we estimate the trend value for each year and subtract it from the actual value (as shown in the table on page 312), we have computed the actual residual by the formula given above. We can then square these residuals and add the squares. This would give us the sum of the squared residuals.

Let us do this for the present example. We have already found that the trend value for 1917 is 292.4 and, for 1929, 989.6 (see page 299). Computing the trend values for the other years by the same process, we get Table 10.15.

It will be noted that the sum of the deviations is very small (−1.0). As a matter of fact had we not dropped decimals in computing the trend equation, the trend values, and the residuals, the sum of the residuals would equal zero. Trends fitted by the method of least squares give residuals whose algebraic sum is zero.

But it is the sum of the last column in which we are particularly interested. The fourth column (the residuals themselves) shows

[1] This definition of the residual is mathematically the same as the definition given on p. 296*n*.

for each year the distance from the trend line to the actual production. In other words, these figures are the lengths of the short vertical lines in Fig. 10.13 on page 310. The last column shows the squares of the residuals, and the sum of the squared residuals is 19,257.42. This seems like a large figure. But if we pass any other straight line through these data, no matter how, and if we find the sum of the squares of the deviations from this new line, the sum will be greater than 19,257.42. This line has been so fitted that the sum of the squared residuals is the smallest that can exist for any straight line. Similarly

TABLE 10.15.—RESIDUALS AROUND THE LEAST-SQUARES TREND

Year (X)	Actual Production (Y)	Trend Value (Y')	Residual ($Y - Y'$)	Residual Squared (d^2)
1917	335	292.4	+42.6	1,814.76
1918	356	350.5	+ 5.5	30.25
1919	378	408.6	−30.6	936.36
1920	443	466.7	−23.7	561.69
1921	472	524.8	−52.8	2,787.84
1922	558	582.9	−24.9	620.01
1923	732	641.0	+91.0	8,281.00
1924	714	699.1	+14.9	222.01
1925	764	757.2	+ 6.8	46.24
1926	771	815.3	−44.3	1,962.49
1927	901	873.4	+27.6	761.76
1928	901	931.5	−30.5	930.25
1929	1,007	989.6	+17.4	302.76
Totals			− 1.0	19,257.42

when we fitted a second-degree parabola by the method of least squares on page 300, we fitted it in such a way that the sum of the squared residuals was smaller than it could be with any other second-degree parabola.

One point should be reemphasized. The straight line fitted by the method of least squares does not give the smallest sum of the squared residuals of all possible lines, but only of all possible straight lines. The straight line fitted to the data is not necessarily (even under our assumption of a chance distribution) the line of best fit; but it is the straight line of best fit. A parabola fitted by these methods might fit better; at any rate, a parabola

so fitted would give the best fit of any parabola, etc. The statistician must decide for himself whether or not a straight line will describe the trend. The method of least squares cannot make any straight line give a good description of petroleum production for the entire period 1906–1929. We know that the straight line fitted to these data by the method of least squares

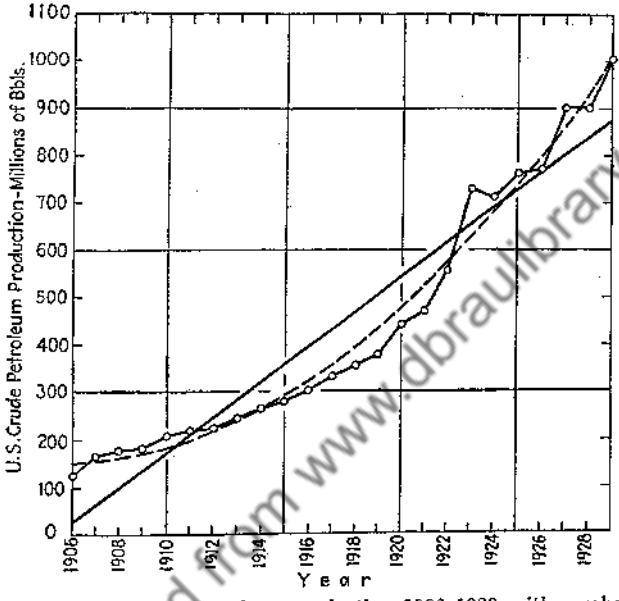


FIG. 10.14.—United States petroleum production, 1906–1929, with parabolic and straight-line trends. For trend formulas see pages 302 and 313*n*.

will fit better than other straight lines, but even so it will fit poorly. The statistician must use judgment in such matters. When we used the data for the entire period, we noted that the trend was curvilinear, so we fitted a curve. The result is shown in Fig. 10.14, where we have the data for the entire period and have fitted to it a straight line by least squares[1] and also the parabola which we computed on page 302. Although the straight line fits better than any other straight line, it obviously does not fit so well as the parabola.

[1] Formula:

$$Y = 447.3 + 18.49X$$

Origin 1917.5; half-year units

10.19. Elimination of Trend.—We noted earlier[1] that one may wish to describe a trend either because one is interested in the trend itself or because one wishes to get rid of the trend and study those movements which are left. We have now discovered how one describes a trend, and it remains merely to see how one might eliminate the trend from the data and why one might wish to do it.

Reverting to the problem discussed on page 312, we find that we have computed not only the trend value of petroleum production but also the deviation of production from the trend. If we take the figures from the fourth column of the table, we find that the deviations from the trend for various years were

Year	Production (deviation from the trend) (millions of barrels)
1917	+42.6
1918	+ 5.5
1919	−30.6
1920	−23.7
1921	−52.8
1922	−24.9
1923	+91.0
1924	+14.9
1925	+ 6.8
1926	−44.3
1927	+27.6
1928	−30.5
1929	+17.4

These figures show, not the actual production in any year, but the amount by which the actual production differed from the trend. If we think of the trend as representing the natural or normal or expected rate of production, then these figures in the table show by how much the actual production differed from normal. We can best illustrate, perhaps, by comparing the productions of two years. In 1928 the production was 901 million barrels, and in 1923 it was 732 million barrels. Evidently 1923 was a year of comparatively low production and

[1] See p. 275.

1928 a year of high production. But this is not so if we consider the trend to represent normal production, for in 1923 the output was 91 million barrels above normal and in 1928 the output was 30.5 million barrels below normal. To be sure, the 1928 output was larger than that of 1923, but there had been a general tendency for production to increase during the interim, and it had not increased so much between 1923 and 1928 as one should expect.

If we took a year of exceptionally large potato production for the United States in the decade 1840–1850 and a year of abnormally low potato production in the decade 1920–1930, we should almost certainly find that the "low" production of the twentieth century was larger (in terms of bushels) than the "high" production of the nineteenth century. Whether a production figure is high or low is a relative matter. When we say that petroleum production was high at any time we mean that it was higher than was to have been expected at that time; that is, it was above the trend. The price of potatoes would, in all probability, have been low in the year of "high" production in the 1840's; and when we had the year of "low" production in the 1920's, the price would presumably have been high. The fact that the 1920 production was greater than the 1840 production would not make prices low; the fact that the 1920 production was below the trend might make the prices high. For this reason we are often primarily interested in getting rid of the trend entirely in order that we may study the deviations from the trend.

Figure 10.15 shows the deviations of petroleum production from the trend for the years 1917–1929. It is a graph of the figures in the table on page 314. One might well be interested in trying to explain how it happened that the 1923 output was so very large as compared with the trend and why the 1928 output (which was actually larger) was so very small as compared with the trend.

Similarly Fig. 10.16 shows the deviations of petroleum production from the trend for the entire period 1906–1929, the deviations being taken from the parabolic trend which was fitted on page 302. It will be noted that the entire tendency for production to increase has disappeared; the tendency for the values on the chart to rise toward the right is gone. The trend has truly been "eliminated," and merely the deviations are left.

Once more we see that the 1928 production was below and the 1923 production above the trend.

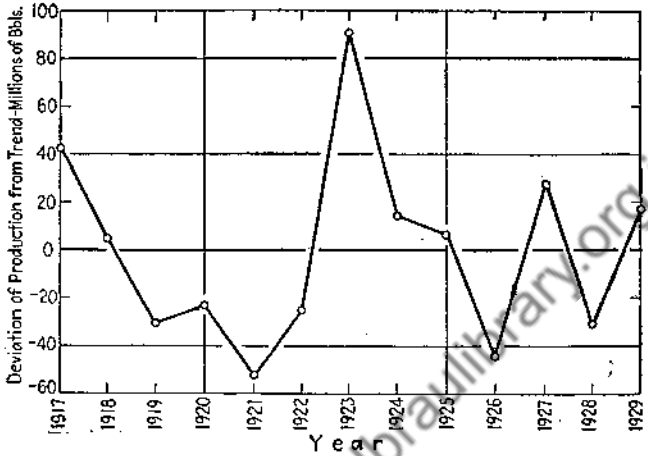


FIG. 10.15.—Deviation of United States petroleum production from straight-line trend, 1917–1929. Figures from table on page 314.

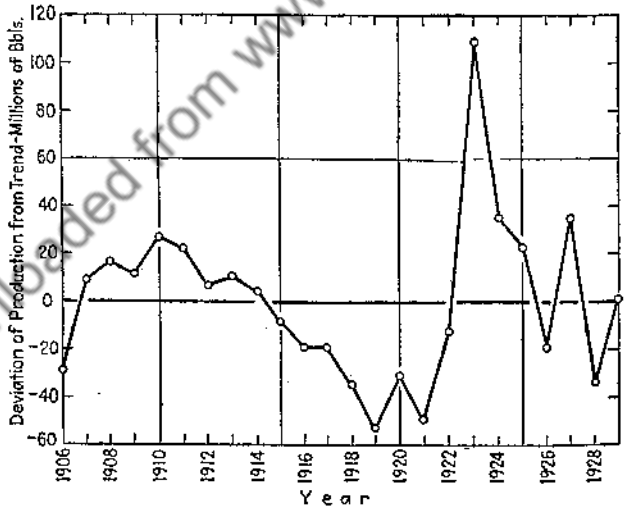


FIG. 10.16.—Deviation of United States petroleum production from the secular trend, 1906–1929. Deviations are from the parabolic trend on page 302.

10.20. Shifting the Origin of the Trend.—We have seen that it is possible to save considerable time in fitting a trend by choosing a time origin at the center of the period being studied.

This sometimes gives us results with an origin between two years and with deviations in half years when the period covered contained an even number of years. Fortunately it is very easy to change the origin of a straight-line trend after it has been computed. In this way we can take a central origin during the process of computation in order to facilitate the mathematical operations, and after the trend has been computed we can shift the origin to any desired year and change the deviations from half years to full years. This cannot be done so easily with curvilinear trends.

When we fitted a straight-line trend to the figures of petroleum production for 1906–1929, we found the following formula:

$$Y = 447.3 + 18.49X$$
Origin 1917.5; half-year units[1]

If we recall the meaning of the figures in this formula, it is understood that the value of Y (petroleum production) in the period of origin (1917.5) is 447.3 million barrels, and that the trend is increasing at the rate of 18.49 million barrels per unit of time.[2] Here the unit of time is the half year; so the trend is increasing at the rate of 18.49 million barrels each half year. If the value of the trend in 1917.5 is 447.3, then a half year later (in 1918) it must be 447.3 + 18.49 = 465.79 million barrels. We thus have the trend value for an even year rather than for a point halfway between two years. We also realize that if the trend rises 18.49 million barrels each half year, then it must rise 2(18.49) = 36.98 million barrels per year. But with these two figures it is easy to state the trend with the origin at 1918 and the units in full years, because we know that in our type equation, $Y = a + bX$, a is the value of the trend at the origin and b is the rate of increase (or decrease) per unit of time. We have just seen that the 1918 trend value is 465.79 and that the yearly rate of increase is 36.98. Substituting these figures for a and b in the type equation, we get the new trend equation

$$Y = 465.79 + 36.98X$$
Origin at 1918

[1] See p. 313*n*.
[2] See pp. 299–300.

Since the time units are years, we need make no special mention of them. The trend values computed from this equation will be exactly the same as those computed from the earlier equation and the figures will be more easily understood, since it is easier to understand the statement that the annual increase tends to be 37 million barrels (36.98) than that the semiannual increase is 18.49 million barrels. Mathematically the two statements are identical, but men are more accustomed to thinking in units of years than in units of half years. It is also true that our original data were annual data and that there is some advantage in having our conclusions in annual form.

When, therefore, one has computed a straight-line trend by the method of least squares from data which cover an even number of time intervals, and the equation of trend has been found with the origin between two periods and the rate of increase stated in units of half periods, the equation can easily be adjusted so that the origin is at an even period and the rate of change is stated in even periods. The process consists in computing the trend for some even year (ordinarily the next even period, but any year will do) and taking that point as the origin. In other words, any year may be taken as the origin, and the trend value for that year can be put in the trend equation as the constant a. The constant b is then doubled so as to give the rate for the full period rather than for half periods. It must be remembered that this simple and easy method will not work with curvilinear trends.

10.21. Suggestions for Further Reading.—The student will find some further discussion of the problem of fitting curves to data, and of determining what sorts of curves should be fitted, in Chap. XIV of this book. For a discussion of the general nature of the problem of the statistical treatment of time series, he is referred to the article on Time Series by Simon Kuznets in the "Encyclopaedia of Social Sciences," The Macmillan Company, New York, 1934. Reference should also be made to Frederick R. Macauley, "The Smoothing of Time Series," National Bureau of Economic Research, Inc., New York, 1931. A more advanced and difficult, but authoritative and exhaustive treatment, may be found in Max Sasuly, "Trend Analysis of Statistics," Brookings Institution, Washington, D.C., 1934. For a simple discussion of the Pearl-Reed and the Gompertz curves, see George R. Davies and Walter F. Crowder, "Methods of Statistical Analysis in the Social Sciences," John Wiley & Sons, Inc., New York, 1933, Chap. VI. The computation of the Gompertz curve is simply illustrated in Theodore H. Brown, Richmond F. Bingham, and V. A. Tem-

nomeroff, "Laboratory Handbook of Statistical Methods," Part II, Chap. VI, McGraw-Hill Book Company, Inc., New York, 1931.

For some understanding of the problems of calendar variation, the student should read the article on the Calendar in the Encyclopaedia Britannica, or the first chapter of Lancelot Hogben, "Science for the Citizen," Alfred A. Knopf, New York, 1938. Discussions of recent proposals for altering the calendar either by setting up 13 months of four weeks each or by setting up a 12-month system of equal quarters (the "World Calendar") can be found in the press and current magazines. The student should consult the Readers' Guide to Periodical Literature in his library. The World Calendar Association, 630 Fifth Avenue, New York City, will send literature relative to its proposals.

The student of a philosophical turn of mind, who is willing to spend some time on a subject usually considered less "practical," will find exciting reading if he investigates the general nature of the time concept. P. D. Ouspensky's "Tertium Organum," Alfred A. Knopf, New York, 1927, is perhaps the best, although Gerald Lynton Kaufman, "The Book of Time," Julius Messner, Inc., Publishers, New York, 1938, is also good. For a direct tie-up of the philosophical side of the time concept to scientific method, see Karl Pearson, "The Grammar of Science," No. 939 of the Everyman's Library, J. M. Dent & Sons, Ltd., London, 1937, Chap. V.; or W. F. G. Swann, "The Architecture of the Universe," The Macmillan Company, New York, 1934, especially Chap. X.

EXERCISES

1. Suppose that a business firm has monthly sales as follows:

January	$1329	July	$1403
February	1275	August	1350
March	1350	September	1226
April	1368	October	1019
May	1420	November	1224
June	1425	December	1283

Correct these figures for calendar variation. Make corrections for the number of business days in the months, computing on the basis of the current year. Subtract Sundays and whatever days are legal holidays in your state (see "World Almanac" for legal holidays). Compare the corrected figures with the original ones.

2. In some businesses it would be foolish to eliminate holidays in correcting for calendar variation? Why? In what businesses, for example?

3. Two proposed reforms of the calendar are known as the 12-month plan and the 13-month plan. What are the differences? What are the advantages and disadvantages of each? Does either of them make it possible for business men or statisticians to forego correction for calendar variation? Discuss.

4. Find at least two historical series which show marked linear upward trends, at least two with marked downward linear trends, at least two with curvilinear trends.

5. Plot on graph paper a historical series in which there is a marked and fairly linear trend. Lay over the graph a piece of tracing paper and ask some other member of the class to draw the trend by freehand methods. Mark the paper in such a way that you can locate again the position of the line. Have several other people draw trends, each working independently. Compare the results.

6. Suppose that you have computed a secular trend by the method of least squares. You have started with data showing weekly figures for 52 weeks. For this reason you have taken as the origin week 26.5, and your formula turns out to be (in half-week units)

$$Y = 17 - 4.3X$$

Shift the origin to week number 30 and the unit to full weeks.

7. A study of car-lot shipments of onions into the state of Connecticut from 1917 to 1924 shows a straight-line trend which can be described as follows:

$$Y = 304.4 + 13.1X$$
Origin 1920

Interpret each of the figures in the equation. What would be the trend value of car-lot shipments in 1926?[1]

8. A student is learning to typewrite. He practices for a fixed period each day, and each day he counts the number of words typed. The numbers of words typed on various days, starting with his first attempt, are as follows:

Day Number	Number of Words	Day Number	Number of Words
1	72	7	108
2	86	8	110
3	93	9	112
4	99	10	113
5	103	11	114
6	106		

Selecting the values for the 1st, 6th, and 11th days, fit a parabolic trend to these data by the method of selected points. Compute the value of the trend for each of the days and see how well it fits the data. How do you explain the relative size of the trend value on the 10th and the 11th days? Project the trend ahead to the 12th and 13th days and determine the number of words which will be written if the trend continues. Would you expect the actual data for the 12th and 13th days to follow the trend? Estimate from the trend equation the number of words which will be written on the 25th day. (If the 6th day is the origin, this will be day +19.)

9. One page 298 is an equation for the trend of petroleum production fitted by the method of least squares. On page 279 is another equation

[1] Data from F. V. Waugh, Connecticut Market Demand for Vegetables, *Storrs Agricultural Experiment Station Bulletin* 138, p. 34.

which is the secular trend of the same data fitted by the method of selected points. In the former case the origin is 1923; in the latter case the origin has not been shifted, but remains at 0 A.D. Shift to the year 1923 the origin of the trend which was fitted by selected points, and compare the two trends when their origins are the same.

10. Compute petroleum production for 1906 as suggested in the footnote on page 279. Explain.

11. Plot a graph of the values of the equation

$$Y = 30 - 15X + 0.2X^2 + 2X^3$$

with the values of X running from -10 to $+10$. How many bends has the curve? Locate the curve at all integral values of X and also at the points halfway between; that is, at $X = -10$, $X = -9.5$, $X = -9.0$, $X = -8.5$, etc.

12. On page 284 is a table showing the computation of the moving average. The last two years are left vacant, for the reasons explained in the text. If, now, you were told that the petroleum output for 1930 was 1090 and, for 1931, 1100, what would be the moving-average figures for 1928 and 1929?

13. The arithmetic average of a group of values is fitted by the method of least squares, although this fact is not known to most people who compute it. Since it is so fitted, it must be true that the sum of the squares of the deviations from the mean is smaller than the sum of the squares of the deviations from any other value. Test this out. For example, the mean of the numbers 5, 7, 12, 2, and 4 is 6. Find the deviations of these values from 6, the squares of these deviations, and the sum of the squares. Compare this sum with the sum of the squares of the deviations from any number other than 6. Try 5 and 7, for example.

14. Plot the data of the table on page 294 on graph paper. Fit two or three freehand trends, all straight lines. Measure on the chart the deviations from the trend, and compute for each line the sum of the squared deviations from the trend. Which line is the best fitting of the lines by the least-squares criterion? Fit a line to these data by the method of least squares. Compute the deviations from the trend by means of the trend equation. Find the sum of the squared deviations and compare it with the others found above.

15. Figure 10.17 shows a straight-line trend. Compute its formula by the method of selected points.

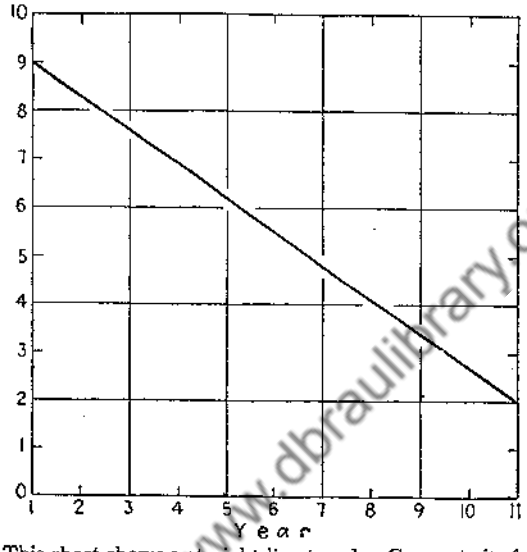


FIG. 10.17.—This chart shows a straight-line trend. Compute its formula by the method of selected points.

16. Fit a curvilinear trend to the following data, deciding first what kind of curve should be used:

X	Y
1860	104
1870	117
1880	132
1890	152
1900	179
1910	218
1920	279
1930	338
1940	633

17. In the preceding exercise, estimate the value of Y in 1933, using your trend equation as the basis of estimate.

18. A geometric trend becomes a straight line on semilogarithmic paper because the horizontal lines, instead of being evenly spaced, are spaced at distances proportional to the logarithms of the numbers. Make a piece of cross-section paper which will straighten out a reciprocal curve. It is merely necessary to space the horizontal lines at distances proportional to the reciprocals of the numbers.

19. Look up monthly figures for mean temperatures at some weather station. (The "World Almanac" gives such figures for New York City if your local data are not available.) These figures will contain a 12-month cycle. Using data for about five years, compute the 12-month moving average, centering by means of a two-month moving average.

20. The statement that the method of least squares gives the line of "best fit," although often made in this bald form, really needs several qualifications before it is correct. What are the qualifications?

21. Fit a straight line to the data of Table 10.10, page 294, using the method of least squares.

22. Using the trend equation computed in the preceding exercise, find the trend value for each of the 11 years of Table 10.10. Compare the trend values with the actual values to find the "errors" or "residuals." Square the residuals, and find the sum of the squares. The sum of the squares should be smaller than that obtained in Table 10.10, since you have used the method of *least* squares.

23. Suppose we have fitted a trend line by some method to figures showing the number of bales of cotton produced in a given county. Our figures are given every decade. In our table and our computations, then, X has represented the number of decades before or after the origin, and Y has represented the number of bales of cotton produced. Our trend equation computed on this basis is

$$Y = 759 + 23X$$
Origin 1870

Imagine, now, that you want the trend equation with the origin in 1900 instead of in 1870, you want Y to represent the number of pounds of cotton instead of the number of bales, and you want X to represent the number of years instead of the number of decades from the origin. A bale of cotton weighs 500 lb. Write the new trend equation.

24. Fit the appropriate curvilinear trend by the method of least squares to the data of Table 10.8. Write out the trend equation.

25. Convert the trend equation of Exercise 24 to the nonlogarithmic form described in the footnote on page 305.

26. Interpret each of the numbers in your trend equation found in Exercise 25.

27. In Table 10.14 we note that the U.S. population grew from 3.9 million to 23.2 million between 1790 and 1850. This was a growth of 19.3 million. If we start with 3.9 million and increase 19.3 million, we have increased 495 per cent. If we increase 395 per cent in 6 decades, we have increased 82½ per cent per decade. Yet our trend equation (see page 307, footnote) tells us that we have increased at the rate of 34 per cent per decade. How do you reconcile the two conclusions?

CHAPTER XI

HISTORICAL DATA—CYCLICAL MOVEMENTS

11.1. The Nature of Cyclical Movements.—In the preceding chapter we have studied movements that continued to act in the same way for a considerable period of time. Cyclical movements differ from secular movements in that the former go through a given routine and then repeat it over and over again. Repetition is the essence of cyclical movement.

Yet in most actual cases the repetitions are not exact. In Table 11.1 we list the monthly mean temperatures in New York City for a period of five years. It is immediately evident that there are seasonal regularities, with high temperatures in July

TABLE 11.1.—MONTHLY MEAN TEMPERATURES, NEW YORK CITY, 1935–1939[1]

Month	1935	1936	1937	1938	1939
January	29.2	29.9	40.4	32.0	32.3
February	31.6	26.6	34.9	35.6	37.4
March	43.2	45.3	36.6	44.2	38.8
April	49.5	47.2	49.0	53.4	47.8
May	58.8	62.6	63.3	59.4	63.7
June	68.6	68.6	70.6	69.0	70.8
July	76.2	74.8	75.4	75.1	74.1
August	73.6	74.1	75.7	76.3	76.8
September	64.2	67.1	65.2	64.9	67.4
October	56.8	57.0	54.6	58.6	56.4
November	48.6	42.4	45.6	47.7	43.2
December	30.6	39.2	35.4	37.2	36.2

[1] Data from "World Almanac," p. 187, 1941.

and August and low temperatures in January. Yet inspection of the data will show that exact repetition even in a single month hardly ever occurs. The only case in the five years where any month had the same monthly mean temperature twice is that of June, 1935 and 1936. It is approximate repetition, and not exact repetition, that we look for in cyclical data.

From the name we are apt to think of a cycle as a rounded, wavelike movement, similar, perhaps, to that shown in Fig. 10.3. Sometimes, especially in the "exact" sciences, we come across data which exhibit such symmetrical regularity. Figure 11.1 shows two complete cycles of this character.[1] When we wish to describe the cycle, we may wish to tell how long a typical cycle lasts. This would be the time elapsing between any point on one cycle and the corresponding point on the next cycle, but since it is hard at most parts of the cycle to say which points "correspond," it is common to measure from one peak to another

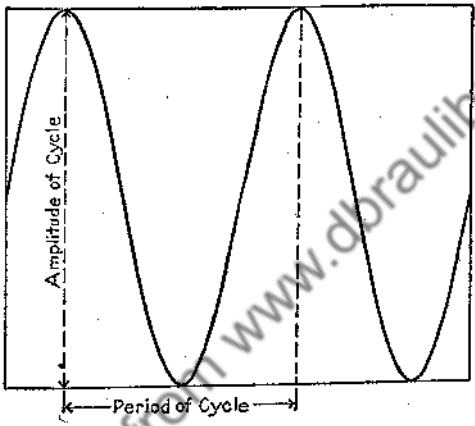


FIG. 11.1.—Two cycles of a sine curve.

or from one trough to another. The distance along our base scale marked with arrows in Fig. 11.1 measures the time from one peak to another. The time that is required for one complete cycle, measured as the horizontal distance from any point in one cycle to the corresponding point in the next cycle, is called the *period* of the cycle. Since it is a measurement on the horizontal base scale, the period of a cycle is always a length of time, such as a year or a month or 5 min.

Cycles differ not only in period, but also in the extent of the "up-and-down" movement—the height of the peaks and the depth of the troughs. This vertical distance, represented by the length of the broken vertical lines in Fig. 11.1, is called the *amplitude* of the cycle. The amplitude is measured at right angles to the base scale, vertically, and therefore it is always in

[1] The curve in the figure is a sine curve.

units of the non-time variable which is being measured. For example, if we were referring to the temperature cycles of Table 11.1, the period of the cycle would be a length of time (12 months in this case), and the amplitude of the cycle would be a number of degrees of temperature (about 43.5° in this case if we take the cycle very roughly as running from a low of 32° in January to a high of 75.5° in August). The usefulness of these two figures—period and amplitude—is evident at once. For example, weather data show that the amplitudes of annual temperature fluctuations in various United States cities are roughly as follows:

Boston	44°
Charleston	31°
Chicago	48°
Miami	15°
Los Angeles	16°
Bismarck	62°

In each of these cases the period of the cycle is 12 months. But the difference in amplitude between Bismarck, North Dakota, and Miami, Florida, is startling.

It is perhaps easiest to illustrate the ideas of period and amplitude with the sinuous, regular, rounded cycles of Fig. 11.1; but the student must not get the idea that all cycles are of this character. It is repetition at approximately equal time intervals and not smooth, flowing regularity which makes a cycle.[1] For example, the sales in a chain grocery store may run along at an approximate level from Monday to Friday, increase greatly on Saturday, and disappear entirely on Sunday. This would be a weekly cycle, even though when plotted it showed none of the wave motion of Fig. 11.1, but looked more like Fig. 11.2. Similarly the annual cycle of sales by a department store might show sudden very sharp increases just before Christmas and Easter, with sales disappearing entirely on Sundays and holidays.

11.2. Common Periods of Cycles.—Cycles of any period can occur, but in practice certain lengths of period are much more

[1] The word "cycle" and the word "circle" come from the same root, and perhaps it would have been better to have confined the idea of cycles to those circular functions which do exhibit the roundness and regularity which the student has come across in his study of trigonometry or the calculus. Usage in the field of statistics, however, justifies the definition above.

common than others for reasons which it is easy to understand. Perhaps the commonest cycle is the annual or seasonal cycle, which is astronomical in origin, but which shows up in the data of almost every science. We have already noticed the annual fluctuations in temperature, but a moment's thought will suggest similar annual cycles in the growth of vegetation,[1] the rates of metabolism among animals, school attendance, volumes of traffic, production of farm products, birth and death rates, etc.

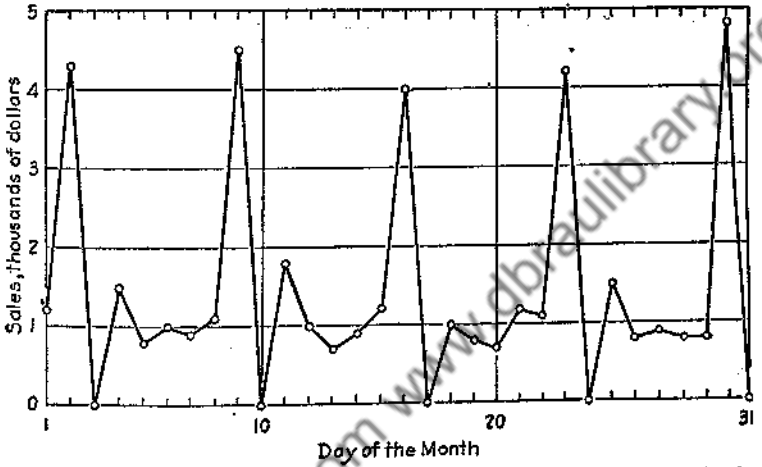


FIG. 11.2.—Daily sales in H. D. Newton's grocery store for the month of March, 1942.

The annual cycle is of importance not only in astronomy, but also in all the biological and the social sciences, where the seasonal changes have important secondary effects. The annual cycle is of less importance, perhaps, in the physical sciences. When a cycle lasts 12 months we call it a *seasonal movement*, so we can say that a seasonal movement is one particular kind of cyclical movement—the one with a 12-month period.

In many sciences there are also important daily or *diurnal cycles*. Some of the phenomena which we have just mentioned as exhibiting seasonal movements also exhibit diurnal cycles.

[1] It is this annual cycle, of course, which produces the "rings" in trees by means of which we ascertain their age. Recent studies of the tree-ring cycles have made it possible to find the dates at which timbers were cut for hundreds of years in the past, and to learn something of weather cycles at times long before weather records were kept.

For example, the temperature of the air, the volume of traffic, birth and death rates, etc., vary from one hour of the day to another. In the field of medicine it is well known that various diseases show typical diurnal cycles in the patient's temperature. A student's mental alertness and his rate of learning vary diurnally. Marketing studies show that almost every kind of retail store has its busy hours of the day and its slack hours, varying for different kinds of stores, to be sure, but exhibiting typical diurnal cycles for any particular type.

Although the week seems to be a far more arbitrary time unit than the day or the year, having far less basis in natural phenomena, nevertheless the week has become firmly enough fixed as part of our lives so that *weekly cycles* are not uncommon, in the social sciences particularly. There are marked weekly cycles in the sales and prices of many perishable farm products. In heavily settled parts of the country there are decided weekly cycles in highway traffic. Certain days of the week are days of large sales in department stores, and other days are days of little business. Studies by the personnel departments of large corporations show what is at first a surprising weekly cycle in the number of employees absent for sickness, some days of the week being chosen for such absences far more commonly than others. (Strange to say, it is just before, and not just after, the week end that these absences are most common.) Even death rates show a weekly cycle, especially in summer when automobile accidents and drownings make their mark. The week-end holiday has become an integral part of our lives, and its effects show up as a weekly cycle wherever men's habits come into play.

These three periods—12 months, 1 week, and 1 day—are by far the most common periods when we look at the problem of cycles in general. In particular cases we may have cycles that last far longer, such as the sunspot cycle of just over 11 years; or cases which fall in between, such as the typical 3-day cycle of tertian malaria. A cycle is just as important, of course, whether it coincides with one of the three common periods or not. The student who is testing data for cyclical movements will do well, however, to look for these periods first. And, of course, it is evident from the examples just given that several periods of cycles may be mixed together in the same data, as in the case of highway traffic outside a large city at 5 o'clock

in the afternoon of a Sunday in the late summer, when we have the coincidence of diurnal, weekly, and annual peaks.

11.3. Preliminary Adjustment of Cyclical Data.—If we are to get a clear picture of the cycle in our data, it is helpful to eliminate as far as possible other sorts of historical movement. For this reason it is common, before studying the cycle in data, to measure and eliminate the trend by means of the methods explained in the preceding chapter. Thus our data for study of cycles are often in the form of residuals from the trend (see Secs. 10.18 and 10.19).

It is also often necessary, especially in the cases of cycles with periods of a year or less, to make allowances for calendar variation (see Sec. 10.2). For some sorts of data it is merely necessary to make allowance for the differences in the lengths of months, while for others it may be necessary to correct for the number and position of holidays, the numbers of Saturdays and Sundays, etc. Only one who has a good understanding of the forces affecting the data in question is qualified to determine what sorts of calendar corrections need to be made.

Where the period covered is a long one, it may be necessary also to make allowances for changes in population, for changes in the method of collecting original data, for changes in the definition of statistical units, etc. The longer the period covered, the more likely it is that some sort of allowance will have to be made for changes in the basic picture.

11.4. Seasonal Variation Measured around the Moving Average.—We shall start our discussion of cyclical movements by the analysis of a case of seasonal variation, since that is perhaps the commonest of all lengths of cycle. Just as is true with the secular trend, we may be interested in the nature of the cyclical movement itself, or we may wish to measure it so that we can eliminate it and study those movements which remain. We shall start by describing a seasonal movement, and then we shall eliminate it.

Table 10.4, page 273, shows monthly egg prices in New York City from 1919 through 1923. These data are shown graphically in Fig. 10.2, page 274. The most noticeable feature of this chart is the fact that there are decidedly regular periodic swings in the data. The seasonal movement dominates the whole chart. The movements are not uniform from year to year, to

be sure; there are variations in the amplitude of the waves. But the similarities of the successive yearly movements are much more striking than are the differences. If one marks the crest of each wave or the trough of each wave, he finds that there are 12 months from crest to crest or from trough to trough. This regular 12-month period is characteristic of the movements which we call "seasonal" movements. Other cyclical movements are characterized by periods of other lengths. The peculiarities of the present movements become even more evident if we plot the prices of the various years one above the other, shifting the vertical scale of the diagram upward each year so that the years will lie in order. The chart shown in Fig. 11.3 has been

TABLE 11.2.—AVERAGE MONTHLY PRICE OF NEAR-BY-HENNERY WHITE EGGS, NEW YORK CITY, 1919–1923

Month	Price (cents per dozen)
January	69.0
February	54.8
March	46.4
April	44.2
May	43.4
June	47.0
July	53.6
August	60.8
September	71.2
October	86.6
November	93.4
December	77.8

constructed in this way. The appearance of each year above the preceding year is not due to the fact that prices were higher, but to the fact that the vertical scale has been shifted. When we look at this chart, we note again the striking similarity between the movements of the various years.

If one were sure that there were no secular trend or long-time cycle in the data, one could describe the seasonal movement easily by computing the average January egg price, the average February egg price, etc. In this way he would get an average price for each month, as in Table 11.2. These figures make the

seasonal movement very clear. They point out the times of high and the times of low prices. If we wished, we could convert

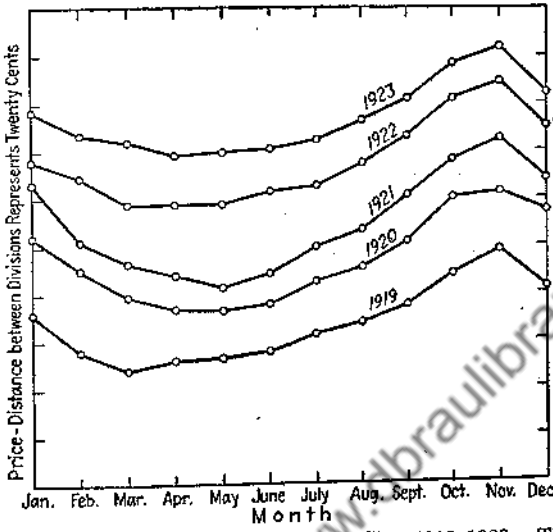


FIG. 11.3.—Monthly egg prices in New York City, 1919–1923. The data for the various years are all on the same vertical scale, but each year the scale has been raised enough so that the cycles will not overlap. No two years are on the same base line.

these figures into an index of seasonal variation by computing the average of the 12 monthly averages (which turns out to be 62.35) and stating each of the 12 monthly average prices as a

TABLE 11.3.—INDEX OF SEASONAL VARIATION IN EGG PRICES

Month	Seasonal Index
January	111.0
February	88.0
March	74.4
April	71.0
May	69.6
June	75.4
July	86.0
August	97.6
September	114.3
October	139.0
November	149.8
December	125.0

percentage of their average. This computation would give us Table 11.3. This table tells us that the January prices were, on the average, 11 per cent above the yearly average price; February prices were 12 per cent below the average of the year; etc.

This method of describing the seasonal variation would not give the correct results, however, if the data included either a secular trend or a cycle other than the 12-month cycle. Hence the statistician commonly uses a method which is slightly more complex. It consists in finding the 12-month moving average of the data first, and in removing this moving average. Since the moving average is a 12-month one, it will have in it nothing of the seasonal movement; the monthly variations will be entirely ironed out. But the trend and cycles other than 12-month cycles will remain in the moving average. When, therefore, we remove the moving average from the original data, we shall be removing the trend and other cycles but not the seasonal variation which we wish to study.[1]

Let us try this plan without data on egg prices. We first tabulate the original data, and we then compute the 12-month moving average. We have discovered that a moving average is placed at the center of the period, and this would make us place the moving average halfway between June and July in the first year. We could adjust our moving average so that it would properly be placed at July instead of halfway between June and July,[2] but since such an adjustment ordinarily makes no significant difference in the final results, we shall omit it here and center each year's moving average at the seventh month. Also, in order to save time, we may omit the division of each 12-month total by 12; that is, we may use the moving total rather than the moving average. This makes no difference whatever in the result, for a reason that will soon be evident.

[1] If we were studying some cycle other than a seasonal one, we should, of course, use a different moving average. If, for example, a study of the graphed data led us to believe that there was a 14-year cycle, we should compute the 14-year moving average, etc.

[2] Since the first figure for the moving average would represent a point halfway between June and July, and the second a point halfway between July and August, the average of these first two figures would represent July. Thus we could take the 12-month moving average and then take a 2-month moving average of the result, centering on the seventh month.

Table 11.4 gives for each month the price of eggs (as shown in Table 10.4, page 273), the moving total of prices centered at the seventh month, and the percentage which the former is of the latter. Each figure in column 2 is the price of eggs for the indicated month. Each figure in column 3 is the sum of the price in the indicated month and the prices in the six months before and the five months after; it is the moving total of 12 months' prices centered at the seventh month. Each figure in column 4 is the percentage which the corresponding figure in column 2 is of the corresponding figure in column 3. If we take as an example the month of July, 1919, we find that the price of eggs was 63 cents, the sum of the July price and the prices of the six months before and the five months after is 811 cents, and 63 cents is 7.76 per cent of 811 cents.

We have seen that one of the difficulties of the moving average is that it cannot be extended to the extremes of the data (see page 291). In this case it means that we have no moving total for the first six or the last five months; the last two columns are therefore vacant for these months. The moving average has the advantage, however, that it is flexible; and while it does in this case eliminate any and all regular 12-month movements, it does not eliminate the trend or cyclical movements. These are still contained in the moving average, and when we remove the moving average they are removed with it.

Had we computed the moving average rather than the moving total, each figure in column 3 would have been divided by 12. Since each figure in column 3 would then be $\frac{1}{12}$ as large, each figure in column 4 would be 12 times as large. But the relative sizes of the figures in this column would be unchanged. We are interested in the relative sizes of these figures and not in their absolute size. For this reason we save time by omitting the division by 12 in column 3—that is, by using the moving total rather than the moving average.

Note the meaning of the figures in our last column. Each is stated as a percentage of the moving total. If a figure is large, the price in that month was high as compared with the moving total. Thus we have eliminated the moving total (this corresponds, as we have seen, to removing the moving average) from our data. The figures in the last column retain whatever seasonal

TABLE 11.4.—ELIMINATION OF MOVING TOTAL FROM EGG PRICES

(1) Month	(2) Price (cents)	(3) Moving Total	(4) Per Cent [(2)/(3)]
1919			
January	72		
February	56		
March	48		
April	52		
May	53		
June	56		
July	63	811	7.76
August	68	823	8.26
September	75	837	8.97
October	88	848	10.37
November	98	850	11.51
December	82	850	9.65
1920			
January	84	850	9.88
February	70	852	8.21
March	59	855	6.90
April	54	862	6.25
May	53	874	6.07
June	56	878	6.38
July	65	891	7.29
August	71	883	8.05
September	82	865	9.47
October	100	849	11.78
November	102	833	12.27
December	95	813	11.69
1921			
January	76	796	9.55
February	52	781	6.65
March	43	767	5.61
April	38	756	5.02
May	33	742	4.45
June	39	735	5.30
July	50	718	6.96
August	57	698	8.17
September	71	695	10.20
October	86	690	12.45
November	95	690	13.77
December	78	695	11.21
1922			
January	56	699	8.01
February	49	694	7.08
March	38	692	5.50
April	38	687	5.54
May	38	683	5.57
June	43	677	6.35
July	45	669	6.75
August	55	670	8.20
September	66	668	9.89
October	82	674	12.18
November	89	675	13.19
December	70	677	10.31
1923			
January	57	675	8.45
February	47	675	6.97
March	44	673	6.55
April	39	669	5.84
May	40	664	6.04
June	41	658	6.23
July	45	652	6.90
August	53		
September	62		
October	77		
November	83		
December	64		

and random movements were present in the original data, but do not include the secular and long-time cyclical movements.

TABLE 11.5.—ARRAYS OF MONTHLY DEVIATIONS

Month	Results				
January		8.01	8.45	9.55	9.88
February		6.65	6.97	7.08	8.21
March		5.50	5.61	6.55	6.90
April		5.02	5.54	5.84	6.25
May		4.45	5.57	6.04	6.07
June		5.30	6.23	6.35	6.38
July	6.75	6.90	6.96	7.29	7.76
August		8.05	8.17	8.20	8.26
September		8.97	9.47	9.89	10.20
October		10.37	11.78	12.18	12.45
November		11.51	12.27	13.19	13.77
December		9.65	10.31	11.21	11.69

Let us now gather together the results for each month for purposes of comparison. For each month other than July we shall have four figures, and for July we shall have five. In assembling them let us arrange the results for each month in order of size—in an array. The results are given in Table 11.5.

We could, of course, take the mean of these figures for each month, but with so few figures for each month this would be likely to give undue weight to extreme items. It is, therefore, more common to take the median. The median will be halfway between the second and third figure (save in the case of July, when it will be the third figure). The medians, then, will be these:

January	9.00	July	6.96
February	7.02	August	8.18
March	6.08	September	9.68
April	5.69	October	11.98
May	5.80	November	12.73
June	6.29	December	10.76

These figures show the median per cent which the price for each month was of the moving total. We can easily make an index of seasonal variation from these figures by computing the average for the year and finding what per cent the figure for each month

is of this average. The average of the 12 figures just given is 8.35. If we state the figure for each month as a percentage of 8.35, we get the index of seasonal variation shown in Table 11.6.

TABLE 11.6.—INDEX OF SEASONAL VARIATION IN EGG PRICES

Month	Index
January	107.8
February	84.0
March	72.9
April	68.1
May	69.5
June	75.3
July	83.4
August	98.0
September	115.9
October	143.5
November	152.4
December	128.7

If this index of seasonal variation is compared with that which was computed by the simpler method on page 331 it will be seen

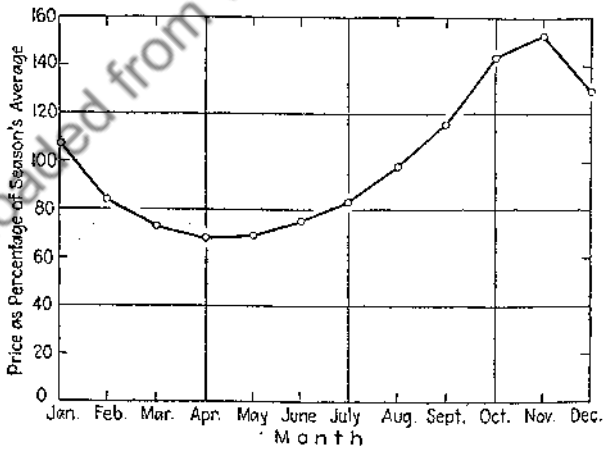


FIG. 11.4.—Seasonal movement of egg prices. Data from Table 11.6

that, although the general nature of the seasonal movement is the same by either method, there are, nevertheless, some fairly sizable differences in results. The method just outlined is, for the reasons already noted, the method to be preferred.

These figures give us a good picture of the seasonal movement itself. They tell us that egg prices tend to be at their peak in November and "reach bottom" in April. They tell us, moreover, that at the peak the prices tend to be over 50 per cent above the season's average, and that in the trough they fall to a point more than 30 per cent below the season's average. We can find the times of high prices and the times of low prices, and we can also get some idea of the amplitude of the movement. The highest prices tend, on the average, to be more than double the lowest prices (see Fig. 11.4).

To summarize the steps necessary for finding an index of seasonal variation based on the moving average, we have the following:

1. Tabulate the original data.
2. Compute a 12-month moving total centered at the seventh month.
3. Divide each original entry by the corresponding moving total. State the result as a percentage.
4. Sort out these percentages by months, and find the median percentage for each month.
5. Express each of these monthly medians as a percentage of the average of all 12 monthly medians. This is the index of seasonal variation.

11.5. Seasonal Variation by Link Relatives.—Another common method of measuring seasonal variation is by means of what are called *link relatives*. Whenever we have a set of figures for a number of months (or other time periods), we find the link relative for any month by dividing the figure for that month by the figure for the preceding month and multiplying the quotient by 100. In other words, a link relative for any month is the percentage which the value that month is of the value in the preceding month. We shall illustrate with the same egg prices with which we have just been working. The original figures appear in Table 11.4. They have been converted to link relatives in Table 11.7. Since a link relative for any month involves comparison with the preceding month, we do not know the link relative for our first month, January, 1919. Each of the other link relatives is found by the method just described. For example, if we want the link relative for September, 1923, we note (see Table 11.4) that the price of eggs in September, 1923, was 62 cents, while in the preceding month the price was 53 cents. Our computation is, then

$$\frac{(62)(100)}{53} = 117.0$$

This answer shows that the price of eggs in September, 1923, was 17 per cent above the price of the preceding month. Whenever prices are rising, the link relative will exceed 100. When the price is falling, the link relative will fall short of 100.

Inspection of Table 11.7 will show that there are some months (such as September) in which the price is almost always higher than it was in the preceding month, while there are other months (such as February) when the price is almost always below that of the preceding month. If we want some idea of the typical situation in any month, it is natural for us to take some sort of average of the link relatives for that month. We might, of course, take the arithmetic mean of the link relatives, but, as

TABLE 11.7.—LINK RELATIVES OF EGG PRICES IN NEW YORK CITY, 1919–1923

Month	1919	1920	1921	1922	1923
January	.....	102.4	80.0	71.8	81.4
February	77.8	83.3	68.4	87.5	82.5
March	85.7	84.3	82.7	77.6	93.6
April	108.3	91.5	88.4	100.0	88.6
May	101.9	98.1	86.8	100.0	102.6
June	105.7	105.7	118.2	113.2	102.5
July	112.5	116.1	128.2	104.7	109.8
August	107.9	109.2	114.0	122.2	117.8
September	110.3	115.5	124.6	120.0	117.0
October	117.3	122.0	121.1	124.2	124.2
November	111.4	102.0	110.5	108.5	107.8
December	83.7	93.1	82.1	78.7	77.1

we pointed out in the preceding section, when there are so few figures one or two erratic values will throw the arithmetic mean very far off. For that reason we take the median link relative for each month. These medians are

January	80.7	July	112.5
February	82.5	August	114.0
March	84.3	September	117.0
April	91.5	October	122.0
May	100.0	November	108.5
June	105.7	December	82.1

We now set up what are called *chain relatives* for the various months. This is done by setting the first month arbitrarily equal to 100.0, and determining the chain relative for any other month by multiplying the median link relative of that month by the chain relative of the preceding month. This gives us the following chain relatives:

January	100.0	July	75.6
February	82.5	August	86.2
March	69.5	September	100.9
April	63.6	October	123.1
May	63.6	November	133.6
June	67.2	December	109.7
		January	88.5

It will be noticed that we have carried the computation clear around to include January for a second time, this second figure for January being obtained, like any other chain relative, by multiplying the month's link relative (80.7) by the chain relative for the preceding month (109.7). If it were not for such things as the influence of trend, the rounding off of numbers, and the fact that we have used the median link relative rather than the arithmetic mean, we should have come back to our original 100.0 with our second January. But things seldom work themselves out so smoothly, and we are therefore usually obliged, as we are here, to make some adjustment. Our final figure, 88.5, is too low by 11.5 per cent. We shall add $\frac{1}{12}$th of the discrepancy (0.9583 per cent) to February, $\frac{2}{12}$ of the discrepancy (1.9166 per cent) to March, etc. This gives us the following index, adjusted for trend:

January	100.0	May	67.4	September	108.6
February	83.5	June	72.0	October	131.7
March	71.4	July	81.3	November	143.2
April	66.5	August	92.9	December	120.2

If the second January chain relative had been larger than 100, it would have been necessary to reduce the various months by their proportionate amounts, just as we increased them in this case because the second January was too small.

It is now common, as a last step, to center the index of seasonal variation, so that the averages of the monthly indexes will be 100. This is done by dividing each of the crude indexes in the preceding table by the average of all 12 monthly indexes. The

average of the 12 indexes in the preceding table is 94.9, and if we divide each of the 12 crude indexes by 94.9 and then multiply by 100, we get the following final index of seasonal variation based on link relatives:

January	105.4	May	71.0	September	113.4
February	88.0	June	75.9	October	137.7
March	75.2	July	85.7	November	150.9
April	70.1	August	97.9	December	116.1

The student will wish to compare the results obtained by this method with those given in Table 11.6, page 336, which we obtained by the moving-average method. While the two sets of figures are by no means identical, nevertheless they do show very definitely the same general sort of seasonal movement.

Perhaps it would be wise to summarize the steps necessary in the link relative method, since the process is not so difficult as it may seem when the illustration has been run through so many pages. The steps involved are these:

1. Convert the original data into link relatives, by dividing each entry by the one which precedes it and multiplying the quotient by 100.

2. Sort out the link relatives by months, and compute the median (or arithmetic mean) link relative for each month.

3. Compute a set of chain relatives, by setting the first chain relative equal to 100, and finding each other chain relative by multiplying the link relative for the period by the chain relative for the next preceding period. Carry this process through to include the first unit of the next period (that is, when dealing with monthly data, carry it through to include the next January).

4. If the last chain relative computed in the preceding step is not 100, adjust for trend by adding or subtracting a correction factor. If the final chain relative is larger than 100, the correction factor is to be subtracted. If the final chain relative is smaller than 100, the correction factor is to be added. The first month is kept at 100, but the correction factor for the next month is $\frac{1}{12}$ the amount by which the last chain relative differs from 100, the correction factor for the third month is $\frac{2}{12}$ of this amount, then $\frac{3}{12}$, $\frac{4}{12}$, etc. When working with other than monthly data, we can work as follows. Let d be the difference between the last chain relative and 100. Let s be the number of subdivisions in our period. (Above it was 12 since there were 12 months in our period. For weekly cycles s might be 7, etc.) Then the correction factors for the succeeding subdivisions are d/s, $2d/s$, $3d/s$, . . . , sd/s.

5. Bring the final index to the level of 100 by finding the average of the indexes adjusted for trend in step 4, and then dividing each of these indexes by their average.

The chain relative method is usually faster than the method based on the moving average, and the results are usually reasonably similar. There is little theoretical advantage of either system over the other for ordinary cases.

11.6. The Elimination of Seasonal Movements.—So much for the methods that are used to describe seasonal movements. Let us now turn our attention to the problem of removing the seasonal movement so that we can study the remaining characteristics of the data without having them obscured by the seasonal swings. As we saw in Fig. 10.2, page 274, the seasonal movements in the prices of eggs are so pronounced that they hide the other movements almost completely. In eliminating this seasonal swing, we shall use the index of seasonal variation based on the moving average, which is tabulated on page 336.

The simplest way to eliminate the seasonal movement is to divide the actual price for each month by the index of seasonal variation. This index is really a percentage, and the January index of 107.8 can therefore be thought of as 107.8 per cent, or 1.078. We should divide all January figures by 1.078 to make them somewhat smaller. The January figures are all too large by 7.8 per cent because of the time of the season. They should be reduced 7.8 per cent to be comparable with the figures for the other months. Similarly the prices are always low in June because of the time of year, and if a price is 75.3 per cent of normal it is just where it belongs. If, then, we divide it by 0.753, we make it comparable with the prices of the other months. If we divide each month's price by the seasonal index for the corre-

Month	1919	1920	1921	1922	1923
January	66.9	78.0	70.6	52.0	52.9
February	66.6	83.4	61.9	58.4	56.0
March	65.8	81.0	59.0	52.1	60.4
April	76.4	79.3	55.8	55.8	57.3
May	76.3	76.3	47.5	54.6	57.5
June	74.4	74.4	51.7	57.1	54.4
July	75.6	78.0	60.0	54.0	54.0
August	69.4	72.5	58.2	56.1	54.1
September	64.9	70.9	61.4	57.0	53.6
October	61.4	69.8	60.0	57.2	53.7
November	64.3	66.9	62.4	58.4	54.5
December	63.8	73.9	60.7	54.5	49.8

sponding month, we get the prices *corrected for seasonal variation* as shown in the table at bottom of page 341.

When these prices are plotted, we discover that the seasonal fluctuations have been entirely eliminated but that the secular and random movements are still present. In fact the latter stand out much more clearly now than before. The prices with seasonal eliminated are shown in Fig. 11.5, which should be compared with the chart on page 274 showing the prices in their original form. Points on the chart in Fig. 11.5 which show high

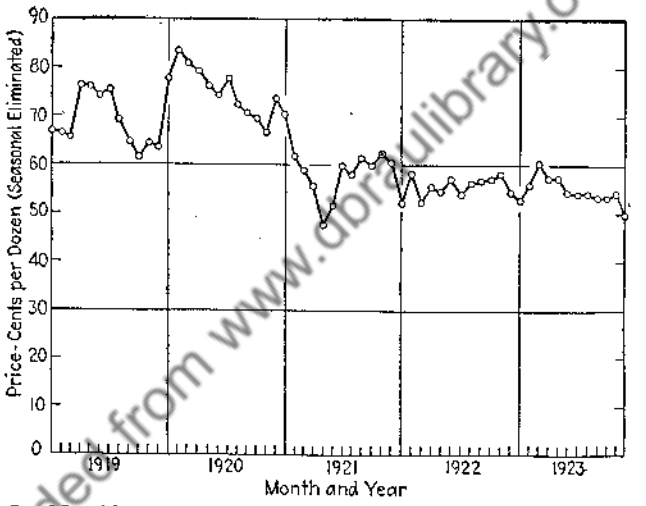


FIG. 11.5.—Monthly egg prices, 1919–1923, with seasonal eliminated. Data from page 341.

prices mean that the prices were high for the time of year in question. A price which is "high" for April might be a "low" price for November. On this chart we have adjusted all prices for the seasonal variation which usually occurs, and the variations which are left are variations from the usual seasonal position. It is common to speak of such prices as prices which have been *corrected for seasonal variation*, or to speak of them merely as prices *with seasonal eliminated*.

To summarize, we eliminate the seasonal from data after the index of seasonal variation has been found by dividing the original data for any month by the seasonal index of that month (remembering that the seasonal index is a percentage and pointing off two places accordingly).

11.7. Random Movements.—We have now found how to describe and to remove both secular and cyclical movements. If we have done the work accurately, there should be no movements left in the data which are coordinated with the passage of time. To be sure, we shall not have removed all variations from the data. There will still be movements reflecting the differences in the quality of eggs, or the quantity shipped to market, or changes in consumer tastes, or changes in the prices of substitute commodities, etc. We have not tried to eliminate all the movement in egg prices, but merely those which show some temporal regularity. The movements which are left should show the effects of changes in nontemporal forces.

Inspection of Fig. 11.5 shows that we have eliminated fairly well the seasonal swings, but we still have the secular trend left. The chart shows a rather definite and fairly linear downward trend of prices. We could fit a straight line to the data of Fig. 11.5 by the method of least squares, but let us eliminate the trend by the easier method of a freehand line. If the student will stretch a string or hair, or lay a transparent ruler over the figure, shifting it until it shows the general direction of the trend, he will see that the trend line crosses the left-hand vertical axis at a point which represents approximately 70 cents, and crosses the right-hand axis at approximately 52 cents. Thus the drop for the entire period is 18 cents. A drop of 18 cents in 60 months is a drop of 0.3 cents per month. We note, then, that our figures are too high by 9 cents in January, 1919. We subtract 9 cents for that month, 0.3 of a cent less than 9 cents for the next month, 2(0.3) of a cent less than 9 cents for the following month, etc. In other words, the amounts that we subtract from the figures in the table on page 341 for successive months starting with the first one are 9.0, 8.7, 8.4, 8.1, 7.8, etc. After we pass the middle of the period, we shall be adding first small fractions of a cent and then more and more until at the end we add 9 cents. This will give us the data of Table 11.8, which contains egg prices with both secular and seasonal movements eliminated.

When we select any figure from Table 11.8, say the figure 70.5 cents for June, 1920, we have to realize that this does not mean that the price of eggs in June, 1920, was 70.5 cents. As a matter of fact, reference to Table 11.4 will show that the actual price was 56 cents. Table 11.8 tells us that, according to our best estimate,

the price in June, 1920, would have been 70.5 cents if the secular and seasonal movements had not been present. June egg prices are only about 75.3 per cent of the annual average on account of the regular seasonal swing (see index of seasonal variation, Table 11.6), so eliminating the seasonal we estimate a price of 56/0.753, or 74.4 cents. But the secular trend is high in June, 1920, and requires a reduction of 3.9 cents, giving us a price, corrected for

TABLE 11.8.—MONTHLY EGG PRICES, 1919–1923, WITH BOTH SECULAR AND SEASONAL MOVEMENTS ELIMINATED

Month	1919	1920	1921	1922	1923
January	57.9	72.6	68.8	53.8	58.3
February	57.9	78.3	60.4	60.5	61.7
March	57.4	76.2	57.8	54.5	66.4
April	68.3	74.8	54.9	58.5	63.6
May	68.5	72.1	46.9	57.6	64.1
June	66.9	70.5	51.4	60.4	61.3
July	68.4	74.4	60.0	57.6	61.2
August	62.5	69.2	58.5	60.0	61.6
September	58.3	67.9	62.0	61.2	61.4
October	55.1	67.1	60.9	61.7	61.8
November	58.3	64.5	63.6	63.2	62.9
December	58.1	71.8	62.2	59.6	58.5

both trend and seasonal, of 70.5 cents as we saw in Table 11.8. The figures of this table are charted in Fig. 11.6. Although there is a noticeable similarity between Figs. 11.6 and 11.5, we note at once that the secular movement has been removed in the new chart.

Suppose, now, that a research man is interested in finding what relationship there is between the quality of eggs and their price. He goes out on the market month after month and candles the eggs to find their quality. He notices the prices at which they sell. When he discovers that the price in November, 1920, was $1.02 per dozen, while in May of the same year it had been only 53 cents per dozen (figures from Table 11.4), he might conclude that the higher November price reflected higher quality of eggs. But when he looks at the figures of Table 11.8, he discovers corrected prices of 64.5 cents for November and 72.1 cents for May, 1920. The May prices were actually considerably higher than the November prices *after allowance is made for time*

factors. The corrected figures of Table 11.8 should be far more useful to this investigator than the actual prices shown in Table 11.4. This is a good example of a case where one has studied the secular movement and the seasonal movement, not because he is interested in them per se, but because he wants to eliminate them and study the relationship between other factors (like quality) and the random or residual movements which are left. The random movements of Fig. 11.6 are not at all noticeable in Fig. 10.2, page 274, which show the actual prices. The secular and

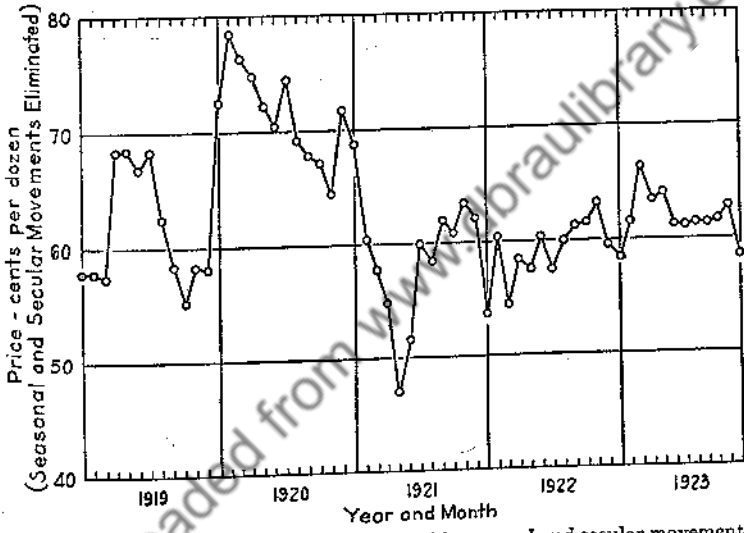


FIG. 11.6.—Monthly egg prices, 1919–1923, with seasonal and secular movements eliminated. Data from Table 11.8.

especially the seasonal movements in that chart obscure everything else. But Fig. 11.6 shows immediately and strikingly the movements in the original data which were not arranged in some temporal pattern—which were not explainable in terms of the passage of time.

11.8. The Concept of the Statistical Normal.—Having described and "eliminated" the secular and cyclical movements, we may now wish to reconstruct them, omitting the random movements. Such a reconstructed idealized series, made up of the secular and cyclical movements, but omitting the random movements, may be thought of as showing the changes that we might have expected

to have occurred under "normal" conditions, in the absence of temporary and sporadic forces.

Suppose, for example, that you are asked, "What number of automobiles might one expect to pass the junction of U.S. Route 11 and U.S. Route 20 in an hour?" Your answer would have to depend on what hour was being considered. The traffic at 2 A.M. and the traffic at 2 P.M. may well be different. A "normal" rate for one time would not be "normal" for another. And in addition to this diurnal cycle, there is an annual cycle, with traffic on Labor Day "normally" different from traffic on "Groundhog's Day." There is likewise a weekly cycle, with Sunday traffic "normally" different from Wednesday traffic. And in addition there is a secular trend, with the "normal" traffic for 1940 far different from the "normal" traffic for 1910, even if we pick the same month, day, and hour of the year.

Our secular trend might tell us that the basic hourly traffic was 42 cars per hour in 1930, with an increase of 8 cars per hour each year thereafter. We could put this in the form of a trend equation, from which we could estimate the "normal" traffic for any ensuing year, with 1930 as the origin. Thus, for 1944 the "normal" traffic based on secular forces alone would be $42 + 14(8)$, or 154 cars per hour. But if we are interested in knowing the facts for Sept. 28, our seasonal index may tell us that the traffic on this day of the year is 114 per cent of that for the normal day of the year. Therefore we should expect not 154 cars per hour, but 114 per cent of 154 cars; that is, we should expect 176 cars per hour. But this is for Sept. 28 on the average. If Sept. 28, 1944, falls on a Thursday, and if our studies show that Thursday traffic is but 78 per cent of "normal" traffic for the week, then we should expect not 176 cars per hour (as on a typical Sept. 28), but 78 per cent of 176 cars; that is, we should expect 137 cars. And finally, if the hour which we choose on Sept. 28, 1944, is the hour from 4 to 5 o'clock in the morning, we realize that we should not use the average traffic figure for the day. Our diurnal index may indicate that the traffic at this hour is but 25 per cent of the average traffic for the day. In that case we shall expect, not 137 cars in our hour, but 25 per cent of 137 cars; that is, we shall expect 34 cars.

We have now come down to making an estimate for the particular hour of 4 to 5 A.M. on Thursday, Sept. 28, 1944. Our estimate

of 34 cars is based on the long-run secular growth in traffic, and on the monthly, weekly, and hourly cycles. When we say that the "normal traffic" is 34 cars an hour, we are indicating that we should expect such an amount of traffic if we consider only those forces which work smoothly and regularly through time. Yet we do not mean that there must be exactly 34 cars passing the intersection in this hour, for there are many forces other than time-connected forces, which we have not considered at all. A heavy rain, an unusually early ice storm, imposition of restrictions on the sale of gasoline or tires, a resurfacing job on the road which detours traffic—these and numberless other "random" forces may come in to upset our calculations. When we say that the "normal" traffic for the particular year, day, and hour is 34 cars per hour, we realize that the actual traffic may fall short of this figure or may greatly exceed it.

These ideas can be expanded to cover any other statistical "normal." When we talk of "normal" department-store sales over the Christmas holidays, or "normal" temperatures on the Fourth of July, or a "normal" price of eggs, or a "normal" yield of hay, or a "normal" number of absences from school—in each of these cases we are setting up by more or less formal statistical means some "expected" value to be used for purposes of comparison. But in computing our "normal" we never include all the forces that may affect the value in question. If we did include all the forces, then there could never be anything "abnormal"—we should always hit the nail exactly on the head. The very concept of "normality" implies its counterpart, abnormality. When people say, as they sometimes do, that there never is a time that is really normal, they are not, as they often seem to think, proving that the concept of normality is useless. The "normal" occurrence is not what happens, but what would have happened if there had been no unusual transitory forces at work to make the result abnormal. And in any field, our idea of this abstract, hypothetical, idealized "normal" is made up by combining the effects of the various sorts of time-connected forces (secular, seasonal, diurnal, etc.), neglecting the "random" forces.

11.9. Suggestions for Further Reading.—Several of the references mentioned in Sec. 10.21, page 318, contain matter on time series in general and, hence, are applicable to cyclical as well as to secular movements. Karl G. Karsten, "Charts and Graphs," Prentice-Hall, Inc., New York, 1923, gives a

simple exposition in Chap. XXI. For dealing with cycles other than the 12-month seasonal cycle, the student may wish to investigate what is known as *periodogram analysis*. A short treatment of the method is given in Harold T. Davis and W. F. C. Nelson, "Elements of Statistics with Applications to Economic Data," pp. 137*ff.*, Principia Press, Bloomington, Indiana, 1935. The method is criticized adversely in the *Journal of the American Statistical Association*, Vol. 18, p. 889; and Vol. 22, p. 289. Many specialized books in the field of business statistics contain discussions of the so-called "business cycle" and its treatment. In this field the student should see particularly Wesley C. Mitchell, "Business Cycles, The Problem and Its Setting," National Bureau of Economic Research, Inc., New York, 1927, particularly Chap. III. Henry L. Moore, "Economic Cycles: Their Law and Cause," The Macmillan Company, New York, 1914; and William L. Crum, Periodogram Analysis, in "Handbook of Mathematical Statistics," edited by Henry L. Rietz, Houghton Mifflin Company, Boston, 1924, are both authoritative and helpful. For a treatment of the problem of random movements, see Gerhard Tintner, "The Variate Difference Method," Principia Press, Bloomington, Indiana, 1940.

EXERCISES

1. Why does it seem desirable to correct for calendar variation in the case of the data of Exercise 1 at the end of Chap. X (see page 319), and yet not to make similar corrections in the egg price data of Table 11.4, page 334?

2. Find at least two historical series showing distinct seasonal movements. Find two others containing cyclical movements of a nonseasonal character.

3. On page 335 we took the median of the figures in the table for each month. Some authors prefer to take the arithmetic mean each month. How much would the index of seasonal variation have been altered if we had followed this other procedure? Compute the index on the latter basis for purposes of comparison.

4. Compute an index of seasonal variation of New York City temperatures, using the basic data of Table 11.1, page 324. Use the moving average method described in Sec. 11.4.

5. Compute the index called for in the preceding exercise, but use the link relative method described in Sec. 11.5.

6. Eliminate the seasonal movement from the data of Table 11.1, page 324, using as a basis your index of seasonal variation computed in one of the two preceding exercises. Interpret your "corrected" data.

7. Annual and diurnal cycles have such an obvious astronomical basis that it is natural to expect to find cycles of these lengths in the data of almost every science. In how many separate sciences can you find illustrations of annual cycles? Diurnal cycles?

8. The weekly cycle has less natural basis than the annual or the diurnal cycle, yet the week has become firmly enough imbedded in our habits so that weekly changes appear in many kinds of data. In how many separate sciences can you find evidences of weekly cycles?

9. What is the amplitude of the cycle of egg prices in Table 11.4, page 334? Use as your basis of measurement one of the indexes of seasonal variation computed in the chapter.

10. In Sec. 11.7 we took a freehand straight-line trend as the basis for eliminating secular trend. Some people would prefer to be more "exact" and use a least-squares straight line, although we know (see Sec. 10.11) that the least-squares method is no sure cure. Compute the least-squares straight line for the data of the table on page 341, and recompute Table 11.8 on the basis of this least-squares trend.

11. Seasonal movements are by no means confined to prices. Table 11.9 gives figures[1] showing the number of strikes beginning each month from 1927–1936. Compute the index of seasonal variation by the link relative method.

TABLE 11.9.—NUMBER OF STRIKES BEGINNING EACH MONTH, 1927–1936

Year	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
1927	35	63	70	84	95	80	55	56	58	50	28	33
1928	45	46	41	69	80	44	56	53	48	60	37	25
1929	50	51	68	121	121	77	81	86	99	73	60	34
1930	49	49	47	68	58	61	79	53	68	42	36	27
1931	58	52	53	78	104	66	67	78	81	68	57	48
1932	88	60	63	89	91	74	72	89	86	50	43	36
1933	83	67	106	89	161	154	237	261	233	145	87	72
1934	98	94	161	210	226	165	151	183	150	187	130	101
1935	140	149	175	180	174	189	184	239	162	190	142	90
1936	167	148	185	183	206	188	173	228	234	192	136	132
Totals....	813	779	969	1171	1316	1098	1155	1326	1219	1057	756	598

12. Are the data of Table 11.9 of a type such that it might be wise to correct them for calendar variation? If so, make the necessary corrections.

[1] From DALE YODER, Seasonality in Strikes, *Journal of the American Statistical Association*, Vol. 33, No. 204, December, 1938, p. 687.

CHAPTER XII

INDEX NUMBERS

The use of index numbers has been pretty largely confined, in practice, to the fields of economics and business; yet the applicability of index numbers seems to be general enough so that there should be some gain in applying them more widely in other fields. To a very considerable extent, index numbers have been used to compare situations at different periods of time; yet they can also be used for making comparisons of different geographical areas, different business units, or almost any other sets of categories. Theoretically, index numbers can be used as broadly and as generally as any of the other statistical measures that we have treated, and the fact that their use to date has been pretty largely confined to a few fields of science should not preclude a consideration of them even in a general, nonspecialized textbook.

Within the fields of economics and business, index numbers have been used to make many kinds of comparisons: comparisons of prices, of volumes of business, of costs of production, of employment, of wages, of volume of output, of buying power, of living costs, etc. Historically the first, and still the most common, use of index numbers is in the making of comparisons of prices at different times. Of course, for a simple comparison of the prices of a single commodity at different times we do not need an index number. If we are told that a pair of shoes cost 50 cents[1] in 1805 and $5 in 1905, we can make the comparison directly without the computation of any complicated statistical coefficients. To be sure, we should still have to satisfy ourselves that the shoes were of comparable quality, but the problem would obviously be far simpler than it would be if we asked what had happened to the cost of living in general between 1805 and 1905. In the latter case, we should doubtless find that during the century the prices of some things had risen, some fallen, and some stayed about the same. It would no longer be possible merely to

[1] This is the actual average cost taken from an old family account book for a large family in upper New York State.

compare two simple figures. As soon as we get to making comparisons between complicated things, such as the cost of living or the prices of farm products, we need some sort of statistical help.

12.1. A Simple Aggregative Index Number.—Suppose you have a list of the prices of 20 commodities in 1930, and another list showing the price of each of these 20 commodities in 1935. You note that some of the commodities have risen in price and some have fallen in price, and you are interested in knowing whether prices in general have risen or fallen. Of course, you might count the cases in which prices had risen and the cases in which prices had fallen, and if you found 14 increases and 6 decreases you might conclude that prices had in general risen. Yet the 14 increases might be small and the 6 decreases large, in which case it is quite possible that the decreases would be more than enough to offset the increases. It is obvious that what you will need is some single summary figure that will characterize these increases and decreases. This is exactly the same problem that we faced in computing averages, and it is handled in approximately the same way.

Let us first take a hypothetical case. We shall assume that there are five commodities: *A*, *B*, *C*, *D*, and *E*. They are of equal importance. Their prices in 1930 and in 1935 are given in Table 12.1. Inspection will show that three of these commodities have risen in price (*A*, *C*, and *E*). One has not changed in price (*D*), and one has fallen in price (*B*). How can we get a single summary figure that will tell us the amount of the change in price?

TABLE 12.1.—HYPOTHETICAL PRICE DATA

Commodity	1930 Price	1935 Price
A	$1.00	$1.17
B	0.40	0.30
C	0.60	0.66
D	1.12	1.12
E	0.60	1.20

We could add the two columns of figures and find whether it takes more or less than before to buy one unit of each com-

modity. This would give us the sum of $3.72 for 1930 and of $4.45 for 1935. In other words, it took somewhat more money to buy this bill of goods in 1935 than in 1930. We might well express the change as a percentage, letting the sum of the prices in 1935 be stated as a percentage of the 1930 sum. In this case we should say that when the 1930 prices are considered as 100 per cent, the 1935 prices are $4.45/$3.72 = 119.5 per cent. More commonly we should say merely that the 1935 index number on a 1930 base in 119.5. We see, then, that the *base* of an index number is the period that is taken as the basis for comparison. We let the prices of the base period be 100 per cent and compute the relation of the prices of other years to the prices in this base period. Instead of saying that the base is 1930, one would usually give the index numbers for the various years with the statement, "1930 = 100." We see also from our simple example that an index number shows the percentage by which a group of values taken at one time or place differs from another group of values taken at another time or place. The index number which we have just computed is called the *simple aggregative index number*. Such index numbers are computed by adding the values for each year and stating the sum for each year as a percentage of the sum in the base year.

12.2. Averages of Relatives.—Another summary figure could be obtained by stating the price of each commodity in 1935 as a percentage of the price in 1930, and then taking some average of these percentage figures. If we convert the figures of Table 12.1 to percentages of the 1930 price, we get Table 12.2.

TABLE 12.2.—RELATIVE PRICES FROM TABLE 43; 1930 = 100

Commodity	Relative Prices	
	1930	1935
A	100	117
B	100	75
C	100	110
D	100	100
E	100	200

Each figure in this table is a percentage showing the price of the particular commodity in each year relative to the price in 1930.

Such figures are called *relatives*, and since these are based on prices they are called *price relatives* or *relative prices*. Any number is a relative if stated as a percentage of some other number.

If we take the arithmetic mean of the price relatives for 1930, we obviously have 100 for an answer. The mean of the five relatives of 1935 is 120.4. We can say that the index number of prices is 120.4. This would be an index number based on the *mean of the relatives*. We might as well have computed the index number based on the *median of the relatives* or the *geometric mean of the relatives* or the *harmonic mean of the relatives*. Such methods are commonly used in index-number work. Our results by these methods would be as follows:

Median relative	110
Geometric mean of relatives	114.0
Harmonic mean of relatives	108.9
Mean of relatives	120.4
Aggregative	119.5

The last two figures are from our earlier computations.

It will be seen, then, that we have computed the index number by five methods and have found five different answers, ranging from a high of 120.4 to a low of 108.9. Each of these answers purports to show the percentage, on the whole, which 1935 prices are of 1930 prices. All are based on the same figures; yet no two agree. This is not surprising, since we discovered when we were studying averages that the mean, the harmonic mean, and the geometric mean always (unless the values averaged are identical in size) differ.[1]

12.3. Bias in Index Numbers.—Again we see that the arithmetic mean is larger than the geometric mean, and that the harmonic mean is smaller than either. With these different results, which are we to use? Suppose that we selected a very large number of commodities for study. We found their prices in the base year, and then the year later we found the prices again. Suppose, moreover, that there had been no general change at all in price level. Individual commodities had changed in price, to be sure, but at random. There had been no movement downward and no movement upward in general. What should we find at the end of the year? Obviously some of the price relatives would

[1] See p. 71.

be larger than 100 and some smaller than 100. It is probably reasonable to assume that if the prices are changing at random (there being no fixed movement upward or downward) a price is as likely to double as to be cut in half. A commodity that doubled in price should just balance one that fell to half its former price. But note that the price relatives of these two commodities would be 200 and 50; that is, 200 for the one that doubled and 50 for the one that was cut in half. The arithmetic mean of these two relatives is $(200 + 50) \div 2 = 125$. Instead of showing no change, the arithmetic mean of the relatives shows a rising price.

Let us, then, try the harmonic mean. The harmonic mean of the two numbers is $2 \div (\frac{1}{200} + \frac{1}{50}) = 80$. By this method we find that prices have been falling; yet we know that there have been merely chance changes of price with no real rise or fall. When we try the geometric mean of the two prices, we find

$$\sqrt{(200)(50)} = \sqrt{10{,}000} = 100$$

This method, then, shows neither a rise nor a fall. In other words, the geometric mean gives a true picture of the situation. Rates of change can properly be averaged only by the geometric method. For this reason there is a decided advantage in using the geometric mean of the relatives in computing any simple index number.

12.4. Weighting of Index Numbers.—Now let us face another problem. Going back to the price changes of the five commodities tabulated on page 351, we may well ask if these commodities differ in importance. If milk rises in price by 2 cents per quart, the effect on the family budget is much greater than is an increase of 5 cents each on hairbrushes. In fact, a 5 per cent increase in the price of milk is much more important to consumers than a 30 per cent increase in the price of hairbrushes. Yet with the methods we have been using they would be given equal weight in the index. If milk prices rose 2 per cent and hairbrush prices fell 2 per cent, the mean of the price relatives would show no change—yet consumers would feel that prices had risen. To them the increased price of milk is not offset by the lowered price of hairbrushes. Obviously the thing for us to do is to take a weighted average of the price relatives rather than a simple average. We could do this by counting the milk price 10 times and the hair-

brush price once. It is simpler merely to multiply the milk price by some figure that represents its importance, the hairbrush price by some figure that represents its importance, etc. Thus for each commodity we shall have a *weight* which indicates the importance of the commodity. If our index number is to be one of the cost of living, we may well weight commodities according to their importance in the budget.

Let us suppose that the relative importance of the five commodities in our price table is as follows:

A	2
B	20
C	1
D	5
E	1

Thus commodity *B* is 10 times as important as commodity *A*, 20 times as important as either commodity *C* or commodity *E*, and 4 times as important as commodity *D*. Let us multiply the price relatives from Table 12.2, page 352, by these weights and divide the sum of the products by the sum of the weights to get the weighted arithmetic mean of the relatives. This procedure gives the figures in Table 12.3. The total of the weights is

TABLE 12.3.—COMPUTATION OF WEIGHTED AVERAGE INDEX OF RELATIVES

Commodity	Price Relative	Weight	Product
1930			
A	100	2	200
B	100	20	2000
C	100	1	100
D	100	5	500
E	100	1	100
1935			
A	117	2	234
B	75	20	1500
C	110	1	110
D	100	5	500
E	200	1	200

$2 + 20 + 1 + 5 + 1 = 29$. The total of the products for the year 1930 (the sum of the five figures in the first group of the last column in the table) is 2900. If we divide the latter figure

by the former, we get $^{2900}/_{29} = 100$, the index number of the base period. In practice it would not be necessary to compute the index number for this year, since it is 100 by definition. If we take next the products for the year 1935, we find that they add to 2544. The index number for this year is, then,

$$^{2544}/_{29} = 88$$

This is the *weighted average index of the relatives* as contrasted with the simple average index computed before.

The most noticeable feature of the present index as compared with those we computed before (see page 353) is that all the others gave values over 100, whereas this gives 88. The reason is easy to see. Before we used no weighting; we pretended that the commodities were of equal value. Yet really commodity *B*, the only commodity which fell in value, was far more important than all the others together. It is proper that it should have more influence on the result. We have now given it influence commensurate with its importance, and as a result the entire index has fallen. We can say, then, that prices are now but 88 per cent of their 1930 value, and we shall be more nearly correct than before. For if the various commodities are given their proper weighting, the change in price has the same effect that a uniform drop of 12 per cent in all prices would have had.

We could, of course, take the weighted geometric mean of the price relatives. This would involve raising each price relative to a power equal to the weight of the commodity, multiplying

TABLE 12.4.—COMPUTATION OF WEIGHTED GEOMETRIC AVERAGE OF RELATIVES

(1) Commodity	(2) Relative Price	(3) Log of Relative Price	(4) Weight	(5) (3)(4)
A	117	2.06819	2	4.13638
B	75	1.87506	20	37.50120
C	110	2.04139	1	2.04139
D	100	2.00000	5	10.00000
E	200	2.30103	1	2.30103
Totals			29	55.98000

together these figures for all the commodities, and taking a root equal to the sum of the weights. The work would be done by logarithms, and we can illustrate it by computing the weighted geometric mean of the relatives for the year 1935 (see Table 12.4). The logarithm of the index number is 55.98000/29 = 1.93034. This gives us a value of 85.2 for the index number, which is again lower than the weighted arithmetic mean.

Likewise we may wish to weight the aggregative index number. We do this by multiplying the price of each commodity by its weight and adding the products. We then divide 100 times the sum for any given year by the sum in the base year, and the quotient is the required weighted aggregative index number. If we illustrate with the data we have just used, we get

(1) Commodity	(2) Price	(3) Weight	(4) Product
1930			
A	\$1.00	2	\$ 2.00
B	0.40	20	8.00
C	0.60	1	0.60
D	1.12	5	5.60
E	0.60	1	0.60
			\$16.80
1935			
A	\$1.17	2	\$2.34
B	0.30	20	6.00
C	0.66	1	0.66
D	1.12	5	5.60
E	1.20	1	1.20
			\$15.80

To find the weighted aggregative index number for 1935 on a 1930 base

$$\frac{100(15.80)}{16.80} = 94.1$$

This, we see, is higher than either of the other two weighted index numbers which we have computed from these data.

12.5. Weight Bias.—We saw in Sec. 12.3 that, with chance variation among the relatives, the geometric mean yields an index

of 100, but that the result by the arithmetic method is too high and that by the harmonic method too low. We say that the arithmetic method is characterized by an *upward type bias* and the harmonic method by a *downward type bias;* that is, one of them tends to give values which are too large and the other values which are too small. For this reason any simple (unweighted) index can be computed by the geometric method to advantage. But when we come to adding weights, as we have done in the more recent examples, we encounter the new difficulty that bias may arise in our index numbers from the weights which are used. Index numbers are commonly computed with fixed weights, as in our examples. That is, the weights for 1930 were the same as those for 1935; and, if we computed indices for other years, we should still use these same *fixed-year weights.*

It is possible, of course, to use as weights values which change from year to year, using in each year a figure that shows the importance of the commodity in that year. Such weights are called *given-year weights.* Now fixed-year weights introduce a *downward weight bias,* and given-year weights introduce an *upward weight bias* in index numbers. If we start with the unbiased geometric method and introduce weights, we bias our results. If we use the arithmetic method with its upward-type bias and combine it with fixed-year weights (which have a downward *weight* bias), we overcome to some extent the type bias with the weight bias. If we use the harmonic method in conjunction with given-year weighting, we overcome the downward type bias to some extent by the upward weight bias.

Our illustration has included but five commodities. If we included a large number of commodities, we might choose as our index number the median or the modal price relative. We have seen that the mode is hard to determine unless there are enough cases so that they can be easily grouped, and that it is not always clearly marked even then. For this reason indices are seldom based on the mode. The median has the advantage that it neglects the extreme cases entirely. But, on the whole, index numbers are computed by the arithmetic, geometric, or harmonic methods or on the basis of aggregates.

12.6. Uses of Index Numbers.—Index numbers can be used whenever one wishes to compare changes in groups of values from time to time. Their commonest uses are in the measurement of

changes in the general price level, the cost of living, the rate of wages, etc. But whenever groups of values vary and we want some single summary figure with which to express the variation, the use of index numbers is indicated. Suppose, for example, that we want to trace changes in the sanitary conditions in a given city. We decide that the healthfulness of the city can be determined in part by the infant death rate, in part by the percentage of dwellings having modern plumbing, in part by the number of absences from the public schools, etc. We determine first which things to include. We must then determine the proper relative weights. The computation of the index number is then simple.

To illustrate, suppose that we are to measure the healthfulness of a city by the three items mentioned above. As the conditions become more healthful, the infant death rate will presumably fall; but we want values which increase with the healthfulness. Let us take, therefore the difference between the infant death rate and, say, 300; that is, we shall subtract each infant death rate from 300. Likewise it is to be expected that the number of absences from school would decrease as the healthfulness of the city increased. There would also be variations in the number of absences according to the number of pupils registered in the schools from year to year. Thus our measure might well be the average number of days attended per pupil divided by the number of days that school was in session. If there were 700 pupils in the town and they attended an average of 200 days each, and if the schools were in session 203 days during the year, our measure would be $^{200}/_{203} = 98.5$. The attendance would be, in other words, 98.5 per cent of the total possible attendance.

Now let us take a hypothetical case. In a given city we have the figures on infant mortality, on school attendance, and on plumbing, for two successive years. Let us call the years 1920 and 1921. In 1920 the infant death rate was 105.2, 73 per cent of the houses had modern plumbing, and the school attendance was 96 per cent of maximum. In the second year the infant death rate was 103.4, the school attendance was 95.5 per cent of maximum, and 75 per cent of the houses had modern plumbing. If we are not to weight our figures, the "index of healthfulness" will be computed as in Table 12.5. The arithmetic mean of the

TABLE 12.5.—COMPUTATION OF "INDEX OF HEALTHFULNESS"

Measure of Healthfulness	Value	Relative Value
1920		
Infant mortality	194.8	100
Plumbing	73.0	100
School attendance	96.0	100
1921		
Infant mortality	196.6	100.9
Plumbing	75.0	102.7
School attendance	95.5	99.5

three relatives for 1921 is 101.0. We could say, then, that the health index of the city rose from 100 to 101 during the year.[1]

Suppose, however, we decide that a change in the infant death rate is ten times as important as school attendance as a health indicator, and that the plumbing conditions are twice as important as school attendance. This assumption would give us weights of 10 for the infant death rate, 2 for plumbing, and 1 for school attendance. If we weight the relatives on this basis, we have the figures shown in Table 12.6. The health index for 1921 is, then, 1313.9/13 = 101.1.

TABLE 12.6.—COMPUTATION OF WEIGHTED "INDEX OF HEALTHFULNESS"

Measure of Healthfulness	Relative Value	Weight	Product
1921			
Infant mortality	100.9	10	1009.0
Plumbing	102.7	2	205.4
School attendance	99.5	1	99.5
Totals		13	1313.9

It should be obvious that such an index might be useful in comparing different cities at a given time as well as different

[1] The figures for infant mortality in the table are 194.8 and 196.6. These are obtained by subtracting the actual rates of 105.2 and 103.4 from 300, as explained in the text.

times in the same city. Thus indices may be geographical as well as chronological.[1]

12.7. Correcting Prices with Index Numbers.—The farm price of potatoes in the United States on Dec. 1 of various years is given in Table 12.7.[2] In the same table are given the index numbers of the wholesale prices of "all commodities" for these years.[3]

TABLE 12.7.—FARM PRICE OF POTATOES IN THE UNITED STATES ON DECEMBER 1, AND INDEX NUMBER OF WHOLESALE PRICES OF ALL COMMODITIES, 1916–1925

Year	Farm Price of Potatoes (cents per bushel)	All-commodity Price Index (1910–1914 = 100)
1916	146	125
1917	123	172
1918	119	191
1919	158	202
1920	113	226
1921	108	143
1922	56	141
1923	76	147
1924	62	143
1925	187	151

This index number is based on the prices of a large number of commodities (well over 800 at present) and is here given with the average of the years 1910–1914 as a base. Thus the index number for 1919 means that wholesale prices were, in general, 202 per cent of their average 1910–1914 level. They had risen to over double their average value for the base period.

[1] This example of an "index of healthfulness" is to be considered by the student as something quite as hypothetical as the price index in which we worked with commodities A, B, C, D, and E. It may well be that the factors here listed are not important as measures of healthfulness, and it must be that their relative importance is far from that here given. Also the subtraction from 300, etc., is quite arbitrary. The author apologizes to vital statisticians for intruding upon a field about which he is ignorant, but it seems worth while to point out to the student that there are uses for the methods here discussed other than uses represented in the analysis of prices.

[2] From *Statistical Abstract*, 1933, p. 597.

[3] From *Farm Economics*, Cornell University, September, 1931. pp. 1586–1587.

It will be noted from the table that the farm price of potatoes was much lower in the years 1922–1924 than it had been before. But it is also well known that this was a time when prices were falling generally in the deflation after the war. The index numbers show that prices reached their postwar peak in 1920 and fell sharply during the following year.

There are at least two things that can account for a change in the price of potatoes. In the first place, the potatoes may have relatively more or less value in terms of all other commodities on account of the size of the potato crop or of changes in people's desires for potatoes. In the second place, there may have been changes in the value of money itself. We commonly measure the value of other things in money terms, but, as has been often pointed out by economists, money is a poor measuring stick because it is not constant in value. During and immediately after the war almost all prices rose. It took more money to buy the same goods; the value of money had fallen. Then came the break in prices, and money suddenly became more valuable again.

Now we are interested in knowing whether these changes in the price of potatoes were due merely to fluctuations in the value of money or whether they show some change in the economic position of the potatoes themselves. Our price index tells us (insofar as it is an accurate measure of changes in the purchasing power of money) that it took $1.25 in 1916 to buy what $1 had purchased during 1910–1914. By 1919 it took $2.02 to buy this same bill of goods; in other words, money had become much less valuable during the period. If the value of potatoes measured in terms of commodities other than money had remained constant throughout this period, the price of potatoes would have risen because the value of money had fallen. During this period the price of potatoes did rise from $1.46 per bushel to $1.58 per bushel. Can this change be accounted for entirely by the change in the value of money? We discover the answer by finding the corrected prices of potatoes.

If the dollar had the same purchasing power in 1916 that it had in the base period, then it would still have taken $1 to buy the goods that were really selling for $1.25 in 1916. In other words, prices would have been reduced from 125 to 100. We can do this by dividing prices for 1916 by 1.25. Similarly,

if we divide the prices for 1917 (when the index number was 172) by 1.72, we shall be putting the prices on a basis of dollars with the purchasing power that dollars had in the base period. And if we divide each price by the index number for the same year (remembering that an index number is a percentage, and therefore pointing off two places when dividing), we shall be stating the price for each year as it would have been had the dollar retained a constant purchasing power equal to that which it had in the base period. Since the base period was the average of the years 1910–1914, we could call these prices "prices in 1910–1914 dollars" to show that we are talking about prices measured in a dollar which supposedly has constant purchasing power. It is also common to call such prices *corrected prices* to indicate that they have been corrected to account for changes in the value of money. A corrected price, then, is a price which has been divided by the index number for the year (or other period that the price may represent).

Since these are Dec. 1 prices of potatoes, it would be better for us to correct them with the index numbers for Dec. 1 of the years given. But, since it is our purpose merely to illustrate the process of correction, we shall not bother to undertake a refined analysis of the data.

TABLE 12.8.—THE CORRECTION OF POTATO PRICES

Dec. 1	Farm Price of Potatoes	
	Actual	In 1910–1914 Dollars
1916	146	117
1917	123	71.5
1918	119	62.4
1919	158	78.3
1920	113	50.0
1921	108	75.6
1922	56	39.7
1923	76	51.7
1924	62	43.4
1925	187	124

We can, then, "correct" the potato prices from Table 12.7, page 361, by dividing each price by the index number. This

would give us the corrected prices. These appear in Table 12.8, and the original prices are given with them for purposes of comparison. These corrected figures are intended to show the relative purchasing power of potatoes in terms of other commodities, not merely in terms of money. They show that, although the price of potatoes fell somewhat between 1920 and 1921, other things in general fell more, so that the same quantity of potatoes would buy relatively more of other things. In many economic problems these corrected prices are far more important than the actual prices. Although no index number has ever been devised which measures changes in the general price level with absolute accuracy and to the satisfaction of everyone, nevertheless correction of prices by what index numbers we have is certainly much better than no correction at all.

12.8. The Choice of a Base Period for Index Numbers.—We have discovered that the base period is the period with which comparisons are made—the period which is taken as 100 per cent and from which all the other index numbers are computed. One can, of course, select for a base period any period he wishes. We could base price indices on the prices which were being paid at 10 A.M. on Jan. 9, 1935. Usually the base period represents the average values for a period of time, such as a year. Surely, if we are going to compare all of our index numbers with that of the base period, it is important to select a base period which is "representative"—which is "normal" in some way. One may well say that no period is normal, but at least one can understand what is meant when it is said that the prices of 1920 were "abnormal." If we were computing a price index, probably we should not select a year such as 1920 for the base.

In addition to choosing a base period which is "normal," there are some advantages in choosing a base period which is not too distant. If we are interested primarily in present-day prices, there are some disadvantages in comparing always with the prices of 1890 or 1913. Here the base period is so far removed that there is no reason at all for considering the prices of that time as "normal" prices for the present. Moreover, even without any general movement of prices up or down, the scatter which would occur in the original prices by pure chance variation would ultimately become so great that the type bias would make

itself felt strongly. Hence one takes, usually, as a base period a "normal" period of the recent past.

In addition there are advantages in taking as a base period some period which is commonly used by others who are computing index numbers, so that your results may be compared easily with theirs. Commonly used base periods are the average of the years 1890–1899, the year 1913, the average of the years 1910–1914, and the year 1926. The last is at present in most common use.

12.9. Link Relatives and Chain Indices.—Writers sometimes use a moving base for their index numbers, hoping to increase the accuracy of the index numbers for year-to-year comparisons. Although it has been shown that such index numbers give no increase in accuracy, but rather the reverse,[1] they do seem to offer some advantage when it becomes necessary to change the list of commodities included in the index, or to alter their weights. Under such circumstances the index for each year (or other time period) is computed with the preceding year as a base, using whatever price and weight data are available. Similarly the index of the third year is computed with the data of the second year as a base, the index of the fourth year with the data of the third year as a base, etc. The index for each year is thus given with the preceding year as 100. Our results might look like those of column 2 in the accompanying table. Our first year is taken as 100, since we have no preceding year with which to compare it. These index numbers, computed on a moving base, are called *link index numbers*. Price relatives so computed would be called *link relatives*.

These index numbers are related to no common base, but usually we wish so to relate them. We do this by "chaining" them together in what is called a *chain index*. It will be noted in the table that the link index for 1931 is 92; this means 92 per cent of the preceding year. Since the preceding year is 100, we have $92 \times 100 = 92$ (these figures all being percentages, and hence having decimal points before the last two digits). In 1932 the index was 105 per cent of the preceding year, or 105 per

[1] ALLYN A. YOUNG, Index Numbers, in "Handbook of Mathematical Statistics," pp. 183–184, H. L. RIETZ, ed., Houghton Mifflin Company, Boston, 1924.

Year	Link Index Number	Multiply by	Chain Index Number	Chain Index (1935 = 100)
1930	100		100	103.6
1931	92	100	92	95.3
1932	105	92	96.6	100.1
1933	102	96.6	98.5	102.1
1934	100	98.5	98.5	102.1
1935	98	98.5	96.5	100.0
1936	96	96.5	92.6	96.0
1937	101	92.6	93.5	96.9
1938	105	93.5	98.2	101.8
1939	104	98.2	102.1	105.8

cent of 92. This gives us 96.6 for the 1932 index. The index for 1933 is 102 per cent of the 1932 index, or 102 per cent of 96.6. This gives us 98.5 for the 1933 index. Similarly we chain together the indices of the other years, multiplying each link index number by the chain index number of the preceding year. Our resulting chain index is based on 1930. If we wish to use any year other than this first year as a base, we can easily convert, as has been done in the table. For example, if we prefer a 1935 base, we can divide each chain index on the 1930 base by 96.5 (the 1935 chain index on the original base). Thus we have the figures in the last column of the table above.

It is evident that this method of computation can be used even though radical changes are made in the commodities covered by the index. It is necessary only that we compute our original links on comparable sets of data; that is, when we compute the link for 1932 it is necessary that we use comparable data for 1931 and 1932. Likewise, when we compute the link for 1933 it is necessary that our data for 1932 and 1933 be comparable. But it is not necessary that the data for 1932 be the same in both these links.

We can summarize as follows the method of computing a chain index:

1. Calculate the index for each period with the data of the preceding period as a base.
2. Chain the links together to form a chain index with the first period as 100. This is done by calling the chain index for the first period 100 and then multiplying each link by the chain index of the preceding period.

3. Shift the chain index to the desired base by dividing each chain index number just found by the chain index number for the desired base period.

12.10. Choosing a Formula for Index Numbers.—The problem of computing index numbers is the problem of describing a universe from a sample. For example, if we wish to compute an index number to show changes in the level of the prices of farm products, we cannot ordinarily include data on all agricultural transactions. We shall be forced to base our index number on the prices of only a part of the possible commodities, and even in the cases of the commodities included we shall have price quotations on only a few of the actual transactions. But we hope that the price movements registered by the commodities and transactions chosen will be typical of the movements of all farm prices.

We have already seen that some methods of computing index numbers will introduce bias into our results—that the method of computation itself will make the index number increase or decrease in size even though the general level of the prices themselves has not changed. Mere increases or decreases in the dispersion of these prices will affect the size of the index number.

In order to eliminate or minimize the various types of bias that may arise, we find that many more or less complicated refinements have been introduced in index-number computation or suggested in index-number theory. Some of the suggestions are in themselves so intricate and time-consuming that they are never applied in practice. The workaday statistician is likely to forego the time and labor involved unless the size of the correction is considerable.

Professor Irving Fisher has made a careful study of various proposals for computing index numbers[1] and has suggested various tests to be applied to any formula to indicate whether or not it is satisfactory. The two most important of these he calls the *time-reversal test* and the *factor-reversal test.* If an index number is to meet the time-reversal test it must be so computed that the index number for any year X to the base year Y is the reciprocal of the index number for the year Y to the base year X. An index number meets the factor-reversal test if a price index and a quantity index computed from the same data will yield, when multiplied together, the value index derived from the same

[1] IRVING FISHER, "The Making of Index Numbers," Houghton Mifflin Company, Boston, 1922.

data. This test likewise requires that the prices and quantities may be interchanged without invalidating the test.

In attempting to meet these tests and others which have been suggested, various statisticians have suggested a multitude of formulas. Most of these are much more complicated than the methods we have illustrated in this elementary text. We can, however, illustrate their complexity and their general nature best by giving one or two of the formulas which seem best adapted in theory to meet the tests.

If we let the price of a commodity in the base year be represented by p_o, while the price in any other given year is p_i; and if we let the weight in the base year be q_o and in the given year q_i, Fisher concludes that the "ideal" index number would be found by means of this formula:

$$\text{Index} = \sqrt{\left(\frac{\Sigma(p_i q_o)}{\Sigma(p_o q_o)}\right)\left(\frac{\Sigma(p_i q_i)}{\Sigma(p_o q_i)}\right)}$$

Another formula, the results of which are in practice almost identical with those of the "ideal" formula, is the aggregative formula of Marshall and Edgeworth, which follows:

$$\text{Index} = \frac{\Sigma(q_o + q_i)p_i}{\Sigma(q_o + q_i)p_o}$$

The computations involved with this formula are much simpler and shorter than those with the ideal formula, and the slight difference in results would seldom make the application of the "ideal" formula worth while.

It is evident that the simple methods discussed earlier in this chapter can all be described by means of formulas. For example, the simple arithmetic mean of relatives is found thus:

$$\text{Index} = \frac{\sum\left(\frac{p_i}{p_o}\right)}{N}$$

The simple aggregative index number would be

$$\text{Index} = \frac{\Sigma p_i}{\Sigma p_o}$$

12.11. Selection of Basic Data.—It will be recalled[1] that the reliability of an arithmetic mean varies, not in proportion to

[1] See Sec. 9.2, p. 239.

the number of cases on which it is based, but in proportion to the square root of that number. Since an index number is a sort of average, showing the typical movement among a whole class of cases, we should remember that its reliability, too, can be expected to vary roughly with the square root of the number of items on which the index is based if these items are of equal importance. Usually, however, it is possible to pick out some items that are far more important than others, and if we choose them first and give them proper weight, we soon come to a point where the remaining items are unimportant enough and the cost of collecting them great enough so that we are warranted in neglecting them. While some indices in actual practical use are based on data concerning several hundred commodities,[1] if the items selected are judiciously chosen it should be possible to compute a worth-while index on a relatively small number.

Troublesome always in selection of basic data is the problem of getting quotations that are comparable. Even with simple staples, it is difficult to compare the packaged butter or flour of today with the bulk commodities of a half-century ago. The problem becomes more difficult still when we are forced to deal with nonstandardized things such as women's dresses or entertainment, either one of which might well be included in an index of the cost of living. And when we come to commodities which may now be important but which formerly were not used at all, such as radios, the problem becomes very difficult indeed. No general rules can be laid down for such cases, and again it is necessary to emphasize the fact that good statistical work is largely nonmathematical in character, involving the use of good judgment by someone who is thoroughly familiar with the facts in his own field as well as with the technical statistical procedures.

12.12. Suggestions for Further Reading.—The student who wishes to go further in the field of index numbers will find Irving Fisher's, "The Making of Index Numbers," Houghton Mifflin Company, Boston, 1922, required reading. It is voluminous, but it is simply and understandably written, and is probably the outstanding book in this field. For a short but very satisfactory treatment, see William L. Crum and Alson C. Patton, "An Introduction to the Methods of Economic Statistics," Chaps. XVIII and XIX, McGraw-Hill Book Company, Inc., New York, 1925. Willford I.

[1] Perhaps the most useful of all indices published is the index number of the wholesale prices of "all" commodities, published by the U.S. Bureau of Labor Statistics. This index is based on over 800 commodities.

King, "Index Numbers Elucidated," Longmans, Green and Company, New York, 1930, is a valuable small book on the subject. King does not agree with Fisher's ideas on the subject of an "ideal" index number, concluding that the formula which is "best" depends on the problem at hand rather than on mathematical considerations. A useful description of a few of the leading current index numbers, followed by a very brief description of 45 such current published indices, may be found in Frederick E. Croxton and Dudley J. Cowden, "Practical Business Statistics," Chap. XVIII, Prentice-Hall, Inc., New York, 1934. Allyn A. Young's chapter on Index Numbers which appears as Chap. XII of the "Handbook of Mathematical Statistics," edited by Henry L. Rietz, Houghton Mifflin Company, Boston, 1924, is very short and very good.

EXERCISES

1. Check the computation of the indices given on page 353, showing your methods.

2. How would you go about the process of computing the weighted harmonic mean of the relatives from which we computed the weighted geometric mean on page 356?

3. On pages 359–360 is described the computation of an "index of healthfulness." Suppose that in 1922 the infant mortality rate was 95, the school attendance was 99 per cent of maximum, and 80 per cent of the dwellings enjoyed modern plumbing. Compute the weighted average index of healthfulness for 1922 to compare with that computed in the text for 1921 (see page 360).

4. Using the fictitious commodities of the example that begins on page 351, suppose that the prices of our five commodites in 1936 are

A	\$1.20
B	0.45
C	0.70
D	0.95
E	1.00

Compute the weighted arithmetic average index number for 1936 to compare with that computed for 1935 on page 356. Compute the weighted geometric average index number for 1936 to compare with that for 1935 computed on page 357.

5. In 1935 the index number of the wholesale price of all commodities was 117, while the index number of the cost of living in the United States was 140. Both numbers are from publications of the U.S. Department of Labor, and both are on a basis of the average figures for the years 1910 through 1914 (that is, 1910–1914 = 100). Compare the two figures, and comment.

6. We hear a good deal about the virtues of a "random sample." If you were selecting items to be included in the computation of an index number, would you select items at random, or would you select them according to a plan? In either case, why?

7. To illustrate how index numbers can be used in fields other than economics and business, outline the items that you would include in an index number to be used for comparing the scholastic standings of various colleges and universities. Select things that can be numerically expressed.

8. Give what you think are approximately correct weights to the items that you have enumerated in the preceding exercise.

9. When comparing nontemporal phenomena, as in Exercise 7 above, there is no "base period." How would you select the base for your index number in such a case?

10. Table 12.9 gives the United States production of petroleum, Pennsylvania anthracite coal, bituminous coal, and coke for the years 1930 through 1939. The value of the outputs of the four fuels in millions of dollars in 1939 was[1]

Petroleum	$1265
Anthracite	187
Bituminous	733
Coke	213

Compute a weighted index number of the volume of fuel production in the United States for each of these 10 years, using the values as weights and using an arithmetic average of the relatives. Use 1935 as a base.

TABLE 12.9.—UNITED STATES OUTPUT OF CERTAIN FUELS, 1930–1939

Year	Petroleum (millions of barrels)	Anthracite (millions of tons)	Bituminous (millions of tons)	Coke (millions of tons)
1930	898	69.4	468	48.0
1931	851	59.6	382	33.5
1932	785	49.9	310	21.8
1933	906	49.5	334	27.6
1934	908	57.2	359	31.8
1935	997	52.2	372	35.1
1936	1100	54.6	439	46.3
1937	1279	51.9	446	52.4
1938	1214	46.1	349	32.5
1939	1264	51.5	393	44.3

[1] All data for this exercise from "World Almanac," 1941, pp. 601–602.

CHAPTER XIII

SIMPLE LINEAR CORRELATION

13.1. The Nature of Relationship.—Any study of the nature of relationship or causation raises philosophical problems which are far too abstruse for discussion here. It seems to be true that no one knows very clearly what is meant by the statement that one thing "caused" another thing. Yet for the purposes of everyday life there is enough meaning to the statement so that it helps men in their thinking, and gives them a convenient method of dodging the philosophical problems involved by means of an elliptical expression.

When we say that two things are "related," we may mean that the connection between them is very definite and unchangeable, or we may intend merely to call attention to some sort of loose connection between the two. For example, we say that the circumference and the radius of a circle are related. Here the relationship is definite and unalterable, and can be expressed by means of the mathematical equation

$$c = 2\pi r$$

For any given radius there is one and only one circumference, and this relationship of r to c remains the same century after century without end. Suppose, however, we say that the price of potatoes and the quantity produced are related, or that a child's age and height are related. The connection in this case is by no means so sharply defined. It is not true that for each child age there is one height and only one. Children of the same age vary in height. Similarly potato prices are not always equal for crops of equal size.

Let us take another case which is somewhat more complicated. If we study the period of the pendulum, we discover that there is a relationship between the length of the pendulum and its period. If we let t represent the time of the vibration, l the length of the

pendulum, and g the attraction of gravity, the relationship can be expressed by the formula

$$t = 2\pi\sqrt{\frac{l}{g}}$$

Now it will be noticed that there is no single period of oscillation which will occur with every pendulum of a given length. We cannot say that for each and every length of pendulum there is one and only one period of oscillation. As long as there are variations in the attraction of gravity we may get changes in the period of oscillation of a pendulum without changes in length. When we say, then, that there is a relationship between the length of the pendulum and its period, we do not mean that the relationship is a simple one. We do not mean that one can tell the exact period from a knowledge of the length. And when we say that there is a relationship between a child's age and his height, we likewise do not mean that a knowledge of the age will make it possible for us to tell the exact height of the individual. There are, of course, other factors which are related to height (just as there was the additional factor of gravity also involved in the swinging of the pendulum), and it is possible that, if we knew them all and knew the facts with regard to their interrelationships, we could tell the exact height of the child just as we can tell the exact period of the pendulum if enough data are given. Some men assume that all facts are so related that if we knew enough about them we could explain them all by methods as satisfactory as those used to explain the swinging of the pendulum.

In most statistical problems there are many variables, and the exact relationships which exist between them are unknown. We have no formulas from which we can give a complete mathematical statement of the problem. As we saw in Chap. I, the statistician commonly deals with problems in which there are many things varying at once, and where it is impossible on account of the nature of the data to hold forces constant. Hence relationships cannot ordinarily be stated so simply or so satisfactorily as can the relationship between the radius and the circumference of a circle.

What, then, do we mean when we say that there is a "relationship" between the height and the age of children? We may mean any one of a number of things. We may mean, for example, that

the *average* height increases (or decreases) with age, so that if we divide children into groups according to age we shall find changes in the average height accompanying changes in the age. We may mean that the *dispersion* of heights differs with age, so that the heights are more widely scattered at some ages than at others. We may mean that there are differences in *skewness* or *kurtosis* of the height distributions at different ages. If the frequency distributions of heights vary (more than they would vary as the result of chance) from one age to another, we should say that the ages and the heights are related.

This principle can be stated to advantage in a somewhat different way. To be sure, the knowledge of a child's age does not make it possible to estimate his height exactly. But does it help at all in estimating the height? Suppose that you have the problem of guessing the height of an unknown child; would it help you at all to be told that the youngster is two years old? This knowledge would not make it possible for you to tell the exact height, but it would make it possible for you to estimate the height with less error than would otherwise exist in your answer. In such cases, where a knowledge of the value of one variable helps us in estimating the value of another variable, we say that the two variables are "related." This does not mean that one of them "causes" the other, but merely that the knowledge of the value of one is an aid to us in estimating the value of the other. Suppose that you have the problem of estimating the price which will be paid for potatoes at retail in New York City next fall. You know the following facts:

1. Easter of the year in question falls on Apr. 4.
2. The Philadelphia Athletics have a team batting average of 0.231 on July 17 of the year in question.
3. The quantity of potatoes harvested in the United States in the year in question is 400 million bushels.
4. The price of rice is lower than it has been for 50 years.

Which of these facts will you consider when making your estimates? It is probable that you will give no weight at all to either of the first two. That will mean that in your opinion the price of potatoes is not related to the date on which Easter falls or to the team batting average of the Athletics. It may well be that you will consider the other two facts in making your estimate. This will mean that you think that there is some rela-

tionship between the price of potatoes on the one hand and the size of the crop and the prices of other sources of carbohydrate food on the other hand. If you can make a better estimate of potato prices by considering any certain factor than you could without considering it, then that factor is related to the price of potatoes. That is all that is meant by "relationship" as the word is used here, and statistical investigations can determine no more. No statistical process can demonstrate cause and effect, but there are statistical processes, which we shall now consider, that show whether or not it is worth while to consider certain particular factors in making estimates of others.

13.2. Simple Methods of Finding Relationships.—In our discussion of the nature of relationships we suggested that if children were classified by age, and if the average height were computed for each class, we could then see whether or not there were variations in the average heights at the different ages and thus infer something as to the existence or nonexistence of relationships. This is one of the simplest, easiest, oldest, and most satisfactory ways of discovering relationships and of presenting evidence as to the nature thereof. The common tables of height and weight are computed on such a basis. But the method can be applied in any field, and is usually one of the first methods used by a statistician who is investigating relationships of any kind. We can illustrate with Table 13.1, which shows the average income of Connecticut farmers growing Havana Seed tobacco in 1926 divided into classes according to their expenditures for fertilizer per acre.[1] This table

TABLE 13.1.—RELATIONSHIP OF FERTILIZER COST AND INCOME ON CONNECTICUT TOBACCO FARMS, 1926

Fertilizer Cost per Acre	Average Income
Under $50	$1092
$ 51–$ 75	1464
76– 100	953
101– 125	660
Over $125	92

shows that incomes (the incomes are net) fell as the fertilizer expense per acre rose. They give evidence of relationship

[1] C. I. HENDRICKSON, Tobacco Farm Organization, *Storrs Agricultural Experiment Station Bulletin* 165, p. 133.

between the size of net income and the per-acre cost of fertilizer. The interested statistician could now analyze his data further to determine, if possible, the degree of the relationship.

In 1916 the United States Public Health Service made a study in seven South Carolina cotton-mill villages, and as a part of the study they investigated the relationship between the size of the family income per person and the amount of sickness in the family. Counting only cases of disabling sickness, and stating the figures in rates per 1000 people, they found the following sickness rates at various income levels (see Table 13.2).[1] These figures indicate

TABLE 13.2.—RELATION OF SICKNESS TO INCOME, SOUTH CAROLINA, 1916

Half-monthly Income per Adult Male	Sickness Rate per 1000 Persons
Less than $6	70.1
$6–$7.99	48.2
8– 9.99	34.4
10 and over	18.5

that there is a relationship between income and sickness, although they do not show whether the people were sick because of privations resulting from small incomes, or whether their incomes were small because they were sick and hence not working regularly. In other words, nothing is shown as to cause and effect, but evidence is presented that a relationship exists.

One more illustration will suffice to show the nature of this method and the wide range of its applicability. Table 13.3 shows the average cost per mile of operating automobiles in relation to their value. The figures are based on a study of 910 automobiles used on New York State farms.[2] Here again there seems to be evidence of a definite relationship. One could, presumably, make a better estimate of the cost of operating a car for a mile if he knew the value of the car than if he did not.

[1] Quoted by DOUGLAS, HITCHCOCK, and ATKINS in their "Worker in Modern Economic Society," p. 318, University of Chicago, Press, Chicago, 1925.

[2] J. M. BANNERMAN, Economic Study of 941 Automobiles on New York State Farms, *Farm Economics*, Cornell University, June, 1931, p. 1565.

In order to show the difference in the results, let us examine a case in which computation of averages for the data when classified into groups gave no evidence of relationship. In the study of the

TABLE 13.3.—RELATION OF COST OF OPERATION TO VALUE OF AUTOMOBILE

Value of Automobile	Cost per Mile (cents)
$ 0–$ 94	3.62
95– 394	4.58
395– 694	6.09
695– 994	6.22
995– 1294	7.67
1295 and over	11.35

use of automobiles on New York State farms which we have just mentioned, figures of Table 13.4 are given showing the distance cars were driven per year in relationship to the distance that the owner lived from a paved road.[1]

TABLE 13.4.—RELATION OF SEASON'S MILEAGE OF AUTOMOBILE TO DISTANCE FROM HARD ROAD

Miles to Hard Road	Season's Mileage
0	4385
0.1–0.9	4152
1.0 and more	4342

Mere differences in the averages of groups give no conclusive evidence of the existence of relationship. As we have discovered in an earlier chapter (pages 250*ff.*), it would be advisable in such cases to compute the standard deviations within the groups and then the standard errors of the differences in the means of the groups. Our results would then be far more significant.

One must not get the impression from the examples which have been used that it is necessary for the group averages to increase or decrease regularly and continuously throughout the table before we can draw the conclusion that a relationship exists. Examine Table 13.5, in which couples are classified according to the length of their married life, and in which for each group the

[1] BANNERMAN, *op. cit.*, p. 1565.

number of divorces per 100 married population has been computed.[1] This table seems to show that the divorce rate rises for the first few years and then falls. But this fact would not deter one from using a knowledge of the length of married life in estimating the likelihood of divorce. The relationship merely turns out to be curvilinear. If the averages differ by a fixed amount, so

TABLE 13.5.—RELATION OF DIVORCE RATE TO LENGTH OF MARRIED LIFE

Years of Married Life	Divorces per 100 Married Population
1	0.70
2	1.20
3	1.32
4	1.32
5	1.27
6	1.10
7	1.00
8	0.97
9	0.84
10	0.67

that when graphed they would fall along a straight line or approximately so, we say that the relationship between the variables is *linear*. If the averages when plotted would fall along a smooth curve, or approximately so, we say that the relationship is *curvilinear*. And if the averages when plotted do not seem to fall along any curve whatever, we say that there is no evidence of relationship between the variables.

This method of discovering relationship by a comparison of group averages is exceedingly useful. Whenever the statistician is beginning to study a problem and is investigating the relationships involved, he is likely to use this method first of all. It tells him whether or not it is worth while to proceed with more complicated methods, and gives him a basis for selecting the type of method to use. It has the advantage that the results are

[1] This is the first part of a long table in A. Cahen, "Statistical Analysis of Divorce," p. 120, Columbia University Press, New York, 1932. In the original the entries continue to 30 years, and the divorce rates continue to decrease throughout the remainder of the table.

readily understood even by the uninitiated and that the computation is short and easy.

13.3. The Scatter Diagram.—We turn now to a second simple method for studying relationships, less common than the one just mentioned, but, nevertheless, a very helpful tool. This is the method of plotting the data on a *scatter diagram*, or *scattergram*, in order that one may see the relationship. It is a graphic method, making its appeal to the eye.

On Jan. 26, 1918, records were made of the temperatures of the skin on the right and left hands of 10 men. The temperatures for each man in degrees centigrade are given in Table 13.6.[1] Is there

TABLE 13.6.—SKIN TEMPERATURES ON EACH HAND OF TEN MEN

Man Number	Temperature of	
	Right Hand	Left Hand
1	25.9	25.4
2	32.3	32.2
3	26.4	25.5
4	32.6	31.3
5	32.7	33.5
6	24.6	25.7
7	32.4	32.3
8	25.5	25.1
9	28.6	27.7
10	30.0	29.7

any relationship between the temperature of the right hand and that of the left? We can tell roughly if we plot these data on cross-section paper, letting the horizontal axis represent the scale of right-hand temperatures and the vertical axis the scale of left-hand temperatures. First we lay off the scales, noting that the right-hand temperatures vary from 24.6 to 32.7 and the left-hand temperatures from 25.1 to 33.5. We then locate on the chart a point for each man, letting its abscissa (horizontal position) represent the temperature of the man's right hand, and its ordinate (height) represent the temperature of his left hand. There will, then, be as many points as there are men, and each

[1] F. G. BENEDICT, *et al.*, Human Vitality and Efficiency under Prolonged Restricted Diet, *Carnegie Institution of Washington Publication* 280, p. 250.

point will represent a combination of right-hand temperature and left-hand temperature. This procedure gives us Fig. 13.1. The dot which appears toward the right and nearest the top is the dot representing man number 5, whose left-hand temperature was highest. Each of the other dots represents one man.

Now the noticeable feature about these dots on the scattergram is that they seem to be arranged in a band running from the lower left of the diagram to the upper right. They are not placed

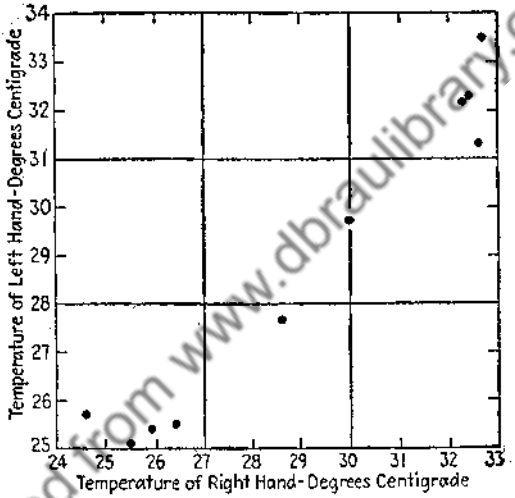


FIG. 13.1.—Scatter diagram of skin temperatures on right and left hands of ten men. Data from page 379.

haphazardly on the paper, as they would be if they had been shaken out of a saltcellar. It becomes evident that whenever the temperature of the right hand was high (as would be shown by a position toward the right in the diagram) the temperature of the left hand tended to be high also (as is shown by a position toward the top of the chart). If after looking at the diagram you were asked to guess at the left-hand temperature of a man whose right-hand temperature was 29°C., you would not neglect the right-hand temperature in making your estimate. In fact, you would probably estimate his left-hand temperature at about 28°C. or a little higher.[1]

[1] The student should note that, although the left-hand temperature and the right-hand temperature are related (that is, a knowledge of the temperature of one hand helps us to estimate the temperature of the other hand),

We have records made the same day of the strength of grip in the right hand of these same men, measured in kilograms. The figures are given in Table 13.7.[1] Let us plot a new scatter

TABLE 13.7.—STRENGTH OF GRIP OF TEN MEN

Man Number	Strength of Grip (kilograms)
1	52.7
2	52.1
3	55.7
4	51.2
5	52.4
6	45.0
7	44.0
8	43.1
9	48.9
10	45.5

diagram with the right-hand temperatures still as abscissas but with the strengths of grip as ordinates (see Fig. 13.2). Here there seems to be no plan at all in the arrangement of the dots; they seem to be scattered quite by chance. If you were asked to estimate the strength of grip of a man whose right-hand temperature is 25.5, what strength of grip would you select? You would get no help from the scatter diagram, except that it would tell you the approximate range of strengths of grip. You would have to guess that the unknown man had an average strength of grip, unless you could get some further information about him. The fact that you knew his hand temperature would be of no help to you.

We see, then, that, when the points on a scatter diagram fall along a definite band, there is evidence that relationship exists

we do not conclude that the right-hand temperature is *caused by* the left-hand temperature, or vice versa. In other words, when we say that the two things are related we do not imply any direct causal connection between them. In the present case it is altogether likely that each of the temperatures is the result of some common outside cause or group of causes. One would have to come to conclusions with regard to cause and effect on other than statistical grounds.

[1] BENEDICT *et al.*, *op. cit.*, p. 583.

between the variables. Under such circumstances the most likely value of one variable depends on the value of the other variable, as the most likely height of a child depends on (or varies with) his age and the most likely temperature of the left hand varies with the temperature of the right hand. The two variables change together. It may be that when one rises the other rises, as in the case of hand temperatures. It may be that when one rises the other falls, as we discovered in the relationship between

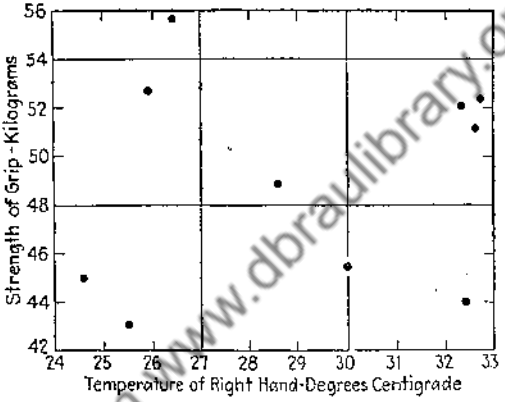


FIG. 13.2.—Skin temperatures and strengths of grip of ten men, showing practical absence of relationship.

income and sickness rates. In the former case we should say that the relationship is *positive* or *direct*. In the latter case we should say that the relationship is *negative* or *inverse*. And in either case we should say that there is *covariance* (since the values of the two variables tend to vary together) or *correlation* (since there seems to be some relationship between the variables). If, then, we say that there is positive correlation between two variables, we mean:

1. A knowledge of the value of one is helpful if we wish to estimate the value of the other. This is correlation.

2. Large values of the one tend to be associated with large values of the other, and small values with small values. This makes the correlation positive.

If we say that there is a negative correlation between two variables, we mean that large values of the one tend to be associated with small values of the other. A positive correlation shows no greater and no less relationship than a negative correlation. The

adjectives "positive" and "negative" refer to the direction of the relationship, not the degree of relationship.

13.4. The Regression Line.—We have seen that the scatter diagram which shows the relationship of right-hand temperature to left-hand temperature would be helpful when we are estimating the one from the other. Suppose we know that a man's right-hand temperature is 31.5°C. and we wish to estimate his left-hand temperature. We look at the scatter diagram and try to get an idea of the way in which left-hand temperature varies as we change right-hand temperature. We let our eye run along the band of dots and make mental note of the point at which the band seems to cross the vertical line representing a right-hand temperature of 31.5°C. We may even draw a line on the diagram to aid us in our computation. If we think that the relationship pictured can be described adequately by a straight line, we may draw a line through the data, trying its position first by stretching a string across the scatter diagram. We might even follow the method suggested for drawing the freehand trend (see page 277), that of computing the two averages and having the line pass through the point of the two averages. Reference to the table on page 379 shows that the average right-hand temperature was 29.1°C. and that the average left-hand temperature was 28.8°C. We could locate on the scatter diagram the point that represents a right-hand temperature of 29.1°C. and a left-hand temperature of 28.8°C. and pass our freehand line through this point. Figure 13.3 shows such a freehand line passed through the data. With the aid of such a diagram the work of estimating is, of course, much quicker.

We could, of course, determine the equation of this freehand estimating line if we wished by the methods outlined under the description of fitting trends by selected points (see page 278). This would involve the selection of two points on the freehand line of estimate. Let us call the variable which is represented on the horizontal scale (right-hand temperature) X, and the variable that is represented vertically (left-hand temperature) Y. In statistical work it is always customary to let the letter X represent the horizontal scale and to call the variable there represented the *independent variable*, while Y always stands for the variable which is represented on the vertical scale, and it is called the *dependent variable*. In this case, then, X = right-hand

temperature = the independent variable, and Y = left-hand temperature = the dependent variable. If we are trying to estimate the value of one variable from that of another, the one we are trying to estimate is always the dependent and is always placed on the vertical scale, while the one on which we are trying to base our estimates is always the independent and is always placed on the horizontal scale.

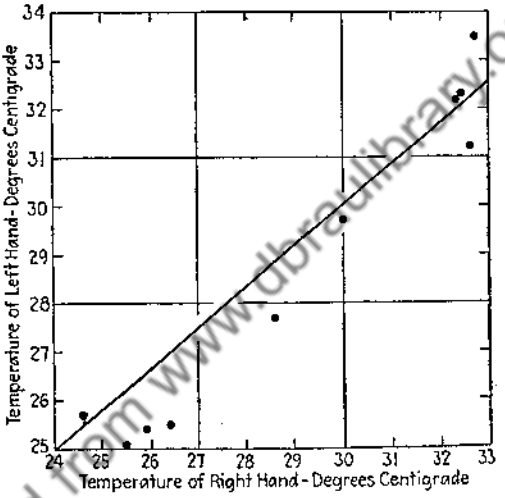


FIG. 13.3.—Freehand regression line describing the relationship of hand temperatures.

Suppose we select the points on the freehand estimating line which correspond to right-hand temperatures of 25 and 32°C. We let $X = 25$ in one case and $X = 32$ in the other. We note the values of Y at these points on the estimating line, and see that when $X = 25$, $Y = 25.8$; and when $X = 32$, $Y = 31.8$. Our pairs of values are, then,

$$\begin{array}{ll} X = 25 & X = 32 \\ Y = 25.8 & Y = 31.8 \end{array}$$

We substitute these two pairs of values in the equation for a straight line ($Y = a + bX$) and get

$$\begin{aligned} 25.8 &= a + 25b \\ 31.8 &= a + 32b \end{aligned}$$

Solving these two observation equations simultaneously, we have the following values of a and b:

$$a = +4.37$$
$$b = +0.857$$

These values are then substituted in the type equation to give the equation of this particular straight line, which is

$$Y = 4.37 + 0.857X$$

These values tell us that when the right hand has a temperature of 0°C. the left hand will have a temperature of 4.37°C., and that for each 1° rise in the temperature of the right hand the left-hand temperature will rise 0.857°. Thus we see that the methods we met when studying historical series and the fitting of trends are likewise applicable in describing the relationship between two variables when neither of the variables is time.

When we have a case such as this one, where we fit a line to a scatter diagram in an attempt to describe the relationship between two variables, we call the line a *regression line* and the equation which describes it a *regression equation.* The line was so named by a great English statistician, Francis Galton. He was very much interested in studies of inheritance, and one of his pioneering studies in the field of correlation was a study of the relationship between the heights of fathers and the heights of their sons. He found that tall fathers tended on the average to have tall sons, but that on the average the sons did not tend to differ from the mean by so much as their fathers. Hence he came to speak about the tendency for sons' heights to "regress" toward the mean, and the line which pictured the relationship of the heights of fathers and sons he called the regression line.

13.5. Least-squares Regression Lines.—The regression line which we have just fitted to describe the relationship of the temperatures of the hands is open to the same criticism that we made of the freehand trend fitted to historical data. No two people would draw exactly the same line, and moreover it can be shown that, if the scatter around the line is a chance affair, with the residuals from the line in a normal distribution, there is one straight line which is more likely to be the line of true relationship

than any other. We have discovered already that this is the line the sum of the squared deviations from which is a minimum.[1]

We have discovered that one can minimize the sum of the squared vertical deviations by fitting a line whose equation contains values of a and b found by solving two normal equations. We have demonstrated the fact[2] that, if the values of a and b are determined by solving these two normal equations, these values will give a straight line the sum of the squared deviations from which will be less than the sum of those from any other straight line. These equations are

$$Na + b\Sigma X = \Sigma Y$$
$$a\Sigma X + b\Sigma X^2 = \Sigma XY$$

Let us compute these values for the problem of hand temperatures and determine the best-fitting straight line. We can organize the work as in Table 13.8. The last column we might omit for our

TABLE 13.8.—COMPUTATION OF LINEAR REGRESSION LINE, TEMPERATURES OF RIGHT AND LEFT HANDS

Right-hand Temperature (X)	Left-hand Temperature (Y)	(XY)	(X^2)	(Y^2)
25.9	25.4	657.86	670.81	645.16
32.3	32.2	1040.06	1043.29	1036.84
26.4	25.5	673.20	696.96	650.25
32.6	31.3	1020.38	1062.76	979.69
32.7	33.5	1095.45	1069.29	1122.25
24.6	25.7	632.22	605.16	660.49
32.4	32.3	1046.52	1049.76	1043.29
25.5	25.1	640.05	650.25	630.01
28.6	27.7	792.22	817.96	767.29
30.0	29.7	891.00	900.00	882.09
291.0	288.4	8488.96	8566.24	8417.36

present purposes, but we shall need it later, and it is customary to compute it while we are computing the regression constants.

[1] It may pay the student at this point to review the statement relative to least squares which was made on pp. 291*ff*.

[2] See p. 296*n*.

For the moment we shall neglect it. We note that we now have the following values:

$$N = 10 \text{ (the 10 pairs of values)}$$
$$\Sigma X = 291.0 \qquad \Sigma Y = 288.4$$
$$\Sigma X^2 = 8566.24 \qquad \Sigma(XY) = 8488.96$$

If we substitute these values in our normal equations, they become

$$10a + 291.0b = 288.4$$
$$291.0a + 8566.24b = 8488.96$$

Solving these two equations simultaneously for a and b, we have

$$a = +0.235$$
$$b = +0.983$$

Substituting these values in our type equation for a straight line ($Y = a + bX$), we get the equation of the straight line which best fits the points on our scatter diagram. It is

$$Y = 0.235 + 0.983X$$

This is the regression equation, and the value of b (which is here equal to 0.983) is called the *regression coefficient* or the *coefficient of regression*. The value of a tells us that, when the right-hand temperature is 0, the left-hand temperature will be 0.235°C. The regression coefficient tells us that each increase of 1° in the temperature of the right hand will be accompanied by an increase of 0.983°C. in the temperature of the left hand. In other words, the regression coefficient tells us the number of units of change in the dependent variable which will accompany a change of one unit in the independent variable.

As we have seen, one can choose the dependent and the independent variables to suit oneself. If we wished we could estimate right-hand temperatures from the left-hand temperatures. We should not do this, however, with the same equation. It would be necessary to go back to our original data and compute another regression equation in which the left-hand temperatures were independent and the right-hand temperatures dependent. This would give us a new equation with a different value of b. To distinguish between the two regression coefficients we call the one computed above (when Y was the dependent variable) the

regression of Y on X and denote it by the symbol b_{yx}. The *regression of X on Y* (used when we are estimating values of X from values of Y) is denoted by b_{xy}. Thus the letter b stands for the regression coefficient, and when it is followed by subscripts the first subscript always denotes the dependent variable.

The values of b_{yx} and b_{xy} will not usually be the same. The two regression lines will not ordinarily coincide. In fact, they will coincide only if all the points on the scatter diagram fall on the

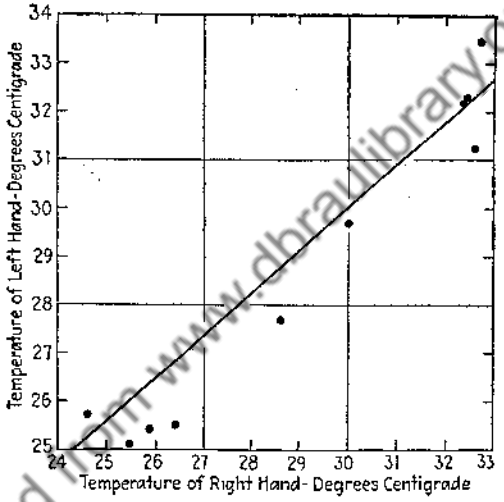


FIG. 13.4.—Relationship of temperature of left hand to temperature of right hand, as shown by the least-squares regression line.

line. (They coincide only if we have what we shall later come to call "perfect correlation.")

If we wish now to estimate the temperature of the left hand of a man whose right-hand temperature is 32°C., we go about it by substituting 32 for X in the regression equation and solving. Thus

$$Y = 0.235 + 0.983(32) = 31.691$$

We round this off to 31.7°C. and say that when the right-hand temperature is 32°C. we expect a left-hand temperature of 31.7°C. We can, of course, draw this regression line on our scatter diagram by locating a point near each end and connecting these points by a straight line. We have already located one point, which would be at $X = 32$ and $Y = 31.7$. We could now estimate the left-

hand temperature (Y) for a right-hand temperature of, say, 25°C. We find the estimated value of Y is 25.8. We can now locate this point on the scatter diagram and connect the two estimated points by a straight line. This will show us graphically the location of the regression line, and we can now estimate by means of the graph or by means of the equation as we choose (see Fig. 13.4).

13.6. Errors of Estimate.—We have said that if we wish to estimate the temperature of a man's left hand it will help us to know the temperature of his right hand. How much will this help us? How much more accurate will our estimates be if they are made from the regression equation than if they are made without it? The answer is very simple, and depends on methods which we have already covered. If we can compute the amount of each error we should make in estimating the values from the regression equation, it is easy to determine the scatter or spread of the errors by computing their standard deviation. If we can also compute the amount of the errors we should make without using the regression equation, we can find the standard deviation of these errors. A comparison of the two standard deviations will tell us whether the errors are smaller when we estimate by the regression equation than they are when we do not.

First suppose that you had no regression equation or you did not know the man's right-hand temperature, and that you were asked to estimate the left-hand temperature. What figure would you give as your estimate? Surely you would not make an estimate of 100°C. or of 5°C. In fact, in the absence of other information, your best chance would be to estimate an average left-hand temperature, that is, a temperature of 28.8°C. As was pointed out in Exercise 13, page 321, the sum of the squared deviations from the mean is smaller than from any other value, so that the dispersion will be less when measured around the mean than when measured in any other way. For this reason if you have to estimate a man's left-hand temperature without any knowledge of other correlated factors, your best possible estimate is the mean of left-hand temperatures.[1]

[1] We can easily show that the sum of the squared deviations is a minimum when taken about the mean—that the arithmetic mean is fitted to a distribution by the method of least squares. The proof follows that used in deriving the normal equations for the straight line on p. 296. We have a series of values, each of which is represented by the letter X. We wish to

Suppose we do estimate that each of these men will have an average left-hand temperature; what will be the amount of the errors? This will depend, as we are aware, on the scatter or dispersion of the actual left-hand temperatures. We can discover it by computing the standard deviation of the left-hand temperatures. We know that

$$\sigma = \sqrt{\frac{\Sigma(X^2)}{N} - \bar{X}^2}$$

This is the formula we used on page 138 to compute the standard deviation. In this case we are calling the variable in which we are interested Y instead of X, so we can restate the formula thus:

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - \bar{Y}^2}$$

We know that $\bar{Y} = 28.8$ (see page 383), and in the table on page 386 we find that $\Sigma(Y^2) = 8417.36$. Substituting these values in

---

find the one most probable value, a, from which the sum of the squared deviations is least. If we represent the differences between the various values and this most probable value by d, then for each value of X there is a value of d which is defined thus:

$$X + d = a$$
$$d = a - X$$
$$d^2 = (a - X)^2$$

If we sum all these terms, we get

$$\Sigma d^2 = \Sigma(a - X)^2$$

This is the function to be minimized, and we shall represent it by f. We minimize by taking the partial derivative with respect to a and setting this equal to 0, thus:

$$\frac{\partial f}{\partial a} = 2\Sigma(a - X) = 0$$

Dividing by 2 and carrying out the summation,

$$Na - \Sigma X = 0$$
$$Na = \Sigma X$$
$$a = \frac{\Sigma X}{N}$$

But $\Sigma X/N$ is the arithmetic mean; hence the arithmetic mean is the value the sum of the squared deviations from which is a minimum.

the formula, we find

$$\sigma_y = \sqrt{\frac{8417.86}{10} - 28.8^2} = 3.51$$

The standard deviation of left-hand temperatures is, then, 3.51°C. This figure answers our question as to how much error we shall have if we estimate all the temperatures at the average.

We know that, insofar as this distribution is normal, 68.27 per cent of the men will have hand temperatures within 3.51°C. of our estimate (since we estimated at the average), 95.5 per cent of them will be within 7.02°C. of our estimate, and practically everyone will be within 10.53°C. of our estimate.[1] These figures measure the amount of the error in case we do not use the regression equation but make all our estimates at the average.

Suppose, now, that we do use the regression equation in making our estimates. Under these circumstances how much error will we have? We can discover it by computing the expected hand temperature for each man and comparing it with the actual hand temperature. This we do by substituting each man's right-hand temperature in the regression equation and solving for the left-hand temperature. For example, the first man had a right-hand temperature (as shown in the table on page 386) of 25.9°C. Substituting this figure for X in the regression equation, we have

$$Y = 0.235 + 0.983(25.9) = 25.7°\text{C.}$$

The estimate of left-hand temperature is 25.7°C., but inspection of the table shows an actual left-hand temperature of 25.4°C. The error is, then, $25.4° - 25.7° = -0.3°$C.; that is, we define our residuals as the actual value minus the estimated value. Our actual values we have called Y, and if we call our estimated values Y' the deviation is defined as

$$Y - Y' = d$$

In the case of the first man, $d = -0.3°$C. If we compute the error in the case of each of the other nine men in the same way, we find the errors indicated in Table 13.9. The average error is

[1] These are the figures for the percentage of the cases which will fall within various numbers of standard deviations from the mean, as was pointed out on p. 146.

TABLE 13.9.—COMPUTATION OF ERRORS OF ESTIMATE

Man Number	Error (degrees centigrade)	Squared Error
1	−0.3	0.09
2	0.2	0.04
3	−0.7	0.49
4	−1.0	1.00
5	1.1	1.21
6	1.3	1.69
7	0.2	0.04
8	−0.2	0.04
9	−0.6	0.36
10	0.0	0.00
Totals	0.0	4.96

zero, as it always must be when we measure the errors around the least squares line.[1] The standard deviation of the errors is, therefore,

$$\sqrt{\frac{4.96}{10}} = 0.704°\text{C}.$$

This tells us that if we use the regression equation as the basis for our estimates, 68.27 per cent of our estimates will be within 0.704° of the correct figure, 95.5 per cent will be within 1.408° of

[1] From the second equation of the footnote, p. 296, we have

$$d = a + bX - Y$$

Hence

$$\Sigma d = na + b\Sigma X - \Sigma Y$$

But the next to the last equation, footnote, p. 297, tells us that

$$na + b\Sigma X - \Sigma Y = 0$$

Hence

$$\Sigma d = 0$$

Since the sum of the deviations is zero, their arithmetic mean must also be zero. Therefore the usual formula for the standard deviation (p. 138) becomes

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

where d is a deviation of the actual from the estimated value.

the correct figure, and practically never will we make an error greater than 2.112°C. When we tried making our estimates without the regression equation (page 391), we found that the standard deviation of the temperatures was 3.51°C. Without using the equation, we had a range of error of 3.51° to include 68.27 per cent of the men. With the regression equation, we can cut this range down to 0.704°. We can say that the error in our estimates has been cut down to 0.704/3.51 or 20 per cent of its former amount. We have eliminated 80 per cent of the error by using the regression equation. To be sure, 20 per cent of the error remains. By using the regression equation, we calculated but one man's left-hand temperature exactly—that of man number 10. But our errors are smaller on the average than they would have been had we made our estimates with no knowledge of the right-hand temperature.

We have just computed the standard deviation of the errors of estimate—the standard deviation of the residuals around the regression line. We know that the sum of the squared residuals (that is, errors) is less around this line than around any other straight line, so that the standard deviation of the residuals, being based on the sum of the squared residuals, must also be smaller around this line than around any other straight line. This standard deviation of our errors of estimate we call the *standard error of estimate*, and we represent it by the letter S. In this case we have been estimating values of Y from values of X, and hence we can call this particular standard error the standard error of estimating Y from X. To distinguish it from other standard errors of estimate, we give it subscripts, thus: $S_{y \cdot x}$. This is read, "The standard error of estimating Y from X." Where it will cause no confusion, we commonly use the symbol S_y alone.

13.7. Correlation.—We have seen that it is helpful to compare the values of S_y and σ_y. In the case we have just discussed, the value of σ_y was 3.51 and the value of S_y was 0.704. Thus we knew that we could estimate by means of the regression equation as many of the cases within 0.704 of the actual figures as we could estimate within 3.51 of those figures without the regression equation. It is evident that if a knowledge of the right-hand temperature were of no use to us in estimating the value of the left-hand temperature, then S_y would be as great as σ_y. Whenever S_y is smaller than σ_y, there has been some advantage in using the regression

equation. In other words, a knowledge of the one variable has improved our estimates of the value of the other variable. But this is just what we said determines the existence of relationship between variables (see page 374). We say that variables are related when a knowledge of the value of one helps us in estimating the value of the other. And when S_y is smaller than σ_y we know that a knowledge of the value of one variable does help us in estimating the value of the other. Hence we can say that if $S_y = \sigma_y$ there is no relationship—no correlation between the variables. But if S_y is smaller than σ_y correlation exists.

This at once leads us to the conclusion that we might measure the degree of relationship by the difference between S_y and σ_y. Yet here there is an immediate difficulty. Suppose we are measuring the relationship of height to weight, estimating the latter from the former. S_y and σ_y will both be in terms of pounds, and the difference between them will be in pounds. Had our original figures been in ounces, the value of S_y and σ_y would both have been 16 times as great, and the difference between them would likewise have been 16 times as great. Yet there would be no more relationship between height and weight when the weights were stated in ounces than when they were stated in pounds. Obviously we need some measure of the degree of relationship which is independent of the units in which the problem is stated.

This leads us to suggest another possible measure of correlation, namely, the ratio of S_y to σ_y. Both S_y and σ_y will always be in the same units, one of them being in pounds if the other is. Multiplying both by 16 will not affect their ratio. Hence we could use as a measure of the degree of correlation their ratio, or S_y/σ_y. In practice, statisticians have preferred to use the ratio of the squares of these numbers, that is, S_y^2/σ_y^2; in other words, they compare the square of the standard deviation of the errors of estimate with the square of the standard deviation of the original figures.[1]

But one minor difficulty still exists with this measure. The greater the relationship between the variables, the smaller will be S_y as compared with σ_y. If we could estimate with perfect accuracy, making no error in any case, it would mean that all the

[1] The square of the standard deviation, as well as the standard deviation itself, is sometimes used as a measure of dispersion. It is called the *variance*. Thus, σ_y^2 = the variance of Y.

points on our scatter diagram fell on the regression line. There would be no errors of estimate, and S_y would equal 0. On the other hand, if there were absolutely no relationship between our two variables, S_y would be as great as σ_y. These are the limiting cases. Perfect correlation would mean that S_y^2/σ_y^2 was equal to 0, and the entire lack of correlation would mean that S_y^2/σ_y^2 was equal to 1. But it is preferable to have a measure which shows the amount of the relationship directly by its size. We should like to have close relationships expressed by large coefficients and lack of relationship expressed by small coefficients. This can be done by subtracting our ratio from 1, to get $1 - (S_y^2/\sigma_y^2)$. And we usually take the square root of this to put it back into the first degree, since S_y and σ_y were originally squared. In this case our measure of correlation will be defined by the formula

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

The letter r always represents the coefficient of correlation around a straight line fitted by the method of least squares. In other words, if we are making our estimates from a line of the form $Y = a + bX$, and if this line was fitted by the method of least squares, then r is defined as indicated and is known as the *coefficient of simple linear correlation.*[1] People often speak of it as the coefficient of correlation, without bothering with the rest of the title. It is true that there are many other types of coefficients of correlation, and it is safer to specify the one in question. If one uses the letter r, however, it is always understood that this is the coefficient to which reference is made.

We have already computed all the values necessary to find the value of r. If we substitute them in the foregoing formula, we find that

$$r = \sqrt{1 - \frac{0.704^2}{3.51^2}} = \pm 0.980$$

Since S_y can never be larger than σ_y, it is evident that the ratio S_y^2/σ_y^2 can never be greater than unity. Hence the value of unity minus this ratio can never be negative, and the square root

[1] This coefficient is also commonly known as the Pearsonian coefficient of correlation after the great English statistician, Karl Pearson, who did much of the early work in the field of correlation.

can always be taken. As when any other square root is taken, the result may be either positive or negative; that is, $\sqrt{4} = \pm 2$. Here, then, we can call our result $+0.980$ or -0.980. It is the accepted practice of statisticians to give to r the same sign that is found for b in the regression equation. In this case we found that $b = +0.983$ (see page 387).[1] Hence we give r a plus sign and say

$$r = +0.980$$

The sign means that the relationship is direct or positive as these terms were defined on page 382.

The smallest possible value of r would occur if there were no relationship at all between the variables—that is, if a knowledge of one helped us not at all in estimating the value of the other. In this case S_y would be as great as σ_y (there would be as much error in using the regression equation as in neglecting it), and hence S_y/σ_y would be equal to 1. Hence

$$r = \sqrt{1 - 1} = 0$$

When there is no relationship between the variables, then, $r = 0$.

The largest possible value of r would occur if the relationship between the variables were perfect—if the points on the scatter diagram all fell on the straight regression line, so that we could estimate each dot with entire accuracy from the regression equation. In this case the errors of estimate would be nonexistent. S_y would equal 0, and hence S_y/σ_y would equal 0. This would give us

$$r = \sqrt{1 - 0} = \pm 1$$

When there is a perfect linear relationship between two variables, $r = \pm 1$. If large values of the one variable are associated with large values of the other and the relationship is perfect, $r = +1$. If large values of the one variable are associated with small values of the other and the relationship is perfect, $r = -1$. These are the limiting values of the coefficient. A value of r beyond the range from -1 to $+1$ means a mistake in computation. Such a coefficient cannot exist in fact, on account of the nature of the definition.

[1] In this case the value of b practically coincides with the value of r. There is no necessary relation between their values except that their signs are always the same. See footnote on p. 405 also.

The student should be warned at this point that a value for r of 0 does not mean that no relationship exists. We have said that if no relationship exists the value of r will be 0, but the statement cannot be turned around and still retain its validity. The coefficient of simple linear correlation measures the degree to which a straight line describes the relationship between the variables. It is quite possible for a close relationship to exist between two variables and to be nonlinear in nature. Attention is again called to the table on page 378 in which one can see a relationship between the variables but in which the relationship cannot be described by a straight line. In such a case the value of r might be very small, and even equal to 0. This would mean merely that if there were any relationship it could not be described by a straight line. Let us make, then, the two statements that can be correctly made, so that the difference between them can be noted:

1. If there is no relationship between two variables, the value of r will be 0.
2. If the value of r between two variables is 0, either there is no relationship between them or whatever relationship exists cannot be described by a straight line.

Failure to realize the qualification in the second statement has led many a careless person to state that no relationship existed in his researches when, as a matter of fact, his researches had not proved the nonexistence of relationship. Computation of the coefficient of correlation can show that relationship (as herein defined[1]) does exist, but it cannot show that it does not exist. A low coefficient of correlation shows merely that if a relationship does exist one has not yet found it.[2]

13.8. Corrections for Small Samples.—We have seen that, when standard errors are computed on the basis of small numbers

[1] See p. 374.

[2] As Charles Kingsley says in his book "Water Babies," "There are no such things as water-babies? How do you know that? Have you been there to see? And if you had been there to see, and had seen none, that would not prove that there were none. If Mr. Garth does not find a fox in Eversley Wood . . . that does not prove that there are no such things as foxes. . . . And no one has a right to say that no water-babies exist, till they have seen no water-babies existing; which is quite a different thing, mind, from not seeing water-babies. . . ."

We can paraphrase Mr. Kingsley: "And no one has a right to say that no relationship exists, till they have seen no relationship existing: which is quite a different thing, mind, from not seeing relationship."

of cases, it is necessary to adjust them somewhat in order to approximate the correct value.[1] The usual formulas are built on the assumption that they will be applied only when the number of cases is large. The same is true of the formulas we have given for r and S_y. In our illustrative problem regarding hand temperatures we have computed the values from 10 cases. Having studied but 10 men, we have tried to reach some conclusion with regard to the relationship between the temperatures of the hands. In practice a statistician would insist on having many more cases than this before he drew conclusions. Our problem has been greatly oversimplified to make it easy for the student to follow the steps involved.

When r and S_y are computed from small samples, there is a tendency for r to be greater than the actual r of the universe and for S_y to be smaller than the true S_y of the universe.[2] In other words, we overestimate the amount of the relationship and underestimate the amount of our errors of estimate. Fortunately it is possible to make a rough correction for this condition. If we let S_y denote the standard error of estimate which has been corrected for the number of cases, and if we let r' denote the coefficient of simple linear correlation which has been corrected for the number of cases, we have the following formulas, which may be used in correcting:

$$(S_y')^2 = (S_y)^2 \left(\frac{n-1}{n-2}\right)$$

$$(r')^2 = 1 - (1 - r^2)\left(\frac{n-1}{n-2}\right)$$

In our illustrative problem we have the following results:

$S_y = 0.704°\text{C.}$ (page 392)
$r = +0.980$ (page 396)
$n = 10$ (the number of cases)

[1] See p. 254.

[2] If we went to the extreme of computing r and S_y from but two cases, it is obvious that a straight line would pass through both points on the scatter diagram and the results would always, of mathematical necessity, show perfect correlation with no error of estimate. This would be true even if the variables from which the two observations were chosen were entirely unrelated.

If we substitute these values in the formula above, we get the following corrected values:

$$(S'_y)^2 = (0.704^2)\left(\frac{9}{8}\right) = 0.558$$

$$S'_y = \sqrt{0.558} = 0.747$$

$$(r')^2 = 1 - (1 - 0.980^2)\left(\frac{n-1}{n-2}\right) = 0.955$$

$$r' = \sqrt{0.955} = \pm 0.977$$

Since r was positive, we make r' positive and call it $+0.977$; that is, r' always has the same sign as r. If the value of $(r')^2$ as computed by the formula turns out to be negative, the value of r' should be given as 0.

It will be noted that in this case the application of the correction formulas served to increase the standard error of estimate from 0.704 to 0.747°C. and to decrease the coefficient of correlation from $+0.980$ to $+0.977$. These changes may seem negligible. Yet the increase in the standard error of estimate was over 6 per cent, and in problems where r was not so great the correction would be greater. Whether or not the change in r was important, we can judge better after we have learned more about the interpretation of r (see page 414).

13.9. Standard Error of Correlation Coefficient.—In Chap. IX we learned that, although one can calculate the average of a group of figures, one can seldom calculate the average of the universe in which one is interested. We can, however, estimate the standard deviation of the means of samples drawn from the universe. This we called the standard error of the mean. Similarly we should like to compute the standard error of the coefficient of simple linear correlation; that is, we should like to estimate the standard deviation of the r's of an infinite number of samples drawn from the same universe. In our sample we found that $r = +0.977$. Had we selected another sample of 10 men, it is probable that we should have found a slightly different value of r. Had we selected 100 such samples, we should have found many values of r. Would these values have been widely scattered, or would they all have fallen close to the value $+0.977$ which we found? The standard error of r would

be our estimate of the dispersion which would occur in the r's of these many samples.

Unfortunately, however, it has now been shown that if we select many samples from the same universe and compute the value of r for each sample, the values we get will not be normally distributed unless the size of the sample is very large and the degree of relationship (the size of r) moderate or small. With small samples or with large values of r, the distributions are badly skewed, and the shape of the distributions varies radically for different values of r. Such non-normal distributions cannot be accurately described by standard deviations, and consequently the standard error of r would give us only a very rough idea of the distribution unless the sample was large and the value of r small. Textbooks usually give the formula for finding this standard error as follows:

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

If we were to apply the formula to our problem of hand temperatures, substituting $+0.977$ for r and 10 for N, we should get

$$\sigma_r = \frac{1 - 0.977^2}{\sqrt{10}} = 0.0145$$

However, our present example is very poorly adapted for this formula, both because of its small size ($N = 10$) and because of the large size of the relationship ($r = 0.977$). Therefore we should not rely on the standard error just computed.

13.10. The z-transformation.—In spite of the difficulties that stand in the way of using the standard error of r, we can fortunately transform any given r in such a way as to derive a coefficient which is normally distributed even in small samples and even when r is large. This method, worked out by R. A. Fisher,[1] seems involved at first glance, but experiment with an example or two shows that its application is in fact very simple. To apply the method we must first find the new function, z, as follows:

$$2z = \log_e (1 + r) - \log_e (1 - r)$$

[1] For a more complete discussion of problems of small samples, the reader should consult R. A. Fisher, "Statistical Methods for Research Workers," Oliver & Boyd, London, 1932.

This formula would prove troublesome in practice to students who were not familiar with natural logarithms. It can be stated in terms of common logarithms thus:

$$z = 1.1513 \log_{10} \left(\frac{1 + r}{1 - r} \right)$$

To apply this to our case of hand temperatures, where $r = +0.977$, we get the following:

$$\begin{aligned} 1 + r &= 1.977 \\ 1 - r &= 0.023 \\ \frac{1 + r}{1 - r} &= 85.96 \\ \log 85.96 &= 1.93430 \\ (1.1513)(1.93430) &= 2.227 = z \end{aligned}$$

Fortunately, however, we do not even have to bother with this calculation, because the values of z for most common values of r have been computed and tabulated. By consulting the table in Appendix VI we can find immediately the value of z corresponding to almost any desired value of r, or the value of r for almost any value of z. In the present case the table shows us that when $r = 0.977$, z falls between 2.22 and 2.23.

Fortunately this coefficient which we call z seems to be normally distributed (or approximately so) for samples drawn from almost every kind of universe. Hence the standard error of this measure will give us an accurate description of its distribution. The standard error of z is easily found from the simple formula

$$\sigma_z = \frac{1}{\sqrt{n - 3}}$$

In our present problem, since $n = 10$, we get

$$\sigma_z = \frac{1}{\sqrt{7}} = 0.3779$$

We can now interpret the significance of our correlation as follows: In our sample we found a z value of 2.227. We should almost never expect the value of z in any sample drawn at random from a given universe to differ from the actual z value of the universe by more than $3\sigma_z$. Therefore we feel confident that the z value of the universe is within $3\sigma_z = 1.134$ of the z value which

we have found, 2.227. Hence the z value of the universe almost certainly lies between $2.227 - 1.134$ and $2.227 + 1.134$; or, between 1.093 and 3.361. Referring again to Appendix VI, we find that these values of z correspond to correlation coefficients of 0.798 on the one hand and well above 0.995 on the other. Computation shows that the upper limit, corresponding to a z value of 3.36, is an r value of approximately 0.998. Therefore we conclude that, although in our sample $r = +0.977$, in the universe from which this sample was drawn the relationship almost certainly fell somewhere above +0.798. Two-thirds of the z's of other samples would be expected to fall within σ_z of 2.227; or, between 1.85 and 2.60. We find from Appendix VI that the chances are 2 to 1 that the value of r in the universe falls between 0.952 and 0.989.

It will be seen from our present example that the common method of computing r and its standard error (or its probable error) by the ordinary formula tends to exaggerate greatly the reliability of the coefficient of correlation. One should, rather, perform the seemingly more complicated process (it is actually no more lengthy when tables are used) of transforming r values to z values and testing the significance of the results by means of the methods we have just described. To summarize the method:

1. Having found r and n from the sample, find the value of z corresponding to the computed value of r. (This is carried out directly from the table in Appendix VI.)

2. Compute the standard error of z by the formula

$$\sigma_z = \frac{1}{\sqrt{n-3}}$$

3. Lay off the area within which the z of the universe will almost certainly be found. The limits of this area are arbitrary, but we have been using three standard errors in each solution. Therefore we add $3\sigma_z$ to the value of z found above and subtract $3\sigma_z$ from the value of z found above to get our two limiting z values.

4. Find from the table in Appendix VI the r values corresponding to the two z values just discovered. These are the limiting r values, between which we feel reasonably certain that the r of the universe will be found.

13.11. Application of Correlation Results.—But what do we mean by the "universe"? The 10 men actually tested, together with all other men who might have been tested, constitute the universe. Since there were no women in our sample, we cannot

feel at all certain that the relationship which we have found would apply to hand temperatures of women or of groups containing both sexes. When we have set up limits between which we think that the relationship almost certainly falls, we have set limits which will apply only for other cases like those we have studied.

Note that our universe contained no children and no men in their dotage. Hence we cannot be sure that the relationship between the temperatures of the hands would be the same among babes or old men. The 10 men measured were all students at Springfield College. Is our universe, then, confined to Springfield College students? Or to Springfield College students who were in attendance on Jan. 26, 1918, when these tests were made? As a matter of fact the statistician never knows the answers to these questions; he has to approximate them as nearly as he can. In a case such as the present one he would be likely to apply his results to other men of about the same age. If the results did not seem to agree, he would try to find out why. But a statistician seldom knows for certain from what universe he has selected his sample.

In this connection the student should be warned against what Whipple calls[1] "the fallacy of concealed classification." As he points out, if we classify males by occupation we might find the death rate of bank presidents much higher than that of newsboys. This finding should not lead us to the conclusion that the difference in death rate is a result of the difference in occupation. We must remember that bankers and newsboys differ not only in the type of their work but also in age. It may well be that the difference in death rate is a natural result of the difference in age, and that if we were to make allowance for the age factor we should find banking a healthier occupation than the peddling of papers. In other words, having selected our samples of newsboys and bankers, we must examine the universes from which they are selected. Surely there are characteristics of the universe other than occupational ones.

Before we can apply our results with any certainty, we have to know within what universe they are applicable. Before we can apply the results of our study of hand temperatures, we have to know to whom the results can be applied. They can be applied

[1] G. C. WHIPPLE, "Vital Statistics," p. 290, John Wiley & Sons, Inc., New York, 1923.

correctly only to the universe from which the sample was drawn, that is, only to people who are like the people in the sample. And it is difficult to make certain that we have discovered all the hidden classifications in the sample.

Let us illustrate once more. Suppose that you wish to discover something about the amount of profit that there is in farming in a given locality. You cannot make studies of all the farms, so you decide to study a few and apply the results to the others. You drive a car around to the most accessible farms and get data from farmers who are willing to give them and who have them to give. You compute the average net income and the standard error of the mean, and you then apply it to all the farms in the county. But all the farms in the country are not in the same statistical universe as the ones that you studied. You chose the most accessible farms, and the chances are that their costs of getting supplies and of marketing their products differ from those of other farms because of their accessibility. You chose farmers who had kept financial records and who were, therefore, able to give you the information which you desired. Such farmers are probably above the average in intelligence and initiative; otherwise they would not know how to keep accounts and would not bother to do so. Thus, since you have selected a sample of the most accessible farms operated by the most intelligent and energetic farmers, the conclusions cannot be applied with safety to farms in general. In this case there are "concealed classifications" in your data which might have led you to mistake the characteristics of the universe. Your results can be applied only to the universe consisting of all farms like those studied—that is, to all accessible farms in the locality which are operated by intelligent, resourceful men.

13.12. Actual Computations.—The methods we have developed for computing r and S_y have been selected for presentation because they help to show the meaning of correlation. They are not the methods which would ordinarily be used in practice, because other methods require less arithmetical work. We shall, then, present other formulas which make the work of computation easier and which give the same results, and for purposes of illustration we shall work out a type problem by these simpler methods.

First we give four formulas which are algebraically equal and from any one of which one can compute the value of r. The first

is the formula with which we are already familiar. In these formulas the symbols used are the same as those we have used before except that the mean of the X's is denoted by M_x instead of by $\bar{X}$.[1]

$$(1) \qquad r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

$$(2) \qquad r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y}$$

$$(3) \qquad r = \frac{\Sigma(xy)}{\sqrt{(\Sigma x^2)(\Sigma y^2)}}$$

$$(4) \qquad r = \frac{\Sigma(XY) - N(M_x)(M_y)}{\sqrt{(\Sigma X^2 - NM_x^2)(\Sigma Y^2 - NM_y^2)}}$$

It will be noted that formulas (2) and (3) are stated in terms of deviations from the mean, as is shown by the small letters. If we were to compute r from formula (3), for example, we should compute the deviation of each X and of each Y from its average, multiply together these deviations, and sum the products. The result would be the numerator of the formula. We should then square the deviations and add them for the X's and for the Y's. We should multiply the two sums together and take the square root. The result would be the denominator. When the numerator was divided by the denominator, the result would be r. In this case the result may turn out to be either positive or negative. It is not necessary to give r the sign of b in the regression equation, because it will already have the correct sign as a result of the computations. In fact, this is true whether one uses formula (2), (3), or (4). It is only when we use formula (1) that we need to look to the sign of b to know the sign of r.

[1] The value of r may also be computed from the two coefficients of regression described on p. 387. The relationship is this

$$r = \sqrt{(b_{xy})(b_{yx})}$$

Obviously if the two regression coefficients are reciprocals of one another (as they would be if the two regression lines coincided), the value of r will be 1.00; that is, the regression line of Y on X and that of X on Y coincide when the correlation is perfect.

The above relationship makes it evident that if we know the value of b_{yx} and the value of r (as we usually do when we have carried out a correlation problem), the value of b_{xy} can be found from the formula without long computation.

Formula (2) is probably the best known of the formulas for r, yet in most cases it is not the easiest to use. Ordinarily formula (4) will give the result more quickly and with less mathematical drudgery than any other formula. When this formula is used, it should be remembered that M_y means the mean of the original Y's—not the mean of the deviations from the mean. The subscript is a small letter, but nevertheless it refers to the original Y's. Of course, the mean of the y's would be 0, as would the mean of the x's. The means of the Y's and of the X's are the values referred to.

Formula (4) gives simple, shorthand directions for the computation of r. When the directions are put into this form, they are much more concise than they could otherwise be. To make sure that they are not misunderstood (and at the same time to show what a saving there is in the statement as given in the formula), we can translate this mathematical statement of the directions into the following longer version, which directs what steps should be taken and in what order:

1. Multiply each X by the corresponding Y (letting X be the independent and Y the dependent variable, as described on page 384).
2. Add the products.
3. Compute the average of the X's.
4. Compute the average of the Y's.
5. Count the number of cases.
6. Multiply the answer to number 3 above by the answer to number 4, and multiply the product by the answer to number 5.
7. Subtract the answer to number 6 above from the answer to number 2. This will give the numerator of the formula, and is equal to $\Sigma(xy)$ of formulas (2) and (3).
8. Square each X and add the squares.
9. Multiply the answer to number 5 above by the square of the answer to number 3.
10. Subtract the answer to number 9 from the answer to number 8.
11. Square each Y and add the squares.
12. Multiply the answer to number 5 by the square of the answer to number 4.
13. Subtract the answer to number 12 from the answer to number 11.
14. Multiply together the answers to number 10 and number 13.
15. Take the square root of the answer to number 14. The result is the denominator of the formula.
16. Divide the answer to number 7 by the answer to number 15. The result is the value of r.

Note what long and complicated directions these make when written out even in this brief form, and contrast them with the simple, short, but identical directions of formula (4).

13.13. Computation of Regression Equations.—Common formulas for the regression equation are as follows:

$$(1) \qquad y = r\frac{\sigma_y}{\sigma_x}x$$

$$(2) \qquad y = \frac{\Sigma(xy)}{\Sigma(x^2)}x$$

$$(3) \qquad Y - M_y = \frac{\Sigma(XY) - N(M_x)(M_y)}{(\Sigma X^2) - N(M_x)^2}(X - M_x)$$

Again it will be noted that the first two of these formulas require that the X's and the Y's be reduced to deviations from their means, while formula (3) is in terms of the original figures. We have already learned that the regression equation can be found by solving the two normal equations

$$(4) \qquad \begin{aligned} Na + b\Sigma X &= \Sigma Y \\ a\Sigma X + b\Sigma X^2 &= \Sigma XY \end{aligned}$$

This fourth method gives results which are identical with those obtained from the other three methods, and it is often the most convenient and easy to use. Usually one will find that either formula (3) or the normal equations of method (4) will require the least time and work.

13.14. Computing the Standard Error of Estimate.—On page 392 we computed the standard error of estimate by solving the regression equation for each value of X, finding the difference between the estimated and the actual values of Y, and computing the standard deviation of these differences. This procedure would require the expenditure of a very large amount of time, and in practice S_y is never so computed. It is important that the student understand that S_y is the standard deviation of the errors of estimate; it is for that reason that we so computed it before. But when one once understands it, there is no longer any reason for taking the long method of discovering the value of S_y. As a matter of fact, this value is almost always discovered from the relationship

$$S_y = \sigma_y\sqrt{1 - r^2}$$

One can see easily whence this formula comes. We know (see page 395) that

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

$$r^2 = \frac{\sigma_y^2 - S_y^2}{\sigma_y^2}$$

$$r^2\sigma_y^2 = \sigma_y^2 - S_y^2$$

$$S_y^2 = \sigma_y^2 - \sigma_y^2 r^2 = \sigma_y^2(1 - r^2)$$

$$S_y = \sigma_y \sqrt{1 - r^2}$$

This is our formula for the standard error of estimate given above. In working out the value of r, we either compute the value of σ_y or almost do so, depending on the formula used. If we work with formula (1) on page 405 we have it all computed; this is also true if we use formula (2). In formula (3) we computed the value of Σy^2, and $\sigma_y = \sqrt{\Sigma y^2/N}$. Hence we can easily compute σ_y. If we use formula (4), we compute the value of $(\Sigma Y^2 - NM_y^2)$. This is equal to Σy^2, and therefore the value of σ_y may be computed just as when formula (3) is used. One merely divides by N and takes the square root.

13.15. Illustrative Problem.—We now apply these methods to an actual problem to see how they work in practice. Suppose that we wish to determine whether or not relationship exists between the production of potatoes and their price. We shall first work out the correlation problem with the original figures, uncorrected from changes in the general price level or for changes in population. Table 13.10 gives figures[1] showing the average farm price per bushel on Dec. 1 of each year and the total United States production for the year in millions of bushels. The data are for the years 1896–1912. (The war years are purposely omitted here on account of the violent fluctuation in the value of money.) The average production will be equal to 4,991/17, or 293.6, million bushels. The average price is 918.6/17 = 54.0 cents. Let us compute r by formula (4) on page 405,

$$r = \frac{269{,}053.7 - (17)(293.6)(54.0)}{\sqrt{[1{,}530{,}629 - (17)(293.6^2)][52{,}487.16 - (17)(54^2)]}}$$

$$= \frac{-471.1}{14{,}145} = -0.033$$

[1] From "U.S. Department of Agriculture Yearbook," 1922, p. 668.

This would show very little relationship, since the lowest possible figure would be 0, and here we have a coefficient of but 0.033.

TABLE 13.10.—COMPUTATION OF COEFFICIENT OF SIMPLE LINEAR CORRELATION BETWEEN POTATO PRODUCTION AND POTATO PRICES, 1896–1912

Year	(1) Production (X)	(2) Price (Y)	(3) (XY)	(4) (X^2)	(5) (Y^2)
1896	272	29.0	7,888.0	73,984	841.00
1897	191	54.2	10,352.2	36,481	2,937.64
1898	219	41.5	9,088.5	47,961	1,722.25
1899	260	39.7	10,322.0	67,600	1,576.09
1900	248	42.3	10,490.4	61,504	1,789.29
1901	199	76.3	15,183.7	39,601	5,821.69
1902	294	46.9	13,788.6	86,436	2,199.61
1903	262	60.9	15,955.8	68,644	3,708.81
1904	352	44.8	15,769.6	123,904	2,007.04
1905	279	61.1	17,046.9	77,841	3,733.21
1906	332	50.6	16,799.2	110,224	2,560.36
1907	323	61.3	19,799.9	104,329	3,757.69
1908	302	69.7	21,049.4	91,204	4,858.09
1909	395	54.2	21,409.0	156,025	2,937.64
1910	349	55.7	19,439.3	121,801	3,102.49
1911	293	79.9	23,410.7	85,849	6,384.01
1912	421	50.5	21,260.5	177,241	2,550.25
Totals...	4991	918.6	269,053.7	1,530,629	52,487.16

Let us correct it for the number of cases as suggested on page 397. This will give

$$r'^2 = 1 - (1 - 0.033^2)(16/15)$$

This equation gives a negative value for r'^2, and, as we noted on page 399, it means that we must call the value of the coefficient of correlation 0. Our conclusion is, then, that $r' = 0$. There is no evidence of linear relationship between the variables.

Yet it seems peculiar, does it not, that there should be no relationship between the production of potatoes and their price? (To be sure, we have not shown that there is no relationship, but we have failed to find any.) How can we explain it? Possibly by the fact that the higher productions of the later years did not bring lower prices because population had increased in the mean-

time, and perhaps in part by the fact that the general price level was rising throughout the period, so that a constant price for potatoes would have corresponded to a falling purchasing power. There is certainly no reason for proceeding further by linear methods with a problem in which there is such a small amount of linear correlation, but it may be worth while to correct our original data to see if there would then be correlation.

It would seem logical to correct the figures on potato production by dividing them by the population of the United States, thus putting them in terms of production per capita. And it would also seem wise to correct the price figures for changes in the value of money by dividing them by the price index number (as we did on page 361). In Table 13.11 are estimates of the population of the United States for each of the years in question[1] and price index numbers for the same years based on the average wholesale prices of all commodities in the period 1910–1914.[2]

TABLE 13.11.—ESTIMATED UNITED STATES POPULATION AND INDEX NUMBERS OF WHOLESALE PRICES OF ALL COMMODITIES, 1896–1912

Year	Estimated Population (millions)	Index Number of Prices
1896	70.9	68
1897	72.2	68
1898	73.5	71
1899	74.8	77
1900	76.1	82
1901	77.7	81
1902	79.4	86
1903	81.0	87
1904	82.6	87
1905	84.2	88
1906	85.8	90
1907	87.5	95
1908	89.1	92
1909	90.7	99
1910	92.3	103
1911	93.7	95
1912	95.1	101

[1] From *Statistical Abstract*, 1933, p. 10.

[2] From *Farm Economics*, Cornell University, September, 1931, pp. 1586–1587.

In Table 13.12 the column headed X contains figures on per-capita potato production, found by dividing the actual production by the population. The figures are in bushels per capita. In the column headed Y the figures represent the price of potatoes in 1910–1914 dollars; that is, an attempt has been made to correct the prices to account for changes in the value of money. The average per-capita production for the period was $59.85/17 = 3.52$ bushels; and the average price (stated in terms of 1910–1914 dollars) was $1063.3/17 = 62.55$ cents. If we substitute these figures in formula (4) on page 405, we get

$$r = \frac{3654.441 - (17)(3.52)(62.55)}{\sqrt{[215.7475 - (17)(3.52^2)][69{,}730.15 - (17)(62.55^2)]}}$$
$$= \frac{-88.551}{128.23} = -0.690$$

TABLE 13.12.—COMPUTATION OF COEFFICIENT OF SIMPLE LINEAR CORRELATION BETWEEN CORRECTED POTATO PRICES AND PRODUCTION PER CAPITA, 1896–1912

Year	(X)	(Y)	(XY)	(X^2)	(Y^2)
1896	3.84	42.6	163.584	14.7456	1,814.76
1897	2.64	79.7	210.408	6.9696	6,352.09
1898	2.98	58.5	174.330	8.8804	3,422.25
1899	3.46	51.5	178.190	11.9716	2,652.25
1900	3.26	51.6	168.216	10.6276	2,662.56
1901	2.56	94.3	241.408	6.5536	8,892.49
1902	3.71	54.5	202.195	13.7641	2,970.25
1903	3.24	70.1	227.124	10.4976	4,914.01
1904	4.26	51.5	219.390	18.1476	2,652.25
1905	3.32	69.5	230.740	11.0224	4,830.25
1906	3.88	56.2	218.056	15.0544	3,158.44
1907	3.70	64.5	238.650	13.6900	4,160.25
1908	3.39	75.8	256.962	11.4921	5,745.64
1909	4.35	54.8	238.380	18.9225	3,003.04
1910	3.76	54.1	203.416	14.1376	2,926.81
1911	3.12	84.1	262.392	9.7344	7,072.81
1912	4.42	50.0	221.000	19.5364	2,500.00
Totals...	59.89	1,063.3	3,654.441	215.7475	69,730.15

Correcting for the number of cases, we get

$$r'^2 = 1 - (1 - 0.690^2)(16/15) = 0.441$$
$$r' = \sqrt{0.441} = \pm 0.664$$

We give this corrected value the sign of the original uncorrected r, which was negative. Thus we state that $r' = -0.664$.

Again the value of r is low. We shall see a little later how it is to be interpreted. In the meantime we shall compute the regression equation and the value of S_y. In computing the latter, we shall use the corrected value of r. Using the formula on page 408, we get

$$S_y = \sigma_y \sqrt{1 - r^2}$$

We compute the value of σ_y from the equation

$$\sigma_y = \sqrt{\frac{\Sigma Y^2 - NM_y^2}{N}} = \sqrt{\frac{3217.61}{17}}$$
$$= 13.8$$

The value of $\sqrt{1 - r^2}$ can be computed from the value of r, but it is much easier to look it up in books of tables.[1] Then we get the following value for S_y:

$$S_y = 13.8(0.7477) = 10.3$$

The standard deviation of the corrected prices themselves was 13.8 cents, so without any regression equation at all we could have estimated the corrected price within 13.8 cents of the correct figure in two-thirds of the cases. Using the regression equation, we can estimate within 10.3 cents of the true figure in two-thirds of the cases. There has been some improvement by using the regression equation; the reduction from 13.8 to 10.3 cents is a reduction of 25 per cent. This is certainly better than nothing, yet a comparison of S_y and σ_y indicates that after using the regression equation we still have 75 per cent of the original error with which we started.

We must still compute the regression equation. We shall substitute the proper values in the normal equations and solve for the values of a and b. This computation gives us the following:

$$17a + 59.89b = 1063.3$$
$$59.89a + 215.7475b = 3654.441$$

[1] If special statistical tables are not available, the desired value may be read from ordinary trigonometric tables. As Holbrook Working has pointed out, if we let $r = \cos \alpha$, then $\sin \alpha = \sqrt{1 - r^2}$. Using the present case as an example, $r = 0.664$. (We neglect the sign of r in this work.) Find the angle whose cos = 0.664. This turns out to be 48° 24′. The sine of this angle is 0.7477, which is the value of $\sqrt{1 - r^2}$ when $r = 0.664$.

Solving, we find

$$a = 130.36$$
$$b = -19.25$$

Our regression equation, then, is

$$Y = 130.36 - 19.25X$$

Suppose that the present population is 100 million, we have a potato crop of 250 million bushels, and the current price index is 150. What is the expected price of potatoes?

A crop of 350 million bushels when the population is 100 million people is a crop of 3.5 bushels per capita. The X in our regression equation represents the crop in bushels per capita. We therefore substitute 3.5 for X in the regression equation and solve:

$$Y = 130.6 - 19.25(3.5) = 63.2$$

This tells us that we should expect a corrected price of 63.2 cents. But we should like to know what actual price to expect. The corrected price is the actual divided by the index number, and the index number is now 150. If we multiply the corrected price by the index number, it will be put back into the form of ordinary prices. This gives us

$$63.2(1.5) = 94.8 \text{ cents}$$

We can say, then, that our best estimate of the actual price is 94.8 cents.

How much error can we expect in the estimate? S_y tells us that in two trials out of three we should be able to estimate within 10.3 cents of the actual corrected figure. But these are 10.3 cents in corrected prices. With the index number at 150, this corresponds to $10.3(1.5) = 15.4$ cents of the current dollar. Hence we can say that the chances are 2 out of 3 that our estimate of 94.8 cents will be within 15.4 cents of the correct figure, and we are practically certain that it will not miss by over $3(15.4) = 46.2$.

It will be noted that preliminary adjustments of our data (correcting for population change and changes in the price level) raised our correlation in this case from none at all to 0.664 (corrected for number of cases). It is possible that further adjustments (such as elimination of trends, etc.) would raise the

coefficient still higher. The student should understand that it is commonly necessary for a statistician in any field to adjust his data before he begins to correlate. It is of great importance that these adjustments be made by someone who is familiar with the data and with the field of knowledge being studied. A thorough acquaintance with statistical methods is a great help to workers in many fields, but it cannot fit men to work in any field. The statistician must know more than statistical method before it is safe to turn him loose. He must know his genetics or his economics or his education or his biology first in order that he may know how to apply his statistical methods within the field.

13.16. Interpretation of Correlation Coefficients.—We have now computed the coefficient of simple linear correlation in three cases. When we were studying the relationship between the temperatures of the two hands, we found that $r' = +0.977$. When we studied the relationship between potato production and the farm price of potatoes, we found that $r' = 0$. Now we have studied the relationship of per-capita output of potatoes to their corrected price and find that $r' = -0.664$. What do these figures mean?

To begin with, we know already the meaning of the sign of the coefficient. We have discovered that r has the same sign as b in the regression equation, and that, if the sign is plus, the correlation is positive, while, if the sign is minus, the correlation is negative (the adjectives being used in the sense explained on page 382). That is all we need to know about the sign of the coefficient, and the remainder of our attention will be given to the absolute size of r. A coefficient of $r = -0.664$ and a coefficient of $r = +0.664$ show exactly the same degree of relationship. The signs merely show whether the regression line slopes up toward the right (positive correlation) or down toward the right (negative correlation).

We know from our past discussion (see pages 396*ff.*) that a coefficient of 1.00 denotes perfect correlation—all the points falling on the regression line. It is possible for us to estimate the values of Y from the values of X without error. We know that a coefficient of 0 denotes the absence of linear correlation, so that we can make no better estimates with the regression equation than without it. Coefficients usually fall somewhere between 0 and 1 in size. Then the problem of interpretation is somewhat

more difficult. We can, however, say two or three things definitely about the nature of the relationship if we know the absolute value of r. Let us suppose, then, that we have a case in which $r = 0.800$ (it is immaterial whether this is $+0.800$ or -0.800). Let us discover as many things about the relationship as we can.

In the first place, we know that

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

In this case, then, we know that

$$0.800^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

$$0.64 = 1 - \frac{S_y^2}{\sigma_y^2}$$

$$\frac{S_y^2}{\sigma_y^2} = 0.36$$

$$\frac{S_y}{\sigma_y} = \sqrt{0.36} = 0.600$$

In other words, we know that the ratio of S_y and σ_y is 0.600, or that we have but 60 per cent as much error in estimate by using the regression line as by trying to get along without it. From any value of r we can find, then, the ratio of S_y to σ_y, and from this ratio we can tell something about the advantage of using the regression line. Since $S_y = \sigma_y \sqrt{1 - r^2}$, it is obvious that

$$\frac{S_y}{\sigma_y} = \sqrt{1 - r^2}$$

To know the ratio it is necessary merely to compute this value.[1] But, as we have seen, one can look up the value in the trigonometric tables and save the time of computing it.[2] If we use the trigonometric tables to look up the values of the ratio of S_y/σ_y for the various values of r which we have computed, we find when

$$r = 0$$

$$\frac{S_y}{\sigma_y} = 1.00$$

[1] The value $1 - r^2$ is called the *coefficient of alienation*, since it measures the extent of departure from perfect correlation.

[2] See p. 412*n*.

In this case we still retained 100 per cent of the error that we should have had by plain guesswork. When

$$r = -0.664$$
$$\frac{S_y}{\sigma_y} = 0.7477$$

In this case we still had 75 per cent of the error which we would have had in guesswork; that is, had we guessed at the corrected price instead of estimating it from the per-capita production,

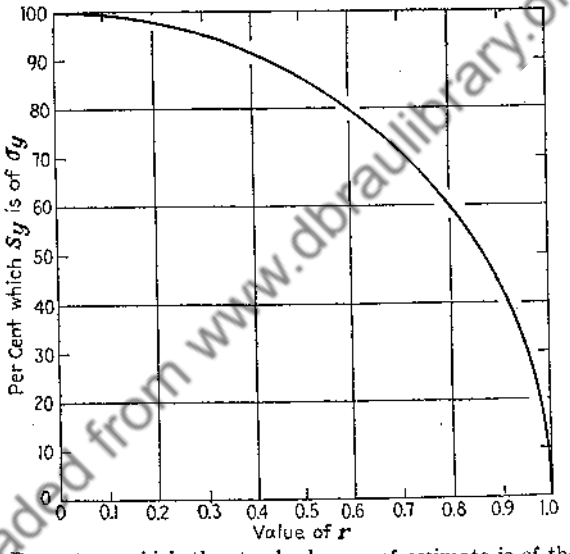


FIG. 13.5.—Percentage which the standard error of estimate is of the standard deviation when the coefficient of correlation has various values.

our error would have been only a little greater. In the case of the hand temperatures, when

$$r = 0.977$$
$$\frac{S_y}{\sigma_y} = 0.213$$

In this case the use of the regression equation cuts the error of estimate to 21.3 per cent of what it would be without the regression equation. One way, then, of interpreting r is to compute the amount of the ratio S_y/σ_y in order to find out whether the error of estimate when we use the regression equation is relatively

large or small as compared with that which we get when we do not use the regression equation.

If one does not have trigonometric tables handy, or wishes to visualize what we have been discussing a little better, he can easily graph the relationship between S_y/σ_y and r. Lay off vertical and horizontal axes on graph paper. On the vertical axis mark off percentages from 0 to 100 per cent. On the horizontal axis lay off decimals from 0 to 1. Make the scales the same size. Then with a compass describe a quarter circle passing through the 100 per cent point of the vertical scale and the point marked 1.00 on the horizontal scale. The horizontal scale represents values of r, and the vertical scale represents values of S_y/σ_y. Figure 13.5 shows the general idea, but the student can easily make a larger scale with the finer subdivisions showing on the graph paper, and from it he can read with sufficient accuracy for practical purposes the values of S_y/σ_y.[1] The chart will also serve to show something about the relative importance of coefficients of correlation of various sizes.

It will be seen that if we start with a coefficient of correlation of zero and continually increase the size of the coefficient, the ratio S_y/σ_y will become smaller; that is, the error around the regression line will become smaller as compared with the error around the mean. But at first the decreases in the ratio are very small. An increase in the size of r from 0 to 0.1 brings almost no advantage. An increase from 0.1 to 0.2 brings somewhat more accuracy of estimate. If we increase the value of r to 0.5, we have eliminated 13.4 per cent of the error of guesswork, and

[1] It will be recalled that the formula for a circle when the center is at the point of origin is

$$x^2 + y^2 = r^2$$

when r is the radius of the circle. We know in our case that

$$r^2 = 1 - \left(\frac{S_y}{\sigma_y}\right)^2$$

or that

$$r^2 + \left(\frac{S_y}{\sigma_y}\right)^2 = 1^2$$

Hence we know that the relationship can be shown by a quarter circle with unity (100 per cent, or $r = 1.00$) as the radius. Hence the construction of our graph. It is easier to construct a graph than to solve the equation if one has coordinate paper and a compass.

S_y/σ_y is 86.6 per cent. In order to get to the point where the error around the regression line is but half the error around the mean, we have to continue to increase the size of r until it reaches 0.866. This means that we get as much additional accuracy by increasing r from 0.866 to 1.00 as we got by raising it all the way from 0.00 to 0.866. The line on the chart falls very rapidly as we approach the value $r = 1.00$. This shows that with large coefficients almost any increase in size will bring a decided reduction in the error of estimate. Many students fail to realize this point, and they get the idea that an increase of r from 0.1 to 0.2 is the same as an increase from 0.8 to 0.9. Inspection of the diagram shows how incorrect this notion is.

On page 399 we found a corrected r of 0.977 when the uncorrected value was 0.980. At that time we raised the question as to whether such a negligible correction was worth while. Later, on page 409, we computed a coefficient of 0.033 which corrected to 0. By changing from 0.033 to 0, we lost but ½ of 1 per cent in accuracy of estimate. In changing from 0.980 to 0.977 we changed from a purported improvement in accuracy of 80.1 per cent to a corrected improvement of 78.7 per cent: a change of 1.4 per cent. Thus it is seen that our correction in the latter case was more significant than the correction which reduced a low coefficient to zero.

Let us state this in another way. By adjusting our data on prices and production in the potato problem, we raised the value of r' from nothing to 0.664. In doing this we cut the error of estimate to 75 per cent of its former amount (actually 74.77 per cent). This cut the error 25 per cent. To cut the error another 25 per cent, so that it would be 50 per cent of the error around the mean, we should have to raise the value of r' to 0.866. Thus we see that raising r' from 0.664 to 0.866 has the same effect as raising it from 0.00 to 0.664. The student should remember that small values of r have little importance, merely because they mean that there is little relationship between the data. But when one gets larger values of r, any adjustment of the data which will bring even a small increase in the size of r is worth while, because it will bring a large increase in the accuracy of estimates.

This method of interpreting the meaning of r is the most useful, but there are one or two other methods worth consideration. We noted on page 394*n*. that σ^2 is called "variance." It is a

measure of dispersion. Now it is plain from the first formula on page 395 that

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

In other words, the square of the coefficient of correlation is equal to unity minus the ratio of the variance of the errors in estimate around the line and the variance of the original figures. Since the largest possible value of S_y^2/σ_y^2 is 1.00, and since we are here subtracting the ratio from 1.00, we are showing the percentage by which the error of estimate has been reduced by the use of the estimating equation if we use variance as our measure of error. For example, suppose that we correlate students' high-school records with the marks made in their freshman year in college and find that $r = +0.600$. In this case $r^2 = 0.36$, and we can say that we have eliminated 36 per cent of the variance by using the regression equation to make our estimates. More commonly statisticians say, "36 per cent of the variation in college marks can be explained in terms of high-school records." When a statement of this kind is made, one can be reasonably certain that it is based on r^2. The value of r^2 is called the *coefficient of determination*. Ezekiel says of this measure:[1]

> Where both X and Y are assumed to be built up of simple elements of equal variability all of which are present in Y but some of which are lacking in X, it can be proved mathematically that r^2 measures that proportion of all the elements in Y which are also present in X. . . . It [r^2] may be said to measure the per cent to which the variance in Y is determined by X, since it measures that proportion of all the elements of variance in Y which are also present in X. . . . Since this is the most direct and unequivocal way of stating the proportion of the variance in the dependent factor which is associated with the independent factor it may be used in preference to the other methods.

In our example of per-capita production and corrected price, we discovered that $r = -0.664$. We might, then interpret it by computing the coefficient of determination, which is $-0.664^2 = 0.441$. This would tell us that 44.1 per cent of the variation in corrected price can be explained in terms of per-capita production, the other 55.9 per cent being tied up with other

[1] M. J. B. Ezekiel, "Methods of Correlation Analysis," p. 120, John Wiley & Sons, Inc., New York, 1930.

factors, such as changes in taste and habit on the part of consumers, changes in incomes, and so on. The coefficient of determination tells us the percentage of the variation in the dependent variable that can be explained in terms of the independent variable.

A third method of interpreting the meaning of r can be deduced from the first formula for the regression equation on page 407. Here we find that

$$y = r\frac{\sigma_y}{\sigma_x}x$$

From this equation it is obvious that

$$r = \frac{y\sigma_x}{x\sigma_y}$$

Therefore

$$\frac{y}{\sigma_y} = r\left(\frac{x}{\sigma_x}\right)$$

But y is the deviation of any value of the dependent variable from its average, and in the regression equation it refers to our computed value of Y. And x is the deviation of any value of X from the mean of the X's. Both x and y are divided by the standard deviation; this means that we are measuring the deviations in units of the standard deviation. Suppose that we have perfect correlation; that is, $r = \pm 1.00$. Then it is evident that $y/\sigma_y = x/\sigma_x$. In other words, we shall estimate a value of Y which is as many standard deviations from its mean as is the value of X in which we are interested. We want to know the most probable value of Y to accompany a given value of X. We find how many standard deviations X is from its mean and place Y exactly the same number of standard deviations from its mean. If, however, $r = 0.500$, we note that $y/\sigma_y = 0.5x/\sigma_x$. In this case if we are trying to estimate the value of Y which is most likely to accompany a given value of X, we shall first determine by how many standard deviations the X differs from its average and then estimate that Y will be half as many standard deviations from its average. If $r = 0.75$, we shall estimate a value of Y that differs from the average of the Y's by three-quarters as many standard deviations as the number by which X differs from the average of the X's. In other words, we can look

upon r as a percentage showing how far Y will be from the average in terms of the deviation of X. We always estimate a value of the dependent variable that is r times as far from its average (in units of the standard deviation) as is the independent variable from its average.

Each of the methods of interpreting r which we have so far discussed depends on standard deviations. Usually we are interested in the standard deviation of the dependent variable and the standard deviation of the scatter around the regression line. These are represented by σ_y and S_y. We learned, however, in Chap. VI that the standard deviation has an exact and known interpretation only when computed from a normal distribution. This means that we cannot use the methods of interpretation given here with any degree of assurance unless the dependent variable and the scatter around the regression line (the distribution of the residuals) are both normal distributions. In correlation work it is desirable to test these distributions for normality. These observations on the importance of normality also hold good in the more complicated problems of multiple and joint correlation which are to be covered in later chapters.

Finally, if we are to interpret r satisfactorily we must know its standard error (or its probable error). We have seen that values of r below 0.5 or 0.7 do not indicate the existence of much relationship. Even where r has a value considerably above these points, we must attempt to discover whether or not the value could reasonably be expected to have arisen by chance from basic data that were actually uncorrelated. In our study of the relationship between per-capita production of potatoes and their corrected price, we found that $r = -0.664$ when $n = 17$. But is this correlation significantly different from zero—that is, is it high enough to make us feel certain that there actually was some degree of negative relationship in the universe? To test this we use as a standard error the measure[1]

$$\sigma_r = \frac{1}{\sqrt{n-1}}$$

In our present example this formula gives us

[1] This formula is used only when testing to see if a given coefficient is significantly different from zero. It is evidently found by letting $r = 0$ in the formula on p. 400.

$$\sigma_r = \frac{1}{\sqrt{16}} = 0.25$$

Dividing our value of r by its standard error, we get

$$\frac{r}{\sigma_r} = \frac{-0.664}{0.25} = -2.656$$

In other words, this coefficient differs from 0 by but 2.656 standard deviations. If there were no correlation at all in the universe, and if we drew many samples of 17 cases each, we should expect in half of the samples to find positive correlation. In another 49.6 per cent of the cases we should expect to find coefficients of correlation between 0 and -2.66σ. But in four-tenths of 1 per cent of the cases we should get just by pure chance negative values of r of greater absolute magnitude than this one. While it is rather unlikely, then, such a coefficient may possibly have arisen by chance, and we cannot call our result significant with entire confidence without studying more cases.

The last statement must not be interpreted to mean that there was really no correlation in the universe. In fact, it is much more likely that there was negative correlation in the universe than that there was not. We do not say that correlation did not exist, but merely that we have not studied enough cases to be sure of it yet. Seventeen cases are not enough on which to make a decision. For this reason we say that the correlation in our sample was not significant, and we either let the whole matter drop or proceed to study more cases to discover whether or not the relationship continues.

We can learn much more about the probable value of the relationship in the universe by carrying through the z-transformation described on pages 400*ff*. We find (by interpolation in Appendix VI) that when $r = -0.664$, $z = -0.7999$. The standard error of z in this case (with $n = 17$) is 0.2673. We should expect to find the true z value of the universe within $3\sigma_z$ of the z value of the sample, or between $-0.7999 + 0.8019$ and $-0.7999 - 0.8019$. Thus we have z limits of -1.6018 and $+0.0020$, corresponding to r values of -0.922 and $+0.002$. From these figures it is again evident that our sample is altogether too small to tell us with any degree of certainty what is true of the universe. Although our sample yielded a negative

coefficient of correlation, we discover that it is quite possible for the coefficient for the universe to be practically zero.

It is important for us to distinguish between two adjectives which are commonly used to describe coefficients of correlation. Sometimes we say that a coefficient of correlation is *high*, or that there is a *high correlation*. Again we may say that a coefficient is *significant*, or that there is a *significant correlation*. A high correlation is one in which the absolute size of the coefficient (that is, disregarding the sign) is close to 1.00. The term would not usually be applied until r exceeded about 0.85 or 0.90, although there is no general rule with regard to its use. A significant correlation is one which, when taken in conjunction with the corresponding value of z, indicates with reasonable certainty that the direction of the correlation in the universe (positive or negative) is the same as the direction of the correlation in the sample. If we accept three standard errors as the limit of fluctuation which can be accepted as being the result of pure chance, we can restate this by saying that one converts any value of r to the equivalent value of z. He then computes the standard error of z, and if z exceeds three times its standard error the correlation is said to be significant. In these cases where z exceeds three times its standard error, it is obvious that the sign of the coefficient of correlation in the universe must be the same as the sign of the coefficient in the sample. Let us suppose that r is very small—say 0.090. In this case the value of z is 0.0902. If the sample contained 5000 cases, the standard error of z was 0.014, and z was more than three times its standard error. We can feel sure in this case that there was correlation in the universe, and that it was positive —but we can also feel sure that any linear correlation in the universe was too small to be of any practical value when we wish to estimate values of Y from values of X. The correlation was low, but significant, in this case.[1]

[1] It is still a very common practice to give the value of r in conjunction with the value of its standard error or its probable error. It was once the accepted rule to consider a coefficient of correlation significant if it exceeded three times its standard error or 4.5 times its probable error. Setting any border line between significance and insignificance is arbitrary, and the particular point selected may be questioned. But the use of the standard error or the probable error at all in connection with coefficients of correlation should be discouraged now that we know that values of the coefficient found

13.17. Correlation of Grouped Data.—We have seen that it is possible, by making certain assumptions relative to the distribution of the items within the classes, to compute averages and measures of dispersion from data which have been classified in frequency tables. It is also possible to compute the value of r (or at least to approximate it) from data which have been classified in a two-way frequency table. Such tables are usually called *correlation tables.* We can illustrate with a table which shows the relationship between the length of the femur (upper bone of the hind leg) and the length of the humerus (upper bone of the front leg) in 370 rabbits. One would rather expect from observation that rabbits with long front legs would tend to have long hind legs, and that those whose front legs were short would tend to be underslung behind also. If this is correct we should get a positive correlation. The data appear in Table 13.13. Each

TABLE 13.13.—NUMBERS OF RABBITS WITH VARIOUS COMBINATIONS OF FEMUR LENGTH AND HUMERUS LENGTH

Humerus Length (millimeters)	Femur Length (millimeters)										Totals
	76–77	78–79	80–81	82–83	84–85	86–87	88–89	90–91	92–93	94–95	
78–79	..	..	..	..	..	..	.	..	..	1	1
76–77	..	..	..	..	..	..	..	..	..	..	0
74–75	..	..	..	..	..	..	..	..	2	..	2
72–73	..	..	..	1	..	1	4	1	3	..	10
70–71	..	..	..	..	3	13	13	4	..	..	33
68–69	..	..	1	10	29	29	4	..	..	..	73
66–67	..	..	13	52	47	4	..	..	..	..	116
64–65	..	9	51	32	4	..	..	..	..	..	96
62–63	2	16	13	4	..	..	..	..	..	..	35
60–61	1	2	1	..	..	..	..	..	..	..	4
Totals.....	3	27	79	99	83	47	21	5	5	1	370

figure in the table shows the number of rabbits whose femur length was within the range given at the top of the column and whose humerus length was within the range given at the left of the row. That is, all the rabbits listed in a given column had the

when repeated samples are taken from the same universe are not distributed normally. Especially when n is small or r is large, the student must learn to distrust the old measures and to rely on the z-transformation.

same (approximately) femur length, but as we pass down the column we come to shorter and shorter humerus lengths. In any row all rabbits have about the same humerus length, but as we pass from left to right we pass to cases of greater and greater length of femur.

It will be seen that the distribution of entries in the table[1] is similar to the distribution of dots on a scatter diagram. In this case the entries fall in a band which rises as we move toward the right. It is immediately evident that long femurs are associated with long humeri. It still remains for us to compute the value of r, the value of S_y, and the regression equation. These values could, of course, be found by assuming that each rabbit was at the mid-point of his class, extracting the data from the table, and computing r in the usual way. Thus, we could list for one rabbit $X = 94.5$ and $Y = 78.5$. (This rabbit would be the one entered at the extreme upper right of the table.) Similarly we could extract the figures for each other rabbit, duplicating the figures as many times as there were rabbits in each class. There is, however, a shorter way of going at the problem which gives the same answers. It involves taking arbitrary origins of classes and taking deviations in class-interval units, as was done when we computed the mean and the standard deviation by the short methods.

We start with the data arranged in a correlation table in the form we have just seen. We total each column and each row. The column totals give us a frequency distribution of the X values just as though we had them arranged in a frequency table, and the totals of the rows give us a frequency distribution of the Y values. We take an arbitrary origin at the center of one of the class intervals (near the center of the X distribution and again near the center of the Y distribution) and then count off the class deviations in each direction as we did when finding σ by the short method (see pages 141*ff.*). If we label the totals of the columns f_x (meaning frequencies of the X classes) and the totals of the rows f_y (meaning the frequency with which various values of Y occurred), and if we call the class deviations of the X's d_x and he class deviations of the Y's d_y, we can proceed to the major part of the computation immediately.

[1] The figures are taken from Castle, "Genetics and Eugenics," p. 68, Harvard University Press, Cambridge, 1916.

We wish to compute the values which are necessary in the derivation of the two standard deviations. These values are Σf_x (or Σf_y, since these two values are equal and each equals N), $\Sigma(f_x d_x)$, $\Sigma(f_x d_x^2)$, $\Sigma(f_y d_y)$, and $\Sigma(f_y d_y^2)$. We can proceed to

TABLE 13.14.—COMPUTATION OF COEFFICIENT OF SIMPLE LINEAR CORRELATION FROM GROUPED DATA, BASED ON TABLE 13.13

f_x	d_x	$f_x d_x$	$f_x d_x^2$	f_y	d_y	$f_y d_y$	$f_y d_y^2$
1	+5	5	25	1	6	6	36
5	+4	20	80	0	5	0	0
5	+3	15	45	2	4	8	32
21	+2	42	84	10	3	30	90
47	+1	47	47	33	2	66	132
83	0	0	0	73	1	73	73
99	−1	− 99	99	116	0	0	0
79	−2	−158	316	96	−1	−96	96
27	−3	− 81	243	35	−2	−70	140
3	−4	− 12	48	4	−3	−12	36
370	Totals	−221	987	370		+ 5	635

compute these one at a time, getting our original values of f_x and f_y from the totals of the columns and rows, respectively, and taking arbitrary origins. The computations are given in Table 13.14.

The values of f_y and f_x come from the table on page 424. To check our arithmetic, we note that Σf_x always equals Σf_y. We can now go ahead with the computations of the means and the standard deviations in accordance with the formulas on pages 89 and 143. The results in this case are

$$\bar{X} = 84.5 + 2\left(\frac{-221}{370}\right) = 83.306 \text{ mm.}$$

The value of the guessed mean was the class mark of the class containing 83 cases. This class mark was 84.5, which appears in the above computation as the guessed mean. In the case of the Y's, we guessed the mean at the class mark of the class containing 116 cases. This class mark is 66.5, which appears in the computation of the average of the Y's:

$$\bar{Y} = 66.5 + 2\left(\frac{+5}{370}\right) = 66.527 \text{ mm.}$$

The two standard deviations are

$$\sigma_x^2 = \frac{987}{370} - \left(\frac{-221}{370}\right)^2 = 3.03$$
$$\sigma_x = \sqrt{3.03} = 1.74$$
$$\sigma_y^2 = \frac{635}{370} - \left(\frac{5}{370}\right)^2 = 1.716$$
$$\sigma_y = \sqrt{1.716} = 1.31$$

These standard deviations are both in units of the class interval (which is 2 mm.) and would have to be multiplied by this class interval if they were to be given in millimeters. For our purposes, however, they can be used as they are in class-interval units, since we merely wish to compare them with other figures which are also in these units.

It is still necessary for us to compute something equivalent to the $\Sigma(XY)$ column of our previous method; that is, we need some total of the cross-products found by multiplying together the various X and Y values. We obtain it by multiplying each entry in the original table on page 424 by its d_x value and by its d_y value, keeping track of signs. If we rewrite the original table, using the class deviations instead of the original class limits, we get Table 13.15. We now multiply each entry in the body of

TABLE 13.15.—COMPUTATION OF COEFFICIENT OF SIMPLE LINEAR CORRELATION FROM GROUPED DATA, BASED ON TABLE 13.13

Humerus Length (millimeters)	Femur Length (millimeters)										Totals
	−4	−3	−2	−1	0	1	2	3	4	5	
6	..	..	..	..	..	..	..	..	..	1	1
5	..	..	..	..	..	..	..	..	..	..	0
4	..	..	..	..	..	..	..	..	2	..	2
3	..	..	..	1	..	1	4	1	3	..	10
2	..	..	..	..	3	13	13	4	..	..	33
1	..	..	1	10	29	29	4	..	..	..	73
0	..	..	13	52	47	4	..	..	..	..	116
−1	..	9	51	32	4	..	..	..	..	..	96
−2	2	16	13	4	..	..	..	..	..	..	35
−3	1	2	1	..	..	..	..	..	..	..	4
Totals.....	3	27	79	99	83	47	21	5	5	1	370

the table by the deviation in the column at the extreme left and by the deviation in the row at the extreme top. For example, in the upper right-hand corner is the figure 1. This would be multiplied by the figure 5 at the top and by the figure 6 at the left to give $1 \times 5 \times 6 = 30$. The highest figure in the next column to the left is 2. This will be multiplied by the figure 4 at the head of the column and the figure 4 at the left of the row to give $2 \times 4 \times 4 = 32$. Some of the results may be negative. For example, the highest number in the fourth row from the left is under the heading -1 and at the right of the heading 3. Hence we get, since the number itself is 1, the following:

$$1 \times -1 \times 3 = -3$$

We carry out such a multiplication for each entry in the table, each time multiplying together three numbers (the number entered in the body of the table, the number at the head of the column in which it is entered, and the number at the left of the row in which it is entered). We add the products so obtained, keeping track of the signs. In the case of the present table, we then have

$$\Sigma f d_x d_y = 627$$

We now solve for r by the formula

$$r = \frac{\frac{\Sigma(f d_x d_y)}{N} - \left(\frac{\Sigma(f_x d_x)}{N}\right)\left(\frac{\Sigma(f_y d_y)}{N}\right)}{(\sigma_x)(\sigma_y)}$$

The values of σ_x and σ_y which are substituted in this formula must be in units of the class interval. In the present case, this formula gives

$$r = \frac{\frac{627}{370} - \left(\frac{-221}{370}\right)\left(\frac{5}{370}\right)}{(1.74)(1.31)} = \frac{1.703}{2.279} = +0.747$$

The corrected value can be found by the usual method, although when there are as many as 370 cases the correction is of little importance. Using the correction formula on page 398, we get

$$r'^2 = 1 - (1 - 0.747^2)\left(\frac{369}{368}\right) = 0.557$$

$$r' = \sqrt{0.557} = \pm 0.746$$

The corrected value is, of course, almost identical with the uncorrected value. When one is working with a large number of cases, as in the present problem, it never pays to use this correction.

We find S_y from the usual formula:

$$S_y = \sigma_y \sqrt{1 - r^2}$$

We found that $\sigma_y = 1.31$ class intervals (see page 427). But the class interval is 2 mm., so we have to multiply by this figure to get the standard deviation in the original units:

$$\sigma_y = 1.31(2) = 2.62 \text{ mm.}$$
$$S_y = (2.62)(0.666) = 1.74 \text{ mm.}$$

We expect, then, to be able to estimate humerus length with an error of 1.74 mm. or less in two-thirds of the cases, and we should never expect to make an error of over $3(1.74) = 5.22$ mm.

We can compute the regression equation from the first formula on page 407. Our two standard deviations (found by multiplying the figures on page 427 by the class interval of 2) are $\sigma_x = 3.48$ mm. and $\sigma_y = 2.62$ mm. If we substitute these values in the equation, we get

$$y = +0.746 \left(\frac{2.62}{3.48}\right) x$$
$$y = 0.561x$$

But x and y are deviations from the respective averages; that is, $x = X - \bar{X}$, and $y = Y - \bar{Y}$. If we substitute these other values in the equation above, we have

$$Y - \bar{Y} = 0.561(X - \bar{X})$$

But we know the values of $\bar{Y}$ and $\bar{X}$ from page 426. Substituting their values in the equation, we get

$$Y - 66.53 = 0.561(X - 83.3)$$
$$Y = 19.8 + 0.561X$$

This is our regression equation in its simplest form. If we wish to estimate the humerus length of a rabbit with a femur 90 mm. long, we find

$$Y = 19.8 + 0.561(90) = 70.3 \text{ mm.}$$

In this case the value of a in the type equation has no sensible interpretation when taken alone; that is, it makes no sense to say that rabbits whose femur length is zero will tend to have a humerus length of 19.8 mm. In such cases, where the usual interpretation makes no sense, one thinks of the value of a merely as a necessary auxiliary to be used in determining the values of the dependent variable, and one does not try to interpret it separately.

The value of b is 0.561. This tells us that there tends to be an increase of 0.561 mm. in humerus length for each 1 mm. increase in femur length; that is, b always tells us the change in the dependent variable which accompanies an increase of one unit in the dependent variable. On page 413 we found that the regression equation for estimating corrected potato prices from the per-capita production was

$$Y = 130.36 - 19.25X$$

Here again it makes little sense to say that if there were no production the price would be 130.36 cents. But we must always interpret the value of b, which is -19.25. This means that there tends to be a reduction of 19.25 cents in the corrected prices for each increase of one bushel in per-capita output.

Finally, we should like to find the reliability of our coefficient of correlation. When $r = 0.746$ we find that $z = 0.964$. Since $n = 370$, we find that $\sigma_z = 1/\sqrt{367} = 0.0522$. We should expect in two-thirds of the cases to find z values between $0.964 + 0.052$ and $0.964 - 0.052$, or between 1.016 and 0.912. These correspond to r values of 0.769 and 0.723. We should almost never expect to find the z value of any sample farther than $3\sigma_z$ from that of the universe. In this case we can say that the practical limits to the z values are $0.964 + 0.157$ and $0.964 - 0.157$, or 1.121 and 0.807. These figures set the limits beyond which we feel certain that the value of r in the universe does not lie at 0.808 and 0.668.

13.18. Suggestions for Further Reading.—Nowhere can the student find a better source of information on correlation methods and procedures than in Mordecai Ezekiel's, "Methods of Correlation Analysis," John Wiley & Sons, Inc., 1941. For a manual which outlines in systematic order the steps to be taken in calculation, the student should see H. A. Wallace and George W. Snedecor, "Correlation and Machine Calculation," published in

paper covers at nominal cost by the Iowa State College Book Store, Ames, Iowa. The student who is particularly interested in the z-transformation should see R. A. Fisher, "Statistical Methods for Research Workers," Oliver & Boyd, Edinburgh, 1930. In Chaps. 20 and 21 of James G. Smith's, "Elementary Statistics," Henry Holt and Company, Inc., New York, the student will find an interesting treatment of correlation in which the approach to the subject matter is very different from that used in this book. The difference in treatment may be helpful to the student who needs another point of view to throw correlation concepts into relief.

EXERCISES

1. Give two or three examples of relationship between variables which is such that it can be stated by a mathematical formula, like the relationship between the circumference and the diameter of a circle. Give two or three examples of relationships that are not ordinarily so described.

2. On pages 375–378 are several tables of figures, each indicating the existence of relationship. In which cases does the relationship seem positive, and in which negative?

3. Table 13.16 on page 432 shows the percentage of the population of each state which filed income-tax returns in 1930 and the number of automobiles registered per 100 population in 1930. The states are arranged geographically, and are in the same order in which they appear in the *Statistical Abstract.*[1]

Divide the states into two groups, putting the first 24 states (Maine through Virginia) into one group and the second 24 states (West Virginia through California) in the second group. Compute r for the first group. Use the per cent filing income-tax returns as the independent variable. Compute the regression equation. How many cars per 100 people would you expect to find in a state in which 4.5 per cent of the people filed income-tax returns? Compute and explain S_y.

4. Using the z-transformation, compute the largest and smallest values of r which you would expect to find in other samples from the universe of Exercise 3. Then compute the value of r for the second group of states in Exercise 3 to see what is actually true of the r of another sample. How can you explain the results? Were not the samples both chosen from the same universe, and at random as far as income taxes and automobile registrations were concerned?

5. In a certain Connecticut dairy region a study was made of farmers' incomes.[2] It was discovered that there was a correlation between the number of cows on a farm and the gross income of the farm of

$$r = +0.519 \pm 0.0638$$

[1] Figures are from *Statistical Abstract*, 1933. Figures on percentage of population filing tax returns are from p. 179, and those on automobiles per 100 persons are derived from 1930 population figures on p. 9 and 1930 automobile registration figures on p. 336.

[2] I. G. Davis and C. I. Hendrickson, Soil Type as a Factor in Farm Economy, *Storrs Agricultural Experiment Station Bulletin* 139, p. 92.

TABLE 13.16.—PERCENTAGE OF POPULATION FILING INCOME-TAX RETURNS, AND NUMBER OF AUTOMOBILES REGISTERED PER 100 POPULATION, BY STATES, 1930

State	Percentage Filing Income-tax Return	Cars per 100 Population
Maine	2.24	23.4
New Hampshire	2.99	24.0
Vermont	2.40	24.2
Massachusetts	4.76	19.8
Rhode Island	3.47	19.8
Connecticut	4.66	20.5
New York	5.65	18.2
New Jersey	4.65	21.0
Pennsylvania	3.35	18.2
Ohio	3.00	26.4
Indiana	2.03	27.0
Illinois	4.29	21.4
Michigan	3.04	27.2
Wisconsin	3.24	26.6
Minnesota	2.24	28.7
Iowa	1.62	31.4
Missouri	2.36	21.0
North Dakota	1.21	26.8
South Dakota	1.36	29.6
Nebraska	1.98	30.9
Kansas	1.74	31.6
Delaware	3.92	23.4
Maryland	4.19	19.7
Virginia	1.57	15.4
West Virginia	1.57	15.3
North Carolina	0.80	14.2
South Carolina	0.70	12.5
Georgia	1.00	11.8
Florida	1.92	26.1
Kentucky	1.19	12.6
Tennessee	1.25	14.0
Alabama	0.85	9.6
Mississippi	0.60	11.8
Arkansas	0.67	11.9
Louisiana	1.57	13.0
Oklahoma	1.36	23.8
Texas	1.80	23.3
Montana	2.16	25.1
Idaho	1.76	26.8
Wyoming	3.02	27.4
Colorado	2.80	29.8
New Mexico	1.49	19.7
Arizona	2.43	25.4
Utah	2.32	22.4
Nevada	4.40	33.0
Washington	1.69	28.5
Oregon	1.13	28.6
California	4.78	35.7

a. Of how much advantage was the regression equation in the reduction of the error of estimate?

b. How many farms were studied?

c. What can you say about the value of r in the universe from which these farms were drawn?

d. If a farm has 7 more cows than the average, and if the standard deviation in number of cows on these farms is 3, how large a gross income would you expect it to produce?

e. What was the standard error of the coefficient of correlation?

f. Assuming still that the standard deviation in number of cows is 3, what is the standard error of estimate, S_y?

6. Suppose that a curvilinear relationship exists between two variables, and that we compute the value of r. Will this coefficient overstate or understate the degree of the relationship? Why?

7. On page 385 is the statement that Galton found sons' heights nearer to the average than were the heights of their fathers.

a. Does this mean that heights are becoming more uniform?

b. What is the relationship between Galton's discovery and the first equation for the regression line on page 407?

8. Look up some data which you think should show relationship. Plot them on a scatter diagram.

9. Give two or three examples of "concealed classifications."

10. On page 406 is a list of directions for computing r in accordance with formula (4) on page 405. Write out a similar list of directions for formula (3) of the same page.

11. Solve the two normal equations (page 386) and show that

$$b = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$

$$a = \frac{\Sigma X^2 \Sigma Y - \Sigma X \Sigma XY}{N\Sigma X^2 - (\Sigma X)^2}$$

12. Show from the normal equations (page 386) that we can also determine the values of a and b from the following formulas (M_x = mean of the X's, etc.):

$$b = \frac{\Sigma(XY) - NM_xM_y}{\Sigma(X^2) - N(M_x)^2}$$

$$a = M_y - bM_x$$

13. Find in trigonometric tables the value of $\sqrt{1 - r^2}$ when $r = 0.758$; when $r = 0.222$. Check the results by longhand computation.

14. By definition $\Sigma y^2 = \Sigma(Y - \bar{Y})^2$. Show, therefore, that

$$\Sigma y^2 = \Sigma Y^2 - N\bar{Y}^2$$

as is stated on page 408

15. Make a chart showing the relationship of S_y/σ_y and r, as suggested on page 417. Check the problems of Exercise 13 above with the chart.

16. Compute and explain the coefficient of determination for Exercise 5 above.

17. Groves and Ogburn studied social phenomena in 170 cities. They determined for each city the sex ratio (number of males per 100 females) and the percentage of women 25 years of age and over who were married. Their results are given in Table 13.17 in which we have listed the number of cities falling in each class.[1]

TABLE 13.17.—NUMBERS OF CITIES WITH VARIOUS COMBINATIONS OF SEX RATIO AND PERCENTAGE OF WOMEN MARRIED

Sex Ratio	Per Cent of Women Married											
	44 to 47	48 to 51	52 to 55	56 to 59	60 to 63	64 to 67	68 to 71	72 to 75	76 to 79	80 to 83	84 to 87	88 to 91
60– 68	1	..	..	..	..	..	..	..	..	..	..	..
69– 77	..	..	2	..	..	..	..	..	..	..	..	..
78– 86	..	1	1	2	2	1	1	..	..	..	..	..
87– 95	..	..	..	5	18	17	..	1	..	..	..	..
96–104	..	..	..	1	5	30	18	6	..	..	..	..
105–113	..	..	..	..	..	3	6	9	1	..	..	..
114–122	..	..	..	..	..	1	7	10	..	..	..	..
123–131	..	..	..	..	1	..	2	7	1	..	..	..
132–140	..	..	..	..	..	..	..	2	..	1	..	..
141–149	..	..	..	..	..	..	..	..	2	1	..	..
150–158	..	..	..	..	..	..	..	..	1	..	..	..
159–167	..	..	..	..	..	..	..	..	..	..	1	..
168–176	..	..	..	..	..	..	..	..	..	..	..	..
177–185	..	..	..	..	..	..	..	..	..	..	..	1
186–194	..	..	..	..	..	..	..	..	..	..	1	..

Would inspection of the data lead you to believe that the relationship, if any, is positive or negative? Note that the scale at the left is reversed, with the small items at the top. Compute r, S_y, and the regression equation.

[1] GROVES and OGBURN, "American Marriage and Family Relationships," p. 481, Henry Holt and Company, New York, 1929.

CHAPTER XIV

SIMPLE CURVILINEAR CORRELATION

14.1. Curvilinearity.—In computing the coefficient of simple linear correlation we determine the usefulness of a straight line for estimating one variable from the other. It is quite possible, of course, that two variables will be closely related but that a straight line will not describe the relationship. The methods of the preceding chapter should be used only when it is fairly evident that the relationship between the variables is linear.

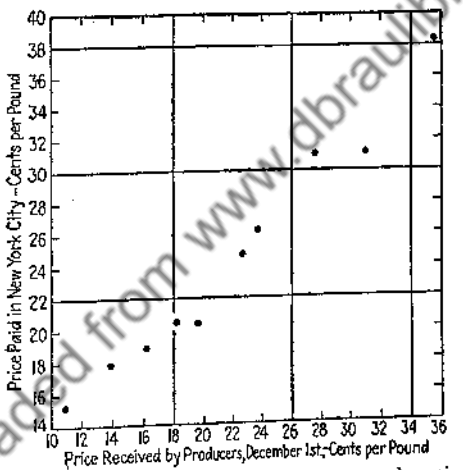


FIG. 14.1.—Relationship between average price per pound received by farmers for cotton on Dec. 1, and the average price per pound in New York City for the same year, 1918–1927. (*Data are from the U.S. Dept. of Agriculture Yearbook*, 1928, *p.* 837, *Table* 250.)

How, then, can we test for linearity? How can we know in advance whether the relationship, if any, between the variables is linear or curvilinear? We have already learned the two most useful methods of testing data in a preliminary way to determine the nature of the relationship between them. These two methods are the computation of group averages and the plotting of scatter diagrams. In Fig. 14.1 it is evident that there is a fairly close and fairly linear relationship between the price which

producers receive for cotton and the price of cotton on the New York market. The dots fall in a narrow, straight band on the chart. The fact that the band is narrow shows that the relationship is marked; the fact that the band is straight shows that the relationship is linear. Contrast this band with Fig. 14.2, which shows the relationship of potato production to potato prices. Here the dots fall in a wide, curved band. The width of the band indicates that the correlation is not so marked as in

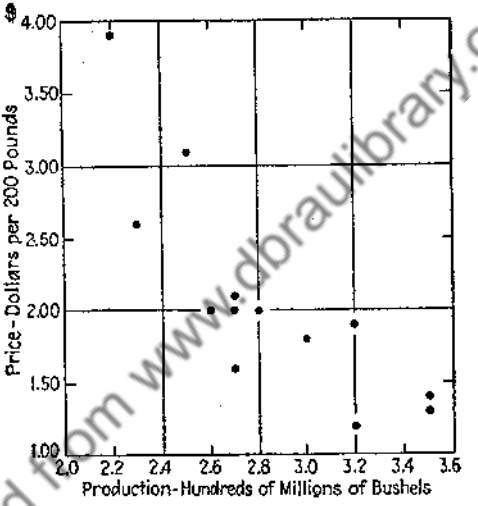


FIG. 14.2.—Relationship between potato production in 27 late-crop states and the price of potatoes at Minneapolis and St. Paul, by years, 1906–1918. Data from page 448.

the cotton case, and the fact that the band is curved indicates that the relationship is curvilinear. Similarly, if we plot the data from the tables on pages 375–378 on cross-section paper, we can discover whether the relationship tends to be linear or curvilinear. Good statisticians follow the practice of testing their data in ways such as these before they apply the more complicated methods. The computation of correlation coefficients, etc., should be the last step in a long series of statistical manipulations. Too often beginners have applied simple linear methods to data which were not at all adapted to description by a straight line.

If the relationship seems to be curvilinear, how can we go about the problem of describing it? The methods are similar

to those with which we are now familiar, and are based on the same logic.

We may, of course, plot the data on a scatter diagram and draw a free-hand curve through the points just as we have earlier drawn a free-hand straight line (page 383). This method has the advantage of simplicity and speed, and is usually quite as logical as any other method we could use. It has the disadvantage that all statisticians will not draw the same curve, so that the results will vary.

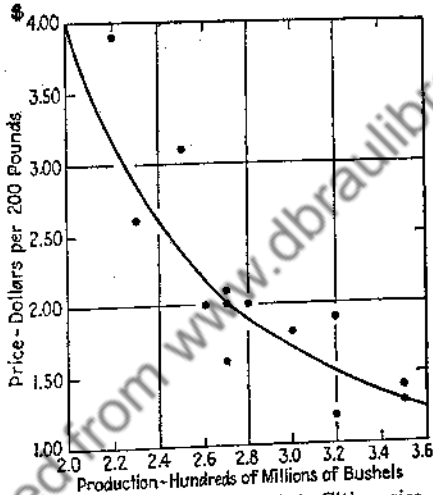


FIG. 14.3.—Late-crop potato production and Twin Cities price, 1906–1918, with free-hand regression line.

Inspection of the scatter diagram which depicts the relationship of potato production and potato prices (Fig. 14.2) reveals that there is a tendency for very marked reductions in price when production is increased at first, but that the price reduction tends to become smaller and smaller as we move toward the right-hand side of the chart. We may draw a curve through the swarm of points on the chart, making sure that the curve also falls rapidly at first and more slowly later on. We should attempt to draw the curve in such a way that it would picture the relationship as accurately as possible if it were used alone without the dots. Such a curve drawn through the potato data is shown in Fig. 14.3.

Now that we have our regression line, the process of computing the remaining correlation constants (coefficient and standard error of estimate) is familiar. We first compute the standard error of estimate. It will be recalled (see page 393) that this measure is the standard deviation of the errors which are made when we use the regression line as a basis for our estimates. In other words, the process is one of estimating each price from the free-hand curve and of comparing the estimates with the actual facts. When the potato production was 2.6, the price was actually 2.0.[1] Our curve gives a value of 2.2 for a production of 2.6. The error is, then, $2.0 - 2.2 = -0.2$; that is, the errors (or residuals, as they are often called) are defined as the differences which result from subtracting the estimated values from the actual values. In the present problem there are 13 points on the diagram, and there will therefore be 13 residuals. The number of residuals is always equal to the number of points on the scatter diagram (although some of the residuals may be equal to zero, which would mean that the line passed through such points). We should find each of the residuals, just as we have now found one of them, and we should then find the standard deviation of the residuals. This standard deviation would be S_y. To find S_y we find the differences between the actual and the estimated values of the dependent variable (using our curve as the basis for the estimates) and then compute the standard deviation of these differences.

When we have once computed the standard error of estimate, the computation of the measure of the degree of relationship is simple. When our relationship is linear, we compute the coefficient of correlation (denoted by r). In order to make a clear distinction between linear and curvilinear correlation, we call the measure of correlation around the curve the *index of correlation* and symbolize it by the Greek letter *rho* (ρ).

The index of correlation is computed in accordance with the formula for the coefficient of correlation which appears on page 395:

$$\rho = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

[1] The data on which the scatter diagram is based appear in the table on p. 448.

In order to compute ρ we must, then, compute the standard deviation of the potato prices and the standard error of estimate around the curve. Just as in the case of the coefficient of correlation, the index of correlation varies between 0 and 1. An index of 0 means that there is no advantage in using the curve in the making of estimates, and an index of 1 means that the points all lie on the curve, so that they can be estimated without error. The index of correlation is not preceded by a plus or a minus sign, because the curve may slope upward at some parts of the diagram and downward at others.

14.2. Curve Types.—The student of algebra will recall the fact that various algebraic equations can be pictured by curves of various characteristics. We have already discovered the fact that any straight line can be described by an equation of the general form

$$Y = a + bX$$

where a and b are constant values. The constants of the curve, such as a and b above, are called the *parameters*, and we have discovered a method of evaluating the parameters—that is, of determining the values of a and b. No matter what the values of the parameters may be, a curve of the type represented by this equation will always be a straight line.[1] This is evident if we understand the meaning of the parameters. The value of a, we have seen, tells us the value of Y when $X = 0$. And the value of b tells us how many units to add to Y for each unit change in X; that is, each time we add a unit of X we add b units of Y. But evidently if we add always the same amount to Y whenever we increase X by a given amount, our line will be straight.

Many other mathematical equations in X and Y can be described by lines. It would be well worth the student's while to draw several of them in order to become acquainted with their characteristics. The simpler and commoner ones are described briefly here.

We may add higher powers of X to the equation for the straight line, so that we have the following:

[1] It may sound peculiar to talk of a straight line as a curve, but mathematically it is one particular kind of curve—the kind that can be described by the formula we are discussing here.

(1) $Y = a + bX$

(2) $Y = a + bX + cX^2$

(3) $Y = a + bX + cX^2 + dX^3$

etc.

These curves are usually spoken of as *parabolas*. The second equation is that of a second-degree parabola, the third equation

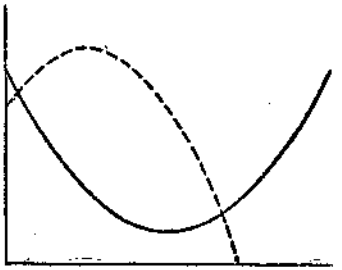


FIG. 14.4.—Typical second-degree parabolic curves.

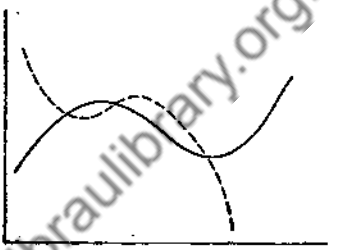


FIG. 14.5.—Typical third-degree parabolic curves.

of a third-degree parabola, and so on. Patently the straight line is a first-degree parabola. It is a characteristic of these curves that they have two fewer bends than the number of parameters. The straight line has two parameters and no bends.

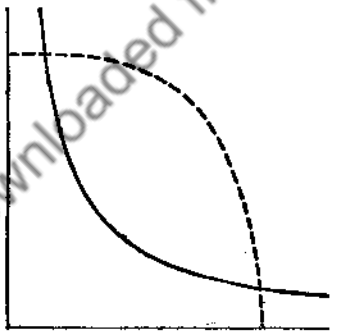


FIG. 14.6.—Typical reciprocal curves.

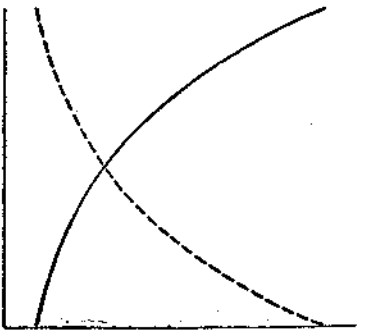


FIG. 14.7.—Typical logarithmic curves.

The second-degree parabola has three parameters and is a continuous curve bending always in the same direction; that is, it has no points of inflection. The third-degree parabola has four parameters (a, b, c, and d, in the equation above) and has two

bends; it has one point of inflection. Graphs of typical second- and third-degree parabolas appear in Figs. 14.4 and 14.5.

It may be that, although Y does not vary directly with X, the reciprocal of Y or the logarithm of Y varies with X. The relationship may be stated by one of the following equations:

$$\frac{1}{Y} = a + bX \tag{1}$$

$$\log Y = a + bX \tag{2}$$

It is evident that these equations give straight lines if used to describe the relationship between X and the reciprocal of Y or X and the logarithm of Y, but give curves if plotted against X and Y. Typical reciprocal curves and logarithmic curves appear in Figs. 14.6 and 14.7.

If the relationship as seen from group averages or from the scatter diagram seems to be similar to one of the curves shown here (either the parabolas or the reciprocal or logarithmic curves), we can use the appropriate mathematical curve to describe the relationship.[1] This gives us the advantage of getting a unique solution and a solution that is easily described by a regression equation. One must not be misled, however, into thinking that these mathematical methods give results which are any more "correct" than those obtained from the free-hand curve. Unless there is some logical reason for expecting a logarithmic or reciprocal or parabolic relationship (as there sometimes is in theory), the freehand curve is quite as "correct" as are these others. But the mathematical curves are commonly more convenient to work with.

The processes of "fitting" mathematical curves to data[2] are merely extensions of the process of fitting the straight line. We shall continue to use the least-squares criterion as that which gives the "best" line of any particular type (see page 291). The normal equations for the best fitting second-degree parabola are

$$\begin{aligned} Na + b\Sigma X + c\Sigma X^2 &= \Sigma Y \\ a\Sigma X + b\Sigma X^2 + c\Sigma X^3 &= \Sigma XY \\ a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 &= \Sigma X^2 Y \end{aligned}$$

[1] See pp. 443*ff*.

[2] To "fit" a curve to data is to find a curve that describes the data.

For the third-degree parabola the equations are

$$\begin{aligned} Na + b\Sigma X + c\Sigma X^2 + d\Sigma X^3 &= \Sigma Y \\ a\Sigma X + b\Sigma X^2 + c\Sigma X^3 + d\Sigma X^4 &= \Sigma XY \\ a\Sigma X^2 + b\Sigma X^3 + c\Sigma X^4 + d\Sigma X^5 &= \Sigma X^2Y \\ a\Sigma X^3 + b\Sigma X^4 + c\Sigma X^5 + d\Sigma X^6 &= \Sigma X^3Y \end{aligned}$$

The symmetry of these equations should make it easy for the student to derive them for parabolas of higher degrees if necessary, but it should already be evident that the labor of computation is multiplied tremendously for each additional power of X which is added.[1] The third-degree parabola is quite complicated, involving the raising of all X values to the sixth power and involving the simultaneous solution of four normal equations with four unknowns. Parabolas of higher degrees require so

[1] The normal equations can be easily derived, like those for the straight line on page 296. Let us illustrate with the second-degree parabola. The type equation is

$$Y = a + bX + cX^2$$

Each actual Y will differ from the estimate by some deviation d (which may equal 0). Thus

$$\begin{aligned} Y + d &= a + bX + cX^2 \\ d &= a + bX + cX^2 - Y \\ d^2 &= (a + bX + cX^2 - Y)^2 \\ f &= \Sigma d^2 = \Sigma(a + bX + cX^2 - Y)^2 \end{aligned}$$

The partial differentials of this function with respect to a, b, and c must be set equal to 0 if we are to minimize the sum of the squared residuals. That is,

$$\begin{aligned} \frac{\partial f}{\partial a} &= 2\Sigma(a + bX + cX^2 - Y) = 0 \\ \frac{\partial f}{\partial b} &= 2\Sigma(a + bX + cX^2 - Y)X = 0 \\ \frac{\partial f}{\partial c} &= 2\Sigma(a + bX + cX^2 - Y)X^2 = 0 \end{aligned}$$

Canceling the 2's, expanding, and summing, we get

$$\begin{aligned} Na + b\Sigma X + c\Sigma X^2 - \Sigma Y &= 0 \\ a\Sigma X + b\Sigma X^2 + c\Sigma X^3 - \Sigma XY &= 0 \\ a\Sigma X^2 + b\Sigma X^3 + cX\Sigma^4 - \Sigma X^2Y &= 0 \end{aligned}$$

Transposing the final terms of each of these equations, we get the normal equations as given above.

much work that they are seldom fitted. We shall shortly fit a parabola of the second degree in order to illustrate the use of the normal equations.

If our data are such that a reciprocal curve (or a logarithmic curve) is indicated, we merely take the reciprocals (or the logarithms) of our original Y values, and correlate them with the original X values by the methods used in simple linear correlation.

14.3. Selection of Curve Type.—When one first approaches a curve-fitting problem, one ordinarily begins by plotting the data on cross-section paper in the form of a scatter diagram. An inspection of this diagram will often show the nature of the relationship existing between the variables, indicating whether it is linear or curvilinear. If it is curvilinear, the statistician compares the shape of the curve with that of parabolas of various degrees and with logarithmic and reciprocal and other curves, and decides from this examination what kind or kinds of curves to fit. It is at this stage of the analysis that statistics is decidedly an art rather than a science, and considerable experience is necessary before one can feel any confidence in his selection of curves.

Although the preliminary steps in statistical analysis are thus graphical, requiring experience and judgment, there are a few "tricks of the trade" which are helpful in determining the type of curve to be fitted. While we cannot go into an extended discussion of such methods at this point, we shall consider some of the commonest and most useful guides which are helpful in the selection of curve types.

The straight line is so familiar that most beginning students have little difficulty in recognizing it. If the data are plotted on a scattergram, a ruler or string will soon show whether or not they fall along a straight line, and if they are distributed linearly it will indicate the approximate location of the line. In other words, when the equation $Y = a + bX$ will work, that fact can be discovered by plotting on ordinary cross-section paper and seeing whether or not the data fall along a straight line. Evidently we could discover the same thing by finding whether or not given increases in one variable were always associated with the same (or about the same) increases in the other variable, because the nature of a straight line is such that a shift of a given number of units to the right will always bring a shift of the same

number of units upward or downward. For example, suppose that we have the following values of the two variables X and Y:

X	Y
7	93
8	90
9	87
10	84
11	81
12	78
13	75
15	69

These data will fall along a straight line, and we can see that fact without plotting them at all. If we omit the last case, we note that each increase of one unit in the value of X is invariably associated with a decrease of three units in the value of Y. And in the last case, where two units are added to X, the decrease in Y is twice as much as before, or six units. Since we shall have to use this method again in more complicated cases, it will pay to become thoroughly familiar with our terminology in this first simple case.

Let us list not only our original figures but also the differences between successive figures in each column. These differences we shall call ΔX and ΔY (meaning "differences in X" and "differences in Y"), and in each case the difference is found by subtracting from any number the number above it in the column. This computation gives us the following new table:

X	ΔX	Y	ΔY	$\frac{\Delta X}{\Delta Y}$
7		93		
8	1	90	−3	$-\frac{1}{3}$
9	1	87	−3	$-\frac{1}{3}$
10	1	84	−3	$-\frac{1}{3}$
11	1	81	−3	$-\frac{1}{3}$
12	1	78	−3	$-\frac{1}{3}$
13	1	75	−3	$-\frac{1}{3}$
15	2	69	−6	$-\frac{1}{3}$

Note that the figures in the last column are all the same. The difference in X divided by the difference in Y is constant. This is a characteristic of a straight line, and where an examination

of the data shows that the quotient of these differences is constant or approximately so, we can feel safe in fitting a straight line to the data.

But now let us examine a somewhat more complicated case. Suppose that we select our data so that $Y = X^2$. Then our table will appear as follows:

X	Y	ΔY	$\Delta^2 Y$
3	9		
4	16	7	
5	25	9	2
6	36	11	2
7	49	13	2
8	64	15	2
9	81	17	2
11	121	40	23

In this table the first two columns are selected so that $Y = X^2$, and the third column shows the differences in Y. The last column does not show the squares of these differences, as one might suppose from the heading of the column, but shows what are called the "second differences of Y." As will be noted, the process of differencing has been gone through twice, so that the fourth column shows the differences of the third column just as the third column shows the differences of the second column. Note that the items in the last column are constant wherever the differences of the first column are constant.

Suppose our original figures had been those given below and our scatter diagram had led us to think that a second-degree

X	Y
1	5
2	11
3	21
4	35
5	53
6	75

parabola (one in which the second power of X was the highest power) was indicated. We could then test by finding the second differences of Y for equally spaced values of X. If such second differences are equal or approximately so, we can

conclude that a second-degree parabola is a proper curve to try. In this case if we carry out the computations we have the following:

X	Y	ΔY	$\Delta^2 Y$
1	5		
2	11	6	
3	21	10	4
4	35	14	4
5	53	18	4
6	75	22	4

In this table the values have been so selected that

$$Y = 3 + 2X^2,$$

and the second differences are constant. Often this method cannot be applied directly to the data in their original form, because the values of X are not equally spaced; that is, the first differences in X are not constant in the original problem. In such a case, however, the data may be plotted on a scattergram and a rough freehand curve drawn through the scatter. Values of Y can then be read from the curve at evenly spaced intervals of X, and the data so discovered can be tested as above.

We have just seen that a second-degree parabola is indicated if the second differences of Y are constant for equally spaced values of X. Likewise a third-degree parabola should be fitted if the third differences of Y are constant for equally spaced values of X; a fourth-degree parabola if fourth differences are constant for equally spaced values of X; and an n-degree parabola if the nth differences of Y are constant for equally spaced values of X. This principle can be illustrated by the following data.

X	Y	ΔY	$\Delta^2 Y$	$\Delta^3 Y$
2	84			
4	108	24		
6	220	112	88	
8	468	248	136	48
10	900	432	184	48
12	1564	664	232	48
14	2508	944	280	48
16	3780	1272	328	48

These data have been so selected that $Y = 100 - 10X - X^2 + X^3$. It will be noted that the third differences of Y are constant for these equally spaced values of X.

Our rule, then, for testing for parabolas of any degree, is to take equally spaced values of X and determine whether or not the nth differences of Y are constant or approximately so. Of course, even if the values of the nth differences are almost all the same size, but are getting continually larger or smaller as we go down the column, we are not justified in fitting the parabola. But if the nth differences are of approximately the same size and show no regular trend, we can safely fit an nth-degree parabola. In all such cases it is necessary that the values of X be equally spaced.

If the scattergram seems to indicate that a logarithmic curve of some kind might fit, we can test by plotting a scattergram on logarithmic or semilogarithmic paper. If $\log Y = a + bX$, the data will give a straight line on semilogarithmic paper if the Y variable is plotted on the logarithmic scale and the X variable on the arithmetic scale. This equation may also be written $Y = ck^X$, where c is the antilog of a and k is the antilog of b in the equation as given before. If $\log Y = a + b(\log X)$, the data would yield a straight scattergram on logarithmic paper.[1]

14.4. Actual Computation.—For purposes of illustration we shall now fit various lines to the scatter diagram showing the relationship of potato production to potato prices. The figures in Table 14.1 show the potato production of 27 late-crop states in hundreds of millions of bushels and the price per 200 lb. at Minneapolis and St. Paul.[2] First we plot these data on a scattergram. This we have already done (see Fig. 14.2, page 436). Next we compare the scattergram with the type curves pictured on page 440. It is obvious that a straight line will not describe the data well, but it appears that the curve is something like one of the second-degree parabolas and also something like one of the reciprocal curves. We shall fit a second-degree parabola and a reciprocal curve to the data and discover which fits the better.

[1] For further consideration of the use of logarithmic and semilogarithmic paper, see pp. 490*ff*.

[2] Figures based on W. C. Waite, "Economics of Consumption," p. 101, McGraw-Hill Book Company, Inc., New York, 1928.

TABLE 14.1.—POTATO PRODUCTION IN 27 LATE-CROP STATES AND POTATO PRICES AT MINNEAPOLIS AND ST. PAUL, 1906–1918

Year	Production (100 million bushels)	Price ($ per 200 lb.)
1906	2.7	1.6
1907	2.6	2.0
1908	2.3	2.6
1909	3.2	1.2
1910	2.7	2.0
1911	2.5	3.1
1912	3.5	1.3
1913	2.7	2.1
1914	3.5	1.4
1915	2.8	2.0
1916	2.2	3.9
1917	3.2	1.9
1918	3.0	1.8

If we are to fit the second-degree parabola, we must find the values with which to solve the normal equations on page 441; that is, we shall need to find the values of

$$\begin{array}{ll} \Sigma X & \Sigma Y \\ \Sigma X^2 & \Sigma XY \\ \Sigma X^3 & \Sigma X^2Y \\ \Sigma X^4 & N \end{array}$$

This is done most easily if we arrange the operations in tabular form. We have given the values of X and Y (in the table above) and we must have columns for the higher powers of X as well as for XY and X^2Y. We must compute these values for each year and add them. The sums (with N) will give us the needed figures for the solution of the normal equations. The figures are arranged in tabular form in Table 14.2.

The values needed for the normal equations are, then,

$$\begin{array}{rlrl} \Sigma X &= 36.9 & \Sigma Y &= 26.9 \\ \Sigma X^2 &= 106.83 & \Sigma XY &= 73.27 \\ \Sigma X^3 &= 315.303 & \Sigma X^2Y &= 203.777 \\ \Sigma X^4 &= 947.9079 & N &= 13 \end{array}$$

Substituting these values in the normal equations given on page 441, we get

$$13a + 36.9b + 106.83c = 26.9$$
$$36.9a + 106.83b + 315.303c = 73.27$$
$$106.83a + 315.303b + 947.9079c = 203.777$$

TABLE 14.2.—COMPUTATION OF PARABOLIC REGRESSION EQUATION BASED ON DATA OF TABLE 14.1

Year	Production (X)	Price (Y)	X^2	X^3	X^4	XY	X^2Y
1906	2.7	1.6	7.29	19.683	53.1441	4.32	11.664
1907	2.6	2.0	6.76	17.576	45.6976	5.20	13.520
1908	2.3	2.6	5.29	12.167	27.9841	5.98	13.754
1909	3.2	1.2	10.24	32.768	104.8576	3.84	12.288
1910	2.7	2.0	7.29	19.683	53.1441	5.40	14.580
1911	2.5	3.1	6.25	15.625	39.0625	7.75	19.375
1912	3.5	1.3	12.25	42.875	150.0625	4.55	15.925
1913	2.7	2.1	7.29	19.683	53.1441	5.67	15.309
1914	3.5	1.4	12.25	42.875	150.0625	4.90	17.150
1915	2.8	2.0	7.84	21.952	61.4656	5.60	15.680
1916	2.2	3.9	4.84	10.648	23.4256	8.58	18.876
1917	3.2	1.9	10.24	32.768	104.8576	6.08	19.456
1918	3.0	1.8	9.00	27.000	81.0000	5.40	16.200
Totals..	36.9	26.9	106.83	315.303	947.9079	73.27	203.777

Solving these equations for a, b, and c, we get

$$a = +18.9$$
$$b = -10.415$$
$$c = +\ 1.549$$

Our equation then becomes

$$Y = 18.9 - 10.415X + 1.549X^2$$

We can now use this equation in estimating values of Y from known values of X. Suppose, for example, that we wish to estimate the most probable price which would accompany a production of 280 million bushels. Our original figures are in hundreds of millions of bushels; hence this would become 2.8.

We substitute 2.8 for X in our regression equation and find

$$Y = 18.9 - (10.415)(2.8) + (1.549)(2.8^2)$$
$$Y = 1.88$$

Our estimate is, then \$1.88 per 200 lb. Similarly we could estimate the price which would accompany any other production. If we estimate the prices for several productions and locate the estimates on our chart, we can connect them with a smooth curve and thus see our regression line on the scattergram. This

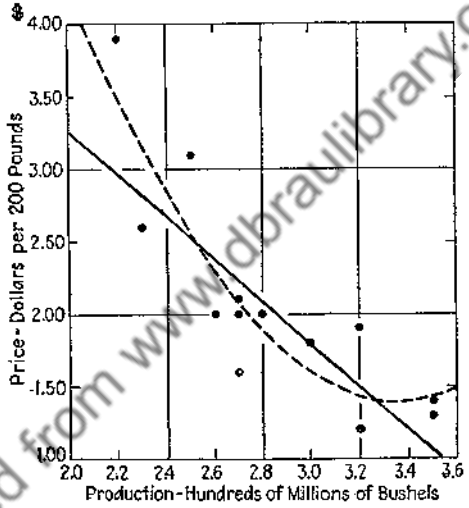


FIG. 14.8.—Late-crop potato production and Twin Cities price, 1906-1918, with straight and parabolic regression lines.

has been done in Fig. 14.8, where the parabola appears on the scatter along with the straight line fitted by least squares.[1]

[1] This chart illustrates well one of the inherent peculiarities of the second-degree parabola. It will be noted that near the right-hand edge of the chart the parabolic trend begins to rise; that is, the trend seems to indicate in this region that larger productions are accompanied by higher prices. This is a peculiarity of the method itself and not a peculiarity of the data. Since various kinds of mathematical curves are subject to various peculiarities of this kind, it is easy to see that the extrapolation of a curvilinear trend (the extending of the trend beyond the limits of the original data) is dangerous. In the present case the parabolic trend would show a continued rise in price, with production rising beyond a production of 3.36 and continuing indefinitely.

One of the advantages of the freehand trend as compared with the trend fitted by mathematical methods is that the former is not subject to these peculiarities.

It will be seen that in computing the constants for the parabolic equation we have computed all the values necessary for the straight line. Substituting the proper values in the normal equations of page 386, we get

$$13a + 36.9b = 26.9$$
$$36.9a + 106.83b = 73.27$$

Solving for the values of a and b, we get

$$a = 6.12$$
$$b = -1.428$$

The equation for the straight line is, then,

$$Y = 6.12 - 1.428X$$

Now let us fit a reciprocal curve. Instead of correlating X and Y we shall correlate X and the reciprocal of Y; that is, we shall let Y stand for the reciprocal of the price rather than for the price itself. Our variables will be

X = production (hundreds of millions of bushels)
Y = reciprocal of price

TABLE 14.3.—FITTING THE RECIPROCAL REGRESSION LINE TO THE DATA OF TABLE 14.1

Year	Production (X)	Reciprocal of Price (Y)	X^2	XY
1906	2.7	0.6250	7.29	1.68750
1907	2.6	0.5000	6.76	1.30000
1908	2.3	0.3846	5.29	0.88458
1909	3.2	0.8333	10.24	2.66667
1910	2.7	0.5000	7.29	1.35000
1911	2.5	0.3226	6.25	0.80650
1912	3.5	0.7692	12.25	2.69220
1913	2.7	0.4762	7.29	1.28574
1914	3.5	0.7143	12.25	2.50000
1915	2.8	0.5000	7.84	1.40000
1916	2.2	0.2564	4.84	0.56408
1917	3.2	0.5263	10.24	1.68416
1918	3.0	0.5556	9.00	1.66666
Totals.....	36.9	6.9635	106.83	20.48809

We proceed just as we would in simple linear correlation (see Table 14.3). The column of X's and the column of X^2's are taken directly from the previous computations (page 449). The column of Y's now represents the reciprocals of the prices, and the column of XY's is changed accordingly. The totals are now substituted in the normal equations for the straight line (page 386), to give

$$13a + 36.9b = 6.9635$$
$$36.9a + 106.83b = 20.488$$

Solving these two equations simultaneously, we get

$$a = -0.44505$$
$$b = +0.3455$$

Our regression equation is

$$Y = -0.44505 + 0.3455X$$

Since we have carried on our computations with the reciprocals of the prices rather than with the prices themselves, the estimates made with this regression equation will be in terms of the reciprocals of prices. If we want the actual prices we must take the reciprocals of these estimates; that is, if we wish now to let Y represent the actual prices, we must change our regression equation to read thus:

$$\frac{1}{Y} = -0.445 + 0.3455X$$

Now let us estimate the most probable price for a production of 280 million bushels. This production makes X equal 2.8, which gives us

$$\frac{1}{Y} = -0.445 + 0.3455(2.8)$$
$$\frac{1}{Y} = 0.52235$$
$$Y = 1.91$$

Our estimated price is, then, \$1.91 per 200 lb. As before, we may estimate the price for several different productions, locate these estimates on the scatter diagram, and draw a smooth curve through the points so found. This will give us a picture

of the reciprocal regression line. Such a curve appears in Fig. 14.9.

We have found the regression equations, and now we must find the standard errors of estimate and the indices of correlation. We shall illustrate the method with the reciprocal curve which we have just found, $1/Y = -0.445 + 0.3455X$. We have already seen that this equation may be used in estimating values of Y from known values of X. Let us estimate the price which would be expected with each of the actual productions and

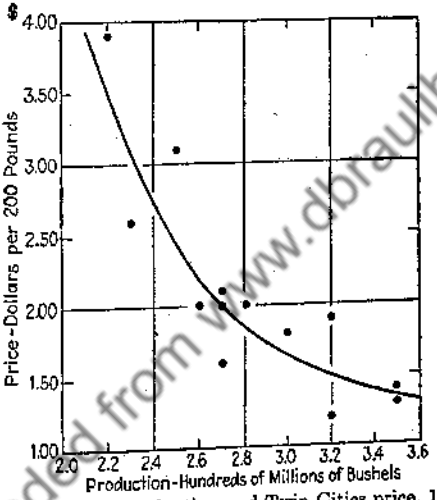


FIG. 14.9.—Late-crop potato production and Twin Cities price, 1906–1918, with reciprocal regression line. Data from page 448.

compare our estimated prices with the actual prices. The comparison is made in Table 14.4. Here the column of estimated prices is found by substituting the production of each year for X in the regression equation and solving. The differences in the last column are found by subtracting the estimated values from the actual values. We next compute the standard deviation of these differences (they are the residuals from the reciprocal curve) and find it to be 0.11. But the standard deviation of the errors in estimating is the standard error of estimate (see page 393). Thus we can say:

$$S_y = 0.11$$

We next compute the standard deviation of the original prices. This turns out to be 0.7211. We need only substitute these two

TABLE 14.4.—DETERMINATION OF ERRORS OF ESTIMATE FROM RECIPROCAL REGRESSION LINE

Year	Production	Estimated Price	Actual Price	Difference
1906	2.7	2.05	1.6	−0.45
1907	2.6	2.21	2.0	−0.21
1908	2.3	2.86	2.6	−0.26
1909	3.2	1.51	1.2	−0.31
1910	2.7	2.05	2.0	−0.05
1911	2.5	2.39	3.1	+0.71
1912	3.5	1.31	1.3	−0.01
1913	2.7	2.05	2.1	+0.05
1914	3.5	1.31	1.4	+0.09
1915	2.8	1.91	2.0	+0.09
1916	2.2	3.17	3.9	+0.73
1917	3.2	1.51	1.9	+0.39
1918	3.0	1.69	1.8	+0.11

values in the equation on page 438 to discover that

$$\rho = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}} = \sqrt{1 - \frac{0.011^2}{0.7211^2}} = 0.988$$

It is well to point out in connection with this illustration the fact that the residuals as shown in the table are peculiarly arranged. In the early years the residuals tend to be negative and in the later years positive. Such a condition would lead one to believe that it would be advantageous to eliminate trends before correlating.

The scatter diagram has shown us that the reciprocal curve is a better fit than is the straight line. The fact that the curve describes the relationship better than does the line can be seen also by comparing the index and the coefficient of correlation. Using the data from page 448, we find that $r = -0.822$. When r and ρ are both computed from the same raw data, the latter will always either equal or exceed the former in size. If the data are actually distributed in a straight band (that is, if they are really linear), the two coefficients will be equal. If the data are curvilinear, the index of correlation will always be greater

than the coefficient of correlation. This is to be expected, since in this case the curve will be a better basis of estimates than the straight line.

If we were to compute ρ from the second-degree parabola, we should follow the same steps we have just taken. First, we should estimate each year's price from the production by means of the regression equation. Second, we should find the differences between the actual prices and the estimated prices. Third, we should find the standard deviation of these differences. This would be the standard error of estimate, and in the case of the parabola it equals 0.338. We should substitute this value, along with the standard deviation of the Y's, in the equation for ρ and these solve the equation. In this case our answer is 0.884. Thus, the parabola gives us a better basis for estimate than does the straight line, but not so good a basis as the reciprocal curve.

14.5. Corrections for Number of Cases and of Parameters.—If we have but two points on our scatter diagram, the best fitting straight line will pass through them both and we shall have perfect correlation. In other words, if we try to correlate two observations, we are bound to get perfect correlation. Similarly, if we have but three points the best fitting second-degree parabola will pass through them all with no error of estimate, and ρ will equal one. Likewise, the best fitting third-degree parabola will pass through four points; and so on. In our most recent illustration we had 13 cases. Had we wished to fit a 12th-degree parabola, it would have passed through all the points and we should have found $\rho = 1.00$.

It is evident, then, that the results of correlation analysis give a somewhat exaggerated picture of the degree of the relationship, and that the extent of the exaggeration depends on two things:

1. The number of cases studied in the problem.
2. The number of parameters in the equation.

Just as we corrected the standard error of estimate and the coefficient of correlation when our relationship was linear, so must we correct our results when we have used curvilinear methods. In fact, the corrections are much more important in the latter case because the number of parameters is larger. The formulas for the corrected values (letting S'_y represent the

corrected standard error of estimate and ρ' represent the corrected index of correlation) are

$$(S_y')^2 = (S_y^2)\left(\frac{N-1}{N-P}\right)$$
$$(\rho')^2 = 1 - (1-\rho^2)\left(\frac{N-1}{N-P}\right)$$

In these two equations P represents the number of parameters in the regression equation if the equation is parabolic. For freehand curves or other nonparabolic curves one uses for P the number of parameters which it would be necessary to use in a parabolic equation to give as many twists as there are in the curve which is used. For example, the reciprocal curve has one bend (see Fig. 14.6, page 440). The second-degree parabola, which is described by an equation with three parameters (a, b, and c, as on page 440), also has one bend. Thus if we have used the reciprocal curve (or a freehand curve with one bend) we let $P = 3$. If one computes ρ from a curve with two bends, one lets $P = 4$, etc. In general the value of P will be greater by two than the number of bends in the curve.

In the case of our reciprocal curve we found that $S_y = 0.11$ and $\rho = 0.988$ (pp. 453–4). Let us correct these values. The reciprocal curve has one bend, and hence is equivalent to the second-degree parabola $Y = a + bX + cX^2$. The equivalent number of parameters is 3. Hence we have

$$(S_y')^2 = 0.11^2\left(\frac{12}{10}\right) = 0.01452$$
$$S_y' = 0.12$$
$$(\rho')^2 = 1 - (1 - 0.988^2)\left(\frac{12}{10}\right) = 0.9713728$$
$$\rho' = 0.986$$

On page 455 we gave the results of the parabolic curve as

$$S_y = 0.338$$
$$\rho = 0.884$$

These give the following corrected values:

$$S_y' = 0.370$$
$$\rho' = 0.858$$

On page 454 the linear correlation results are given as

$$r = -0.822$$

Hence we get the corrected value $r' = 0.804$. In this latter case $P = 2$.

We can now compare the results of the various methods which we have applied to the potato study. We find with the reciprocal curve that $\rho' = 0.986$. With the second-degree parabola $\rho' = 0.858$. With the straight line $r' = 0.804$. The reciprocal curve is the best to use for purposes of estimating the price, and the straight line is the worst of those tried; but even the straight line is better than nothing. With the reciprocal curve $S'_y = 0.12$, which means that two-thirds of our estimates should fall within 12 cents of the actual price and that we should never be in error by over 36 cents. With the parabolic curve $S'_y = 0.37$; hence two-thirds of our estimates should be within 37 cents of the actual price and we should practically never be in error by over \$1.11. The standard error around the straight line is 0.429. This means that, if we use the straight line as the basis of our estimates, we should be within 43 cents of the correct price two-thirds of the time and we should almost never err by over \$1.29.

It is necessary to add one word of explanation to the formula for correcting ρ. If the value of $(1 - \rho^2)\left(\frac{N-1}{N-P}\right)$ should turn out to be greater than 1, so that the value of ρ^2 is negative, the corrected result should be called zero; that is, there is no evidence of relationship.

We noted a moment ago that if the number of cases and the number of parameters are equal, the line will pass through all the points and the uncorrected index of correlation will equal unity. It will be mathematically necessary for us to get perfect correlation. But note what will happen when we correct the index of correlation. If the number of cases and the number of parameters are equal, $N - P = 0$. Thus we have a value of zero for the denominator of the last part of the correction formula. But division by zero is not allowed in algebra; the idea is absurd. Likewise the idea of fitting a complicated curve to a small number of points is absurd.

14.6. Linear and Curvilinear Correlation Compared.—It will be noted that the formula we have used for ρ is exactly the same as the basic formula that we first derived for r on page 395; that is, the concepts are the same. The coefficient of correlation tells us whether or not it is advantageous to use a straight line in estimating values of the dependent variable. The index of correlation tells us whether or not it is advantageous to use some particular curve in estimating values of the dependent variable.

We have, however, made a slight departure from our former practice in fitting the reciprocal curve, and we should make a similar departure if we fitted a logarithmic curve. In fitting the reciprocal curve we found values of the parameters in such a way that we minimized the sum of the squared deviations. But these were not, as before, deviations of the Y's from the average Y. This time they were deviations of the reciprocals of Y from the average reciprocal. Thus we minimized the sum of the squared deviations of the reciprocals from the mean reciprocal. In fitting a logarithmic curve (which we should do just as we fitted the reciprocal curve, save that we should use logs of Y rather than reciprocals of Y), we should minimize the sum of the squared deviations of the logs of Y from the mean log of Y. In such cases one must admit that the least-squares criterion of goodness of fit is no longer being used, since we are not minimizing the sum of the squared deviations in the sense which we originally described. In other words, while there is a theoretical advantage in using the least-squares approach when fitting straight lines and parabolas, this advantage disappears when we fit reciprocal curves, logarithmic curves, and the like. In these cases the use of the normal equations which are derived in such a way as to minimize the sum of the squared deviations is arbitrary and based merely on the fact that they are convenient. Theoretically one may as well use the freehand regression line, and as a matter of fact this has the advantage that the man using it is not so likely to be led astray by misplaced reliance on mathematical computation as he is when using the more formal methods (see last paragraph of Chap. I, pp. 5–6). One can, of course, determine the equation of his freehand regression curve if he wishes by the methods already described for finding the formula of freehand trend lines (see page 278).

14.7. Standard Errors in Curvilinear Correlation.—We should naturally like to test the reliability of our index of correlation to find whether the index is apt to be indicative of the nature of the relationship in the universe or whether it may have arisen from peculiarities of the particular sample studied. For this purpose some authors have suggested the use of the formula

$$\sigma_\rho = \frac{1 - \rho^2}{\sqrt{N - P}}$$

In the case of the reciprocal curve with which we have just been working, we found that $\rho = 0.988$, $N = 13$, and $P = 3$. From these figures and the above formula we should compute $\sigma_\rho = 0.00755$. However, we saw on page 400 that these common formulas are often misleading when applied to small samples or to instances in which the relationship is large. The situation is even worse in curvilinear than in linear correlation. The curvilinear coefficients found in various samples from the same universe are decidedly non-normal in distribution, and the use of ordinary methods of standard or probable error is out of the question. Hotelling has said,[1] "The probable error of the correlation ratio may now be considered as an obsolete concept; the assumption of a normal distribution is in this case an extremely crude approximation." It is possible that the reliability of the curvilinear regression line could be determined by methods called "analysis of variance," but these methods are too advanced for us to take up here. The interested reader is referred to the standard work on the subject by R. A. Fisher.[2] We are probably safe in assuming at least that there is some correlation in the universe when the value of ρ is as high as it is in our present case.[3]

[1] *Journal of the American Statistical Association*, Vol. XXVI, No. 173A, p. 82.

[2] "Statistical Methods for Research Workers," Oliver & Boyd, Edinburgh, 1938.

[3] In this connection we should note that it is rather difficult to define our universe in the present case. We have based our computations on figures showing potato production and prices for the years 1906–1918. But it can hardly be said that these are a random sample of the years before and the years after. They were not chosen by lot from all the years in existence (or from all the years that ever did or ever will exist). Only insofar as the years 1906–1918 are representative of other years can we apply our con-

14.8. Suggestions for Further Reading.—The problems of curvilinear correlation combine the concepts of curve fitting and the concepts of correlation. For this reason the references already cited under these two separate headings (see Secs. 10.21 and 13.18) are again useful here. The standard work on correlation, Mordecai Ezekiel's, "Methods of Correlation Analysis," John Wiley & Sons, Inc., New York, 1930, should again be given especial mention. A quick review of the material covered in Chaps. 10 and 13 of this book may help to clear up doubtful points.

EXERCISES

1. The scatter diagram on page 436 shows the relationship of potato production to the price of potatoes. Make a scatter diagram showing the relationship of potato production to the reciprocal of the price of potatoes. This will be a scatter diagram of the figures in the table on page 451. Note the change in the distribution of points. What does this tell about the possibility of fitting a reciprocal curve to the points?

2. Plot a scatter diagram of the data in the table on page 448 and draw a freehand curve through the points. Compute ρ from the freehand curve, reading your estimated prices from your diagram.

3. Draw sample curves of each of the following equations. Save them for use in determining the type of curve that will describe particular distributions of points on scattergrams.

$$Y = a + bX$$
$$Y = a - bX$$
$$Y = X^2$$
$$Y = a + bX + cX^2$$
$$Y = a - bX + cX^2$$
$$Y = a + bX - cX^2$$
$$Y = a - bX - cX^2$$
$$Y = a + bX + cX^2 + dX^3$$
$$Y = a - bX + cX^2 - cX^3$$
$$Y = \log X$$
$$\log Y = X$$
$$Y = a - \log X$$

clusions to other periods. This fact is but one of the difficulties which arise when we correlate time series. Ordinarily we should also hesitate to correlate times series without first eliminating trends and seasonal movements; otherwise we are likely to get spurious correlation. If we correlate monthly egg prices in New York City with monthly mean temperatures in Bangkok, we shall almost certainly get a sizable coefficient of correlation merely because both series are characterized by seasonal fluctuations, and they are almost certain to be fluctuating either together or in opposite directions. The dangers inherent in correlation problems involving historical data should be evident; it is impossible to give more space to them here.

$$a - \log Y = X$$

$$Y = \frac{1}{X}$$

$$Y = a - \frac{1}{X}$$

$$\frac{1}{Y} = X$$

$$a - \frac{1}{Y} = X$$

In each equation replace a, b, c, and so on, by arbitrary numbers, substitute various values of X (both positive and negative), and solve to find the values of Y. Plot these values of X and Y to see the shape of the curve.

4. How many parameters would be necessary in a parabolic equation to give eight bends to a curve?

5. Using the normal equations of the second- and third-degree parabolas as guides, write the normal equations of the fourth-degree parabola. Students of the calculus should derive the equations to check their results.

6. If, in a given problem, the number of parameters exceeds the number of points on the scatter diagram, what happens to the correction formulas on page 456? Explain.

7. On page 449 we found the regression equation of the second-degree parabola. Using this equation, estimate the price which would accompany a production of 3.3; a production of 3.6; a production of 4.0. What is happening to the price as the production increases? Is this result to be expected a priori? Note the danger of projecting regression lines beyond the extremes of the data studied. At what production would the price be lowest if this equation really described the relationship?

8. On page 449 is a table which was used in computing the constants for the normal equations of the second-degree parabola. Draw up a table such as would be used if we were fitting a third-degree parabola.

9. On page 451 are given two normal equations. Solve them simultaneously and check the results given in the text.

10. On page 454 is a table which includes a column of estimated prices. Using the appropriate equation, estimate these prices and check the figures given in the table.

11. Using the figures in the table on page 454, find S_y and σ_y. Check your results against those given in the text.

12. On page 455 is given the value of ρ computed around the second-degree parabola. The computations are not given. Carry out the computations and check the results.

13. Compare the correction formulas of page 456 with those given for linear correlation on page 398. What is the relationship between them? How many parameters are there in the equation for a straight line?

14. On page 457 we are told that $r = -0.822$. On page 454 we are told that $\sigma_y = 0.7211$. From these two figures compute the value of S_y and compare it with the value given on page 457.

CHAPTER XV

MULTIPLE CORRELATION

15.1. Nature of Multiple Relationships.—We have discovered that it is often possible to make use of a knowledge of the value of one variable when we are trying to estimate the value of another. When the knowledge of one variable makes possible more accurate estimates of the value of another variable than would be possible without it, we have said that the two variables are related or correlated. Yet it is evidently true that when we are trying to make an estimate of the value of some one variable (say a person's weight) we may wish to consider not one other variable but several others. For example, if you are asked to estimate the weight of an unknown person, what data will help you to make your estimate? You will want to know the person's age, height, sex, nationality, etc. Obviously weight is correlated with several other things, not merely with one, and the regression equations which we have used up to this point have enabled us to estimate the value of one variable by substituting the value of but one other variable. How much handier it would be if we could find an equation in which we could substitute figures for both a man's height and his age in making an estimate of his weight!

Problems that involve the determination of the relationship between one variable and several other variables acting together are called problems of *multiple correlation*. The chances are good that most relationships are in fact multiple relationships; most effects probably have multiple causes. In some cases some one of the connected variables is so far and away more important than any of the others that we can neglect the others and determine the relationship of the effect to but one of the causes.[1] For example, we often estimate the period of vibration

[1] Here we fall into the easy circumlocution of "cause and effect" without taking the trouble to explain their meaning. Again we mean merely that a knowledge of one or more variables is helpful in estimating the value of

of a pendulum from the length alone, neglecting variations caused by gravity, air resistance, etc. The length is relatively so important under ordinary circumstances that we feel safe in assuming that the period varies with the length alone. Such a relationship would be one of simple correlation. If we desired greater accuracy, however, and tried to estimate the period of a pendulum from its length, the pull of gravity, and the resistance of the surrounding medium, we should have a problem in multiple correlation.

15.2. Dependent and Independent Variables.—In problems of multiple correlation, we are dealing with situations that involve three or more variables. We are trying to make estimates of the value of one of these variables based on the values of all the others. The variable whose value we are trying to estimate is called the *dependent variable*, and the other variables, on which our estimates are based, are known as the *independent variables*. Again we emphasize, as we did in the case of simple correlation (see pages 374 and 381*n*.) that no problem of cause and effect is involved in the dependence or independence or variables. It is merely a question of the usefulness of one variable in making estimates of another. The variable that we wish to estimate is automatically the dependent variable. We select as independent variables all the other variables which, in our opinion, will be of significant help to us in estimating its value.

It should be obvious that the statistician himself chooses which variable is to be dependent and which variables are to be independent. It is merely a question of the problem being studied. If we are trying to determine the most probable weight of men, we make weight the dependent variable and height, age, etc., the independent variables. If, on the other hand, we are interested in explaining (or estimating) height, we will make height the dependent variable and age, weight, etc., the independent variables.

Problems of multiple correlation always involve three or more variables (one dependent and two or more independents). In order that we may distinguish them easily, we follow the custom of representing them by the letter X with various subscripts.

another variable. As to which variable is the cause and which the effect, or what is the meaning of cause and effect, we leave the problem to texts on logic and philosophy.

The dependent variable is always denoted by X_1, and the others by X_2, X_3, etc. Thus in the height, weight, and age problem which we have just used for purposes of illustration, if we are trying to estimate men's weights (that is, if weight is the dependent variable) we might say:

X_1 = weight in pounds
X_2 = height in inches
X_3 = age in years

Any statistician who read these statements would know that weight was the dependent variable, since its subscript is the number 1.

Other correlation symbols are also changed where necessary. In simple correlation we have designated the coefficient of correlation by r and the index of correlation by ρ. The coefficient of multiple linear correlation is represented by R, and it is common to add subscripts designating the variables involved. Thus $R_{1.234}$ would represent the coefficient of multiple linear correlation between X_1 on the one hand and X_2, X_3, and X_4 on the other. Likewise $\rho_{1.234}$ would represent the index of multiple correlation (used when the relationships are curvilinear) between X_1 on the one hand and X_2, X_3, and X_4 on the other. The subscript of the dependent variable is always to the left of the point. The standard error of estimate when the value of X_1 is computed from the values of X_2, X_3, and X_4 would be represented by $S_{1.234}$. The variable estimated is designated by the number to the left of the point, and the variables on which estimates are based are to the right of the point. The standard deviation of the X_1 variable would be represented by σ_1; and σ_2 and σ_3 would represent the standard deviations of X_2 and X_3, respectively. The arithmetic means of X_1 and X_2 would be shown as $\bar{X}_1$ and $\bar{X}_2$ or as M_1 and M_2, respectively.

15.3. Multiple-regression Equations.—The regression equations which we used in simple linear correlation were rather simple algebraic equations which (after we had evaluated the parameters) had but two unknowns, X and Y. For example, such an equation might be

$$Y = 12 - 3X$$

This equation showed us how to vary our estimate of Y when the value of X varied. In fact, the above equation tells us to reduce

the value of Y three units whenever we increase the value of X one unit.

When we came to parabolas of higher degrees, the regression equations became somewhat more complex because we included in them one or more of the higher powers of X as well as the first power. But still there were but two variables involved, X and Y. For example, such an equation might be

$$Y = 7 - 2X + 5X^2$$

This equation also tells us exactly how to vary our estimates of Y as the value of X varies.

The multiple-regression equation must obviously be altered so that we can account for changes in all the independent variables. The value of the dependent variable is to depend on the values of several other variables. This fact, however, requires no major alteration in the regression equation. It requires merely that we add terms for the new variables. If we are to use two independent variables, with the dependent variable being represented by X_1 and the other two variables by X_2 and X_3, our equation might be something like this:

$$X_1 = 3 + 2X_2 - 3X_3$$

Now if the value of X_2 is 10 and the value of X_3 is 7, we can substitute to get

$$X_1 = 3 + 2(10) - 3(7) = 2$$

In this case we have estimated the value of X_1 from known values of X_2 and X_3. The nature of the problem is exactly the same as in simple correlation.

It will be noted that in the above sample multiple-regression equation we are told that if $X_2 = 0$ and $X_3 = 0$, X_1 will equal 3. (Try substituting 0 for X_2 and X_3, and solving.) We are told likewise that each increase of one unit in X_2 will bring an increase of two units in X_1 and that each increase of one unit in X_3 will bring a decrease of three units in X_1. In other words, the parameters of the multiple-regression equation tell us the same type of thing that we are told by the parameters of the simple-regression equation.

The multiple linear regression equation is always of the general form

$$X_1 = a + b_2X_2 + b_3X_3 + b_4X_4 + \cdots$$

The parameters b_2, b_3, b_4, etc., are called the *regression coefficients*. Strictly speaking, we should give them more subscripts so that we can tell not only with which independent variable they are connected, but also which variable is dependent and what other variables are independent. For example, if we have two independent variables and one dependent variable (the latter being X_1), we should write our complete multiple linear regression equation thus:

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

The parameter $b_{12.3}$ is called the regression coefficient of X_1 on X_2 with X_3 held constant. It tells us the amount by which X_1 will vary for each unit's change in X_2 if there are no changes in the value of X_3. Likewise $b_{13.2}$ tells us the number of units by which X_1 will change for each unit of change in X_3 if there are no changes in X_2 at the same time. Consequently $b_{13.2}$ is called the regression coefficient of X_1 on X_3 with X_2 held constant.

These terms sound formidable, but they represent no ideas which we have not encountered in the case of simple linear correlation except the idea of "holding constant." Obviously we cannot say what will happen to X_1 when we change X_2 unless we know that there are no variations in X_3 at the same time. We do not know what will happen to the period of the pendulum when we change its length unless we hold gravity constant. But if we can hold gravity constant, then we can say that the addition of 1 in. to a pendulum whose present length is 13 in. will have a definite effect on the period. It may be that we shall have to hold two or three or more other factors constant. If we had five independent variables, one of our regression coefficients would be $b_{12.3456}$. This would be the regression coefficient of X_1 on X_2 with X_3, X_4, X_5, and X_6 held constant. It would represent the same type of thing as a statement regarding the effect of changing the length of a pendulum while we held constant the pull of gravity, the temperature, the barometric pressure, and the resistance of the surrounding medium.

If we use all the subscripts, then, our typical regression equation for multiple linear correlation involving two independent variables will be

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

Our problem would be that of evaluating the parameters in order that we might state the equation in a form like this:

$$X_1 = 6 + 17X_2 - 1.543X_3$$

With three independent variables, the equation would become

$$X_1 = a_{1.234} + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

The extension for a greater number of variables is obvious. As long as the relationship is linear, each X (that is, X_2, X_3, etc.) will appear in the equation but once and always in the first power. The major problem is that of determining the values of the regression coefficients and of the a term so that it will be possible to make estimates.

15.4. Types of Relationship.—In the regression equations of which we have just been speaking it is evident that every addition of one unit to any independent variable has the same effect regardless of the size of the dependent variable. Thus, if our regression equation is

$$X_1 = 5 + 2.4X_2 + 1.5X_3$$

it is clear that whenever we add one unit to X_2 (X_3 remaining the same) we increase the value of X_1 by 2.4 units. This is true regardless of the size of X_2. Similarly each increase of one unit in X_3 brings an addition of 1.5 units in X_1 regardless of the size of X_3. When our multiple-regression equation is of this type, we say that the relationship is linear. The equation corresponds to that for simple linear correlation, in which the independent variable appears but once and in the first power. Just as the simple linear regression equation can be represented by a straight line, so the multiple linear regression equation involving two independent variables can be represented by a plane.

For example, suppose that we are studying the effect of variations in temperature and in rainfall on the yield of potatoes. If the relationship is linear, it can be described by some such plane as that in section A of Fig. 15.1. In this chart is shown a solid. The height of the solid at any point represents the value of the dependent variable (potato yield). The scale along the lower right-hand edge represents the values of X_2 (rainfall), and the scale along the lower left-hand edge represents values of X_3 (temperature). It will be noted that the values of X_1

are represented by a plane which slopes from the back corner toward the front corner. As we move along the X_3 scale from left to right (that is, as the value of X_3 increases while the value of X_2 is fixed), the values of X_1 (heights of the solid) diminish.

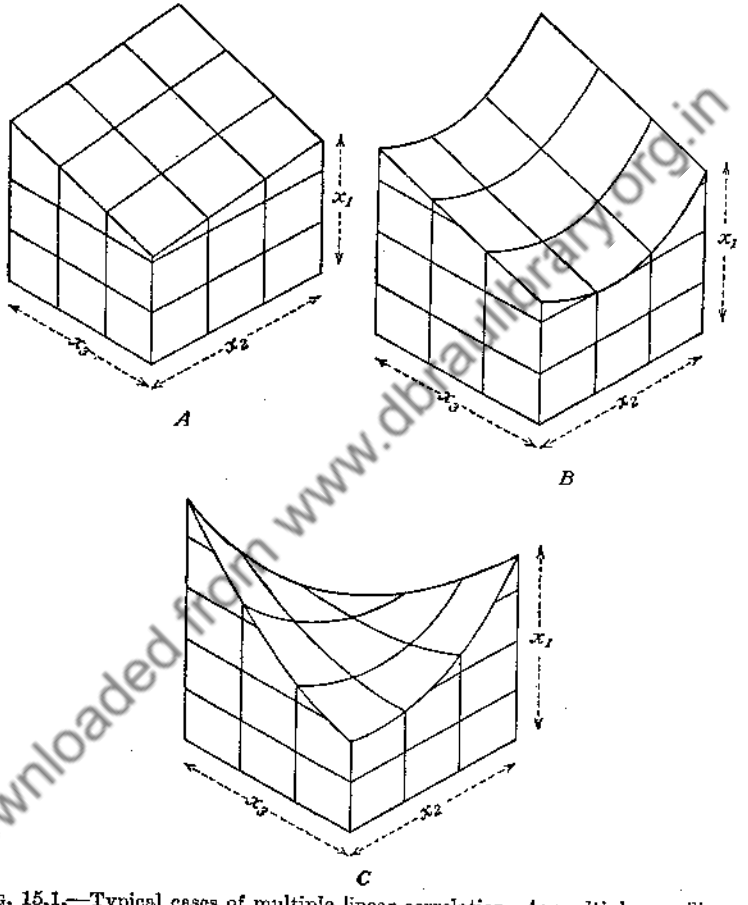


FIG. 15.1.—Typical cases of multiple linear correlation, *A*; multiple curvilinear correlation, *B*; and joint correlation, *C*.

Moreover, they diminish regularly along a straight line, and everywhere along lines of the same slope, no matter at what point we hold X_2 constant.

Now let us move along the X_2 scale from the front of the solid to the back. It will be seen that, as the values of X_2 increase, the value of X_1 likewise increases regularly along a straight line.

Regardless of the point at which we hold X_3 constant, the slope of this line is the same; that is, if we pass many vertical planes through the solid parallel to the X_2 scale, they will all cut the upper plane in parallel straight lines. This is always true of *linear multiple correlation.* Just as the parameter b in the simple linear regression equation $Y = a + bX$ shows the slope of the regression line, so the parameter $b_{12.3}$ in the multiple linear regression equation shows the slope of any line on the regression plane which is parallel to the X_2 scale, and the parameter $b_{13.2}$ shows the slope of any line on the regression plane which is parallel to the X_3 scale. If the correlation is linear, all these lines will be straight—that is, the regression surface (the surface of the solid pictured) will be a plane surface.

Multiple curvilinear correlation exists when the relationship between one or more of the independent variables and the dependent variable is curvilinear. The situation can be pictured as in section *B* of Fig. 15.1. The variables here are the same as before, but now it will be noted that the surface is no longer a plane. It will be seen that as one moves along the X_3 axis from left to right the values of X_1 diminish, and that they diminish everywhere along straight lines which are parallel. Thus, the relationship between X_3 and X_1 is linear, as in the preceding case. But when one moves along the X_2 axis from front to back, the surface is curved. At first the values of X_1 fall somewhat, but thereafter they rise more and more rapidly. It will be noted, however, that no matter where we start to cross the surface, as long as we cross it in a direction parallel to the X_2 axis we shall always pass over the same type of curve. We shall always start by walking downhill for a way, after which we shall start up a hill which becomes increasingly steep. In other words, all the paths across the surface which are parallel to the X_2 axis describe parallel curves. In multiple correlation this is always true. We can generalize by stating that in multiple correlation the relationship between any one independent variable and the dependent variable (all other variables being held constant) is always the same regardless of the point at which the other variables are held constant.

In the present case it looks as though the relationship between X_2 and X_1 when X_3 is constant could be described by a second-degree parabola. All the curves on the surface are the same,

and each has but one bend. Hence the relationship between X_2 and X_1 could presumably be shown by a regression equation of the general type

$$X_1 = a_{1.2} + b_{12}X_2 + b_{12'}X_2^2$$

This is the general formula for the second-degree parabola adapted to our new symbolism. The relationship between X_3 and X_1 can, on the other hand, be depicted by a straight line of the general formula

$$X_1 = a_{1.3} + b_{13}X_3$$

The relationship of X_2 and X_3 together to X_1 could, then, be shown by taking the sum of these two relationships. However, since this operation would give us two constant terms ($a_{1.2}$ and $a_{1.3}$), and since the sum of the two constant terms will likewise be constant, we can substitute the new constant $a_{1.23}$ for the sum of the other two. Now if X_1 is to be estimated from both together, it will be estimated by some such equation as

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{12'.3}X_2^2 + b_{13.2}X_3$$

This would be a regression equation for multiple curvilinear correlation. The equation for the plane surface of multiple linear correlation would be of the general form

$$X_1 = a_{1.23} + b_{12.3}X_2 + b_{13.2}X_3$$

It will be seen that this is the same type equation that we mentioned on page 466.

There still remains one type of correlation to be described. It is quite possible that rainfall and temperature will both be related to potato yield, but that the relationship between rainfall and potato yield will vary according to the temperature. It is possible that greater rainfalls might be advantageous if the temperature were high but disadvantageous if the temperature were low. Thus the line or curve showing the relationship between rainfall and yield, with temperature constant, would differ according to the point at which we held the temperature constant. In such a case we should say that there was *joint correlation* between the variables. Such a situation is pictured in section *C* of Fig. 15.1. It will be noted that, as one passes along the right-hand front edge of the surface parallel to the X_2

axis, the values of X_1 increase. When one passes along the left-hand back edge of the surface parallel to the X_2 axis, the values of X_1 decrease. If we let X_1 represent potato yield, X_2 the rainfall, and X_3 the temperature as before, this is equivalent to saying that with high temperatures increasing rainfall increases the yield, but with low temperatures increasing rainfall lowers the yield. In such a case the surface is *warped*, and we say that the correlation is *joint*. Joint correlation exists, then, when the nature of the relationship between an independent variable and the dependent variable differs according to the size of some other independent variable, just as here the relationship between potato yield and rainfall differs at different temperatures. There is no simple type of regression equation which we can use to describe the joint correlation surface, and the methods of handling problems of joint correlation must be left to more advanced treatises.[1]

We can classify relationships, then, as follows:

1. Simple correlation (but two variables involved).
   a. Linear (regression line straight).
   b. Curvilinear (regression line curved).
2. Multiple correlation. (Two or more independent variables. Relationship of each independent to the dependent constant regardless of size of other independents.)
   a. Linear (described by a plane).
   b. Curvilinear (described by a regular curved surface).
3. Joint correlation. (Two or more independent variables. Relationship of one or more of the independent variables with the dependent variable varies according to the value of other independent variables. Described by warped surface.)

15.5. Methods of Computation.—It will be impossible for us to give a detailed description of multiple-correlation analysis here, but we can sketch briefly the processes involved.

The regression equation for linear multiple correlation is discovered by solving normal equations which are similar to, and derived by the same process as, the normal equations of simple correlation. If we have two independent variables, X_2 and X_3, and if the dependent variable is, as is customary, represented by X_1, the normal equations for linear relationships are

[1] See especially M. J. B. Ezekiel, "Methods of Correlation Analysis," Chaps. 21 and 22, John Wiley & Sons, Inc., New York, 1941.

$$Na + b_2\Sigma X_2 + b_3\Sigma X_3 = \Sigma X_1$$
$$a\Sigma X_2 + b_2\Sigma X_2^2 + b_3\Sigma(X_2X_3) = \Sigma(X_1X_2)$$
$$a\Sigma X_3 + b_2\Sigma(X_2X_3) + b_3\Sigma(X_3^2) = \Sigma(X_1X_3)$$

One must find the values of N, ΣX_1, ΣX_2, ΣX_3, $\Sigma(X_2^2)$, $\Sigma(X_3^2)$, $\Sigma(X_1X_2)$, $\Sigma(X_1X_3)$, and $\Sigma(X_2X_3)$. These values are substituted in the normal equations, and the equations are then solved for values of a, b_2, and b_3. The two latter parameters are really the coefficients of regression $b_{12.3}$ and $b_{13.2}$, but here we are using the shorter form. We can illustrate the method of applying the equations by solving the hypothetical case in Table 15.1. We shall use but five sets of observations, although no one would in practice apply such complicated methods to so few cases. Here the figures are given merely for illustrative purposes. One can assume, if one wishes, that X_1 is potato yield,

TABLE 15.1.—COMPUTATION OF MULTIPLE LINEAR REGRESSION EQUATION

X_1	X_2	X_3	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3
1	3	2	9	4	3	2	6
2	4	4	16	16	8	8	16
3	2	9	4	81	6	27	18
4	1	13	1	169	4	52	13
5	5	12	25	144	25	60	60
15	15	40	55	414	46	149	113

X_2 is rainfall, and X_3 is temperature, as in the illustrations on page 467. Substituting these totals in the normal equations (and remembering that $N = 5$), we get

$$5a + 15b_2 + 40b_3 = 15$$
$$15a + 55b_2 + 113b_3 = 46$$
$$40a + 113b_2 + 414b_3 = 149$$

Solving these three equations simultaneously, we have the following values of the parameters:

$$a = -0.66666\ldots$$
$$b_2 = +0.33333\ldots$$
$$b_3 = +0.33333\ldots$$

Substituting these values in the type equation, we get

$$X_1 = -0.6667 + 0.3333X_2 + 0.3333X_3$$

In this case the equation could more easily be put in fractional form, thus:

$$X_1 = -\frac{2}{3} + \frac{X_2}{3} + \frac{X_3}{3}$$

If we have three independent variables, the normal equations become

$$\begin{aligned} Na + b_2\Sigma X_2 + b_3\Sigma X_3 + b_4\Sigma X_4 &= \Sigma X_1 \\ a\Sigma X_2 + b_2\Sigma(X_2^2) + b_3\Sigma(X_2X_3) + b_4\Sigma(X_2X_4) &= \Sigma(X_1X_2) \\ a\Sigma X_3 + b_2\Sigma(X_2X_3) + b_3\Sigma(X_3^2) + b_4\Sigma(X_3X_4) &= \Sigma(X_1X_3) \\ a\Sigma X_4 + b_2\Sigma(X_2X_4) + b_3\Sigma(X_3X_4) + b_4\Sigma(X_4^2) &= \Sigma(X_1X_4) \end{aligned}$$

The extension of these equations for a larger number of independent variables is obvious.

The method of handling curvilinear multiple correlation depends on the type of curve to be fitted. One would have to determine this, as in other cases, by preliminary classification of the data. It would be necessary, however, to choose cases in which the values of X_3 (for example) were about equal, and to determine by group averages or scatter diagrams of these cases the type of curve that described the relationship of X_2 and X_1. If a parabolic curve could be used, the method would be a simple extension of that just given. For example, if we have a relationship similar to that pictured in section *B* of the chart on page 468, we know that the relation between X_1 and X_2 (when X_3 is constant) can be represented by a second-degree parabola, while the relationship between X_1 and X_3 (when X_2 is constant) can be represented by a straight line. As was pointed out on page 470, under such circumstances the type equation will be of the general form

$$X_1 = a + b_2X_2 + b_{2'}X_2^2 + b_3X_3$$

The normal equations for such a surface are

$$\begin{aligned} Na + b_2\Sigma X_2 + b_{2'}\Sigma(X_2^2) + b_3\Sigma X_3 &= \Sigma X_1 \\ a\Sigma X_2 + b_2\Sigma(X_2^2) + b_{2'}\Sigma(X_2^3) + b_3\Sigma(X_2X_3) &= \Sigma(X_1X_2) \\ a\Sigma(X_2^2) + b_2\Sigma(X_2^3) + b_{2'}\Sigma(X_2^4) + b_3\Sigma(X_2^2X_3) &= \Sigma(X_1X_2^2) \\ a\Sigma X_3 + b_2\Sigma(X_2X_3) + b_{2'}\Sigma(X_2^2X_3) + b_3\Sigma(X_3^2) &= \Sigma(X_1X_3) \end{aligned}$$

Again we shall illustrate with a short hypothetical example. Letting X_1 represent potato yield, X_2 the rainfall, and X_3 the temperature, suppose that we obtain the data of Table 15.2.

TABLE 15.2.—COMPUTATION OF MULTIPLE CURVILINEAR REGRESSION EQUATION

X_1	X_2	X_3	X_2^2	X_2^3	X_2^3	X_2^4	X_2X_3	X_1X_2	$X_2^2X_3$	$X_1X_2^2$	X_1X_3
9	1	2	1	4	1	1	2	9	2	9	18
12	2	3	4	9	8	16	6	24	12	48	36
13	3	6	9	36	27	81	18	39	54	117	78
32	4	1	16	1	64	256	4	128	16	512	32
29	5	8	25	64	125	625	40	145	200	725	232
95	15	20	55	114	225	979	70	345	284	1411	396

Substituting these totals in the normal equations, we have

$$\begin{aligned} 5a + 15b_2 + 55b_{2'} + 20b_3 &= 95 \\ 15a + 55b_2 + 225b_{2'} + 70b_3 &= 345 \\ 55a + 225b_2 + 979b_{2'} + 284b_3 &= 1411 \\ 20a + 70b_2 + 284b_{2'} + 114b_3 &= 396 \end{aligned}$$

If we solve these equations we find the following values of the parameters:

$$\begin{aligned} a &= 10 & b_{2'} &= 1 \\ b_2 &= 2 & b_3 &= -2 \end{aligned}$$

The regression equation, then, becomes

$$X_1 = 10 + 2X_2 + X_2^2 - 2X_3$$

With more complicated curvilinear relationships it becomes necessary to derive the necessary normal equations each time for the particular problem in hand. With the aid of the calculus and the examples here given, this should not be a difficult task.

The constants of the regression equation are interpreted as in simple correlation. In the numerical illustration of multiple linear correlation on page 473, we found the equation

$$X_1 = \frac{-2}{3} + \frac{X_2}{3} + \frac{X_3}{3}$$

In terms of decimals this gives approximately

$$X_1 = -0.67 + 0.33X_2 + 0.33X_3$$

The first constant, a, is equal to -0.67. This tells us that X_1

will have a value of -0.67 when the two independent variables are both equal to 0. In the terms of our problem this would mean that with a rainfall of 0 and a temperature of 0 the potato crop would be -0.67 bushels. This makes no sense. In such cases we think of this constant merely as one which defines the height of the regression plane, and we do not try to give it further significance. The next constant ($b_{12.3}$) is 0.33. This tells us that every increase of one unit in the size of X_2 brings an increase of one-third of a unit in the size of X_1 if there is no change in the size of X_3. That is, if temperature is constant, each additional inch of rainfall will add one-third of a hundred bushels to the potato yield (assuming now that 100 bushels is the unit in which potato yield is measured). Similarly the third constant ($b_{13.2}$) is 0.33. This means that each unit of increase in the size of X_3 (X_2 remaining fixed) is accompanied by an increase of one-third of a unit in the value of X_1. In the terms of our problem, if we consider only changes in temperature, with no variations in rainfall, each additional degree of temperature is accompanied by an increase of one-third of a hundred bushels (that is, one-third of a unit) in the size of the potato yield.

15.6. Effects of Variables Separately.—Here rainfall and temperature seem equal in their effect on yield, since a change of one unit in either is accompanied by a change of a third of a unit in yield. Their effects seem to be exactly the same. But it may be true that a difference of 1° in temperature from season to season is a very minor change, while a difference of 1 in. in rainfall is large. In other words, the fact that a change of one unit in the one brings the same result as a change of one unit in the other does not tell us enough. It is hard to compare a change of 1° in temperature with a change of 1 in. in precipitation. We have discovered, however, that measurements which were originally taken in different units can be compared if stated each in terms of its own standard deviation (see page 171). It is not uncommon so to state the regression coefficients; in this way we make them comparable. When regression coefficients are stated in terms of their standard deviations, they are called *beta coefficients*, and they are represented by the Greek letter beta followed by whatever subscripts followed the regression coefficient. For example, $\beta_{12.34} = b_{12.34}(\sigma_2/\sigma_1)$. Similarly $\beta_{13.2} = b_{13.2}(\sigma_3/\sigma_1)$. In our recent short numerical example we found the following

values of the regression coefficients:

$$b_{12.3} = 0.33$$
$$b_{13.2} = 0.33$$

Inspection of the original data in the table on page 472 shows that the standard deviations of the variables are

$$\sigma_1 = 9.5$$
$$\sigma_2 = 1.4$$
$$\sigma_3 = 2.6$$

In this hypothetical case, then, a variation of one unit in X_3 is much more likely to occur than is a variation of one unit in X_2. If we put the regression coefficients in terms of the standard deviations, we get the following beta coefficients:

$$\beta_{12.3} = 0.33\left(\frac{1.4}{9.5}\right) = 0.049$$
$$\beta_{13.2} = 0.33\left(\frac{2.6}{9.5}\right) = 0.091$$

The first of these figures tells us that each increase of one standard deviation in the value of X_2 will be accompanied (if X_3 stays constant) by an increase of 0.049 standard deviations in the value of X_1. The second of the beta coefficient tells us that an increase of one standard deviation in the value of X_3 (X_2 staying constant) will be accompanied by an increase of 0.091 standard deviations in the value of X_1. Since a standard deviation of change is equally likely to occur in X_2 and in X_3 (if the distributions are normal), it is evident that X_3 probably accounts for much more of the actual change in X_1 in our problem than does X_2. In other words, variations in temperature have more effect than variations in rainfall in bringing about changes of potato yield. The multiple-regression equation may be written in terms of the beta coefficients rather than in terms of the coefficients of regression,[1] and the beta coefficients also become

[1] For example, the regression equation for multiple linear correlation with two independent variables is equally written

$$X_1 = a + b_2X_2 + b_3X_3$$

It may be written in terms of the beta coefficients thus (where k is a constant):

$$\frac{X_1}{\sigma_1} = \beta_{12.3}\frac{X_2}{\sigma_2} + \beta_{13.2}\frac{X_3}{\sigma_3} + k$$

very useful in more advanced statistical work. Here we note merely the fact that they are useful in determining the relative importance of the various independent variables. The relative importance of the several independent variables is also shown by coefficients of partial correlation and by coefficients of part correlation. It is impossible for us to describe these coefficients here; the interested student is referred to more advanced works.[1]

15.7. Other Correlation Constants.—Having covered briefly the multiple regression equation, we now turn to the correlation coefficients or indices and the standard errors of estimate. These are determined by methods already familiar to us (see pages 453–454*ff.*). Having found our regression equation, we estimate each value of X_1 from the known values of X_2, X_3, etc. We compare the estimated and the actual values of X_1 and compute the differences between them. The standard deviation of these differences is the standard error of estimate. If we have computed the linear multiple-regression equation, we usually denote the standard error of estimate thus:

$$S_{1.234}$$

This would mean, "The standard error of estimating X_1 from X_2, X_3, and X_4." The first subscript indicates the dependent variable, and the subscripts following the decimal point indicate the independent variables. If the regression surface is curved, the standard error of estimate is denoted by the symbol

$$S_{1.f(234)}$$

The letter *f* signifies that we have used some function of variables 2, 3, and 4, but does not state what function. For further information one would have to consult the regression equation itself, which would usually be given.

After we have found the standard error of estimate, it is easy to find the coefficient of multiple correlation (if the relationship is linear) or the index of multiple correlation (if the relationship is curvilinear). The same formula is used for either:

$$R_{1.234} = \sqrt{1 - \frac{S^2_{1.234}}{\sigma_1{}^2}}$$

$$\rho_{1.234} = \sqrt{1 - \frac{S^2_{1.f(234)}}{\sigma_1{}^2}}$$

[1] Especially good is Ezekiel, *op. cit.*

Thus the coefficient of multiple correlation (or the index of multiple correlation) is based on a comparison of the variability around the regression line (the errors made in estimating by means of the regression equation) and the variability about the mean of the dependent variable. These coefficients tell us, as have the others which we have studied, the extent to which our errors of estimate are reduced if the estimate is based on the regression equation rather than on chance.

Enough has been said here as to the methods of computing the constants of multiple-correlation problems to give the student a fairly good idea of the concepts involved. It is not hoped to do more. Methods have been developed by which a considerable part of the mathematical work of multiple-correlation problems can be saved, and by which the accuracy of the work can be checked step by step. These methods are of tremendous importance to anyone who goes at the problems of multiple correlation seriously. Other sources must be consulted for a description of these methods.[1] The purpose of the present chapter is merely that of acquainting the student with the concepts of multiple and joint correlation in the hope that he will understand simple correlation and the concepts of relationship, in general, better for having taken this brief journey into more complicated fields.

15.8. Corrections for Numbers of Cases and Parameters.—In multiple-correlation problems the corrections for the number of cases and the number of parameters are as important as before. In fact, the corrections are likely to be larger in these cases because the addition of variables makes for the addition of parameters. The formulas by which the unadjusted coefficients are corrected are the same as those given heretofore on page 456.

15.9. Standard Errors of Coefficients of Multiple Correlation.—The standard errors of coefficients of multiple correlation or indices of multiple correlation are computed by formulas similar to those used before, and their interpretation is unchanged. For example:

$$\sigma_{R_{1.234}} = \frac{1 - R^2_{1.234}}{\sqrt{N - P}}$$

$$\sigma_{\rho_{1.234}} = \frac{1 - \rho^2_{1.234}}{\sqrt{N - P}}$$

[1] Especially helpful is H. A. Wallace and G. W. Snedecor, Correlation and Machine Calculation, *Iowa State College Bulletin* 35, 1931.

The probable errors are, as before, found by multiplying the standard errors by 0.6745.

Here, as before, we are troubled by the fact that standard errors computed by these formulas are misleading unless the numbers of cases are very large. As the number of parameters is increased, it becomes more and more important that the number of cases studied be sizable. The large amount of arithmetical work involved makes it certain that statisticians will seldom work with more than five or six independent variables, although cases can be found where many more have actually been used. Possibly we can make a rough statement about the reliability of multiple-correlation results, without being too far from the facts, if we say that in problems where the number of cases runs from 50 to 100 or more (preferably more) and where the number of independent variables is not over four or five, coefficients of multiple correlation greater than $R = 0.5$ are probably significant. However, such crude rules of thumb are seldom satisfactory, and should be used merely for a first rough check. For more accurate tests we fall back on the method of analysis of variance, which is based on a comparison of the dispersion in the original data and the dispersion around the regression surface. Unfortunately the methods involved are too complicated for us to take up here.[1]

15.10. Suggestions for Further Reading.—The problems of multiple correlation have been covered here in but the briefest summary form, in an attempt to give the elementary student some idea as to the possibilities of the method without making him expert in its application. Any attempt to apply these methods to actual problems should be preceded by further reading and study. The most helpful books are listed in Secs. 13.18 and 14.8.

EXERCISES

1. Give examples from the fields of physics, geometry, and other fields of simple and of multiple relationships. For example, if C represents the circumference of a circle and if D represents its diameter, we are told that

$$C = 3.1416D$$

This is an example of simple linear correlation. Find others, both simple and multiple.

[1] For complete treatment see R. A. Fisher, "Statistical Methods for Research Workers," Oliver & Boyd, Edinburgh, 1938; or Wallace and Snedecor, *op. cit.*

2. In the example given in the preceding exercises, the circumference is treated as the dependent variable. How can you tell that this is true? Is the circumference any more dependent on the diameter than the diameter is on the circumference? If you were to treat the diameter as dependent, how would you change the statement of the problem?

3. Suppose we are studying a problem in which there are three variables, as follows:

X_1 = yield of milk in pounds
X_2 = pounds of grain fed per cow per day
X_3 = age of cow in years

Suppose that we find the following regression equation:

$$X_1 = 3 + 2.5X_2 - 0.1X_3$$

a. Exactly what is the meaning of each of the three numbers in the regression equation?

b. Suppose that we have the following values:

$$\sigma_1 = 4 \qquad \sigma_2 = 2 \qquad \sigma_3 = 1$$

What are the beta coefficients? Interpret them. Does age or grain ration play the larger part in the fluctuations of milk yield?

4. In a problem of linear multiple correlation we find the value of

$$b_{12.3} = 3.65$$

a. What is the meaning of each of the subscripts?

b. What is meant by "holding a factor constant"?

c. Interpret the figure 3.65 above.

5. Describe in a paragraph each of the sections of the chart on page 468.

6. The formula showing the space (s) passed over by a falling body in a given time (t) under various gravitational attractions (g) is

$$s = \tfrac{1}{2}gt^2$$

Is the relationship linear, curvilinear, or joint? Test it to see which it is by holding g constant at two widely different values and solving for various values of t. Plot the results. Do they give the same curve?

7. On page 467 is the type equation for multiple linear correlation with three independent variables. Write the type equation for four independent variables.

8. Using the normal equations given in the text as samples (see pages 472 and 473), write the normal equations for multiple correlation with four independent variables.

9. In the table on page 472 are the figures for a multiple-correlation problem. The problem was solved with X_1 as the dependent variable. Suppose that we wish to use X_3 as the dependent variable (that is, to substitute the column heading X_1' for the heading X_3). Work out the multiple linear regression equation for these same figures but with the new dependent variable. Most of the needed totals are already computed in the table.

10. On page 472 are given the data of a multiple-correlation problem. On page 473 is given the regression equation for these data. Estimate the values of X_1 from the regression equation, find the differences between the actual and the estimated values of X_1, and compute $S_{1.23}$ and $R_{1.23}$.

11. In the light of the correction formula on page 456 and the regression equation on page 473, explain why one would not use such complicated methods with but five cases, as in the illustration.

12. The figures of Table 15.3 show the average weight of men of various heights and ages.[1] This is obviously a case of either multiple or joint correlation, since there are two independent variables (age and height) and one dependent variable (weight). Cross-classifications such as this would be used by a statistician to determine the nature of the relationship.

TABLE 15.3.—AVERAGE WEIGHT OF MEN OF VARIOUS HEIGHTS AND AGES

Age Group	Height (feet and inches)							
	5′	5′2″	5′4″	5′6″	5′8″	5′10″	6′0″	6′2″
15–19	113	118	124	132	140	148	158	168
20–24	119	124	131	139	146	154	163	173
25–29	124	128	134	142	150	158	169	181
30–34	127	131	137	145	154	163	174	186
35–39	129	133	140	148	157	167	178	191
40–44	132	136	142	150	159	169	181	194
45–49	134	138	144	152	161	171	183	197
50–54	135	139	145	153	162	172	184	198

a. Is the relationship linear or curvilinear? Plot the figures of one or two of the columns on cross-section paper. Do they yield straight lines or curves? (Minor random variations from straight lines would not give evidence of curvilinearity. Look for regular and consistent deviations from a straight line.) Plot the data of two or three of the rows of figures. Do they yield straight lines or curves? Are the data linear in both directions, in one direction, or in no direction?

b. Is the relationship multiple or joint? Test by plotting the data from various rows on the same piece of cross-section paper, seeing whether they yield approximately parallel lines. Plot also the data of the various columns on another piece of cross-section paper and see if they yield approximately parallel lines. If in both cases the lines are approximately parallel, the relationship is multiple. If in either case (or both cases) there is a tendency for the shape of the lines to change as we go from row to row or column to column, the relationship is joint.

c. If in either case there are consistent and parallel curves, what is their general nature? Are they parabolic or logarithmic, or do they resemble any of the type curves we have plotted elsewhere?

[1] Based on figures in "World Almanac," p. 809, 1934.

d. Look up the corresponding data for women's weights and test them similarly. (The "World Almanac" usually carries these figures. Or suitable figures can be copied from the chart on the scales next time you get weighed.)

e. Explain how we "hold one factor constant" in this table; that is, how we can tell what happens to weight when we vary height but hold age constant. How can we tell what happens to weight when we vary age but hold height constant?

CHAPTER XVI

TABULATION AND GRAPHIC PRESENTATION

16.1. The Use of Tables.—The statistician arranges his data in tabular form to save space and to make it easy to find particular items or to make desired comparisons. One who was interested in studying the growth of population in the United States could, if he wished, record his data thus:

> The population of the United States at the time of the 1790 Census was 3,929,214. By 1800 it had grown to 5,308,483. In 1810 the population had become 7,239,881. . . .

Such a system of recording data would be wasteful of space, and it would also require considerable searching to find the figures which were needed. It is much easier to assemble comparable figures in tabular form and give the above information as in Table 16.1.

TABLE 16.1.—UNITED STATES POPULATION (OFFICIAL CENSUSES), 1790–1930

Year	Population
1790	3,929,214
1800	5,308,483
1810	7,293,881
1820	9,638,453
1830	12,866,020
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	38,558,371
1880	50,155,783
1890	62,947,714
1900	75,994,575
1910	91,972,266
1920	105,710,620
1930	122,775,046

16.2. The Form of Tables.—Although it is not absolutely necessary that any one form of table be adhered to uniformly, nevertheless there are some practices which have become common

in the making of tables. In many cases the conventions which have been adopted have good reason for their use, although sometimes it must be admitted that they have been determined more or less by chance.

In ruling tables it is common to use a double line at the top, between the title and the figures of the table. This is often the only double line used, although if the table is divided into distinct sections which should be separated sharply it is not uncommon to use double lines for the purpose (see Table 3.2, page 25). There is always a single line under the column headings, and there are single lines between columns. No line is used at the extreme right or left of the table; that is, the first and last columns are left open at the outside. If a heading of the table applies to two or more columns, it is separated by a single line from subordinate headings which apply to a smaller number of these columns. The makeup of the table can best be understood by reference to Table 16.2.

TABLE 16.2.—THE PLANETS AND THE SOLAR SYSTEM[1]

Planet	Sidereal Revolution (days)	Distance from Sun (millions of miles)		Distance from Earth (millions of miles)	
		Maximum	Minimum	Maximum	Minimum
Mercury	88.0	43.3	28.6	136	50
Venus	224.7	67.7	66.7	161	25
Earth	365.3	94.5	91.3	.....	.....
Mars	687.0	154.8	128.3	248	35
Jupiter	4,332.6	506.7	459.9	600	367
Saturn	10,759.2	935.6	836.7	1,028	744
Uranus	30,685.9	1,866.8	1,698.8	1,960	1,606
Neptune	60,187.6	2,817.4	2,769.6	2,910	2,677

[1] Based on figures in "World Almanac," p. 73, *New York World-Telegram*, New York, 1934.

It will be noted first that the orderly arrangement of the table makes it very easy to find any particular figure, and also very easy to make comparisons. As to rulings, one will note the double line under the title, the open ends of the table, the single line at the bottom, and the line under the column headings. The words "Distance from Sun" refer to two different columns, and these words are separated from the specific headings of these two columns ("Maximum" and "Minimum") by a single line.

It will also be noted that the vertical lines which divide the columns from one another do not in all cases extend to the top of the table. Each vertical line is carried as far as possible without dividing a heading which belongs to several columns. In this way the upper part of the table is divided into "boxes" of various shapes, and the headings in these boxes are called the *box headings* of the table. The left-hand column of the table is called the "stub," and the headings contained therein are called the *stub headings*. Thus in Table 16.2 the box headings are

Planet	Sidereal Revolution (days)	Distance from Sun (millions of miles)		Distance from Earth (millions of miles)	
		Maximum	Minimum	Maximum	Minimum

The stub headings of this table are

Mercury.....
Venus.....
Earth.....
Mars.....
Jupiter.....
Saturn.....
Uranus.....
Neptune.....

16.3. Order of Headings.—The box headings should be arranged like the items in the outline of a book; that is, there should be major headings with the proper subheadings under them. The entries in the stub of the table should always be arranged in some definite and obvious order. In Table 16.2 we could have arranged the planets in alphabetical order, in the order of their discovery, in order of the number of letters in their names, or by some other arrangement. As a matter of fact, they are there arranged in the order of their distances from the sun. When there are several possible arrangements for the entries in a table, the arrangement that will be most useful should always be used if feasible. Few people are interested in comparing the planets with respect to the length of their names; many people are interested in comparing them with respect to their distances from the sun. Hence distances from the sun offer a much better basis for classification in the stub.

Stub entries are usually arranged chronologically, geographically, alphabetically, numerically, in order of size or importance, or on some similar basis. If the table is to be as useful as possible, and if we are to get from it the maximum possible saving of time, it is imperative that the box headings be arranged so as to make easy the comparison of those things most likely to be compared, and that the stub headings be classified and arranged in the most convenient manner.

It is always extremely helpful to have each table accompanied by a concise but complete title, so that people using the table can tell from the title exactly what they can find in the body of the table. It is never possible, of course, to include in the title all the information in the table, but the more information that can be included without making the title itself unwieldy or confusing, the better. It takes considerable skill and practice to write good titles for tables, and good titles are a part of the advantage of tabulation (saving of time).

16.4. The Purpose of Graphic Presentation.—When data are presented graphically, it is with the purpose of appealing to the eye. Many people learn better through the visual approach than through other approaches. In addition the visual approach is usually very much faster. Graphic presentation attempts to transfer ideas quickly, but unless they are transferred correctly they are better not transferred at all. Hence we must include accuracy as well as speed among the prerequisites of good graphic method.

16.5. Standards for Graphic Presentation.—The principal rules to be followed in making graphs are well stated and illustrated in the report of the Joint Committee on Standards for Graphic Presentation.[1] The charts and directions on pages 487–489, have

[1] These rules, published in 1915, were promulgated by a committee representing 15 scientific societies and 2 government bureaus. The Joint Committee's report appears in the *Publications of the American Statistical Association*, December, 1915. In April, 1936, the Sectional Committee on Standards for Graphic Presentation, sponsored by the American Society of Mechanical Engineers and including in its membership representatives of many learned societies, industrial concerns, and government bureaus, released a tentative draft of a "Code of Preferred Practice for Graphic Presentation: Time Series Charts," This code, which is too lengthy and detailed for inclusion here, should be consulted by any specialist in the graphing of time series.

Fig. 1a Fig. 1b Fig. 1c

1. When possible, linear magnitudes should be used (Fig. 1a) rather than areas (1b) or volumes (1c).

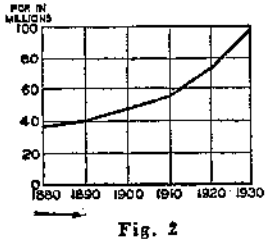


Fig. 2

2. The general arrangement should proceed from left to right (Fig. 2).

3. If the graph is to show sizes directly, the complete scale and zero line should be included (Fig. 3).

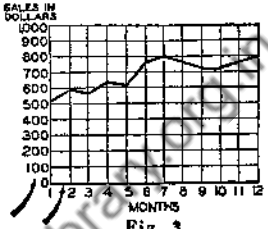


Fig. 3

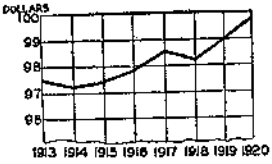


Fig. 4a

4. If the complete scale is not included, the fact should be emphasized by a horizontal break as in Fig. 4a or 4b.

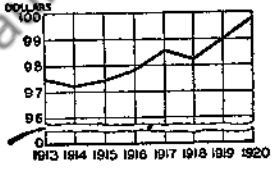


Fig. 4b

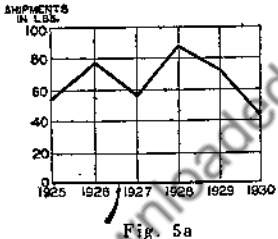


Fig. 5a

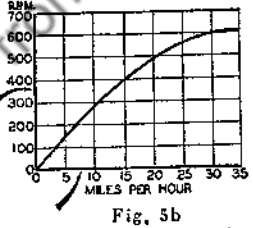


Fig. 5b

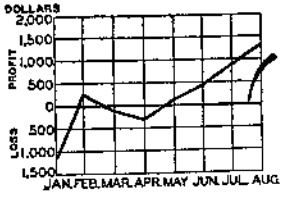


Fig. 5c

5. The zero lines of curve scales should be sharply distinguished from the other coordinate lines (Figs. 5a, 5b and 5c). In showing time series, the first vertical line should not be emphasized as a zero line since it does not represent the beginning of time (Fig. 5a).

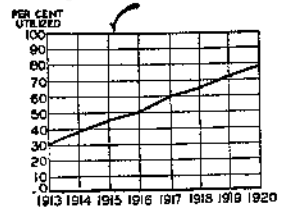


Fig. 6a

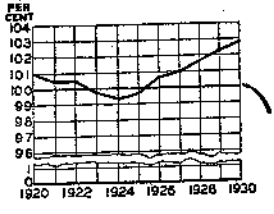


Fig. 6b

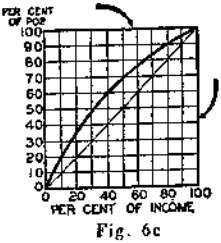


Fig. 6c

6. When the 100 per cent line is used as a basis of comparison, it should be emphasized (Figs. 6a, 6b and 6c).

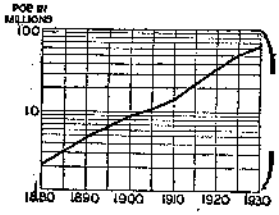


Fig. 7a

7. When a logarithmic chart is made, the top and bottom limiting lines should each be at some power of ten if convenient (Fig. 7a). If not convenient, as when only a small portion of the range is used (Fig. 7b), any other lines may be used as top and bottom limits.

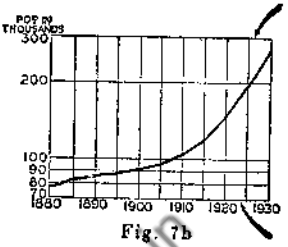


Fig. 7b

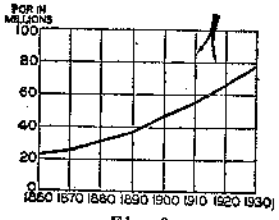


Fig. 8a

8. It is advisable to show no more coordinate lines than are necessary to guide the eye in reading the diagram. Fig. 8a is easier to read than Fig. 8b.

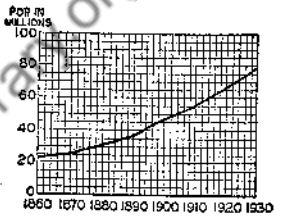


Fig. 8b

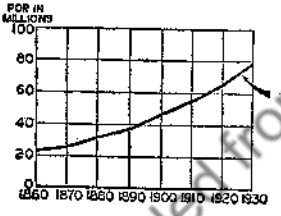


Fig. 9

9. The curves should be sharply distinguished from the background of coordinate lines (Fig. 9). Since the attention is centered on the curve, it should be made more prominent than the zero line as shown in Fig. 9.

10. When lettering cannot be made to read horizontally as in Fig. 9, it should be made to read from bottom to top as in Fig. 10—never from top to bottom. That is, if it cannot be placed like this: ABC, it should be placed like this: ABC, but never like this: CBA.

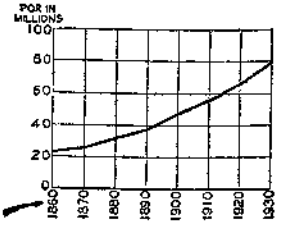


Fig. 10

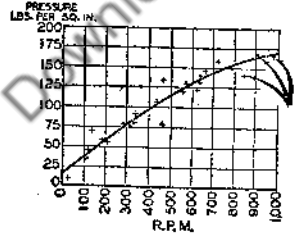


Fig. 11a

11. When irregular or scattered data are plotted, it is often desirable to indicate the points representing the different observations as in Figs. 11a and 11b.

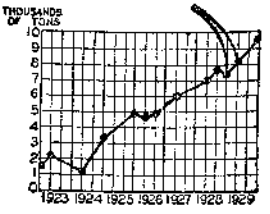


Fig. 11b

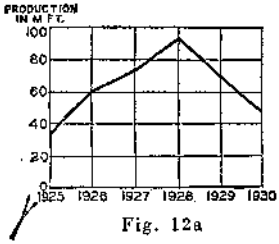


Fig. 12a

12. The scales should be placed at the left and at the bottom of the graph as in Figs. 12a and 12b. If desirable they may be placed at the top and on the right also, but they should never be omitted on the left and on the bottom.

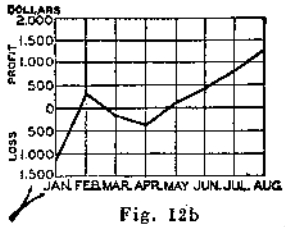


Fig. 12b

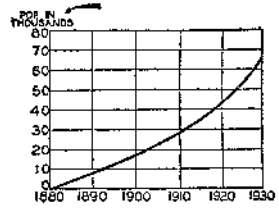


Fig. 13a

13. A caption describing the units used in the vertical scale should be placed at the top of the scale as in Fig. 13a. (Another method is to place the designation along the side but this is less preferable.) When desirable, the horizontal scale should be designated as in Fig. 13b. (Ordinarily it is not necessary to designate years or months.)

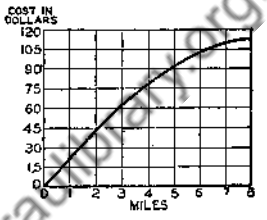


Fig. 13b

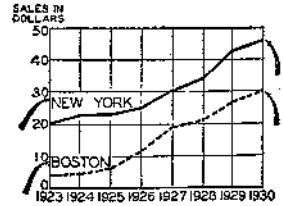


Fig. 14a

14. When two or more curves are presented, they should be differentiated (solid, dashed, dotted lines, etc.) and properly designated (see Figs. 14a and 14b). The method of placing the designations along the curves (14a) is preferred.

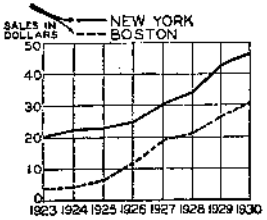


Fig. 14b

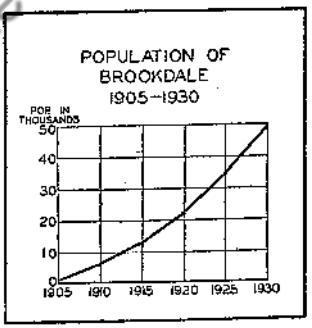


Fig. 15

15a. The title of a chart should be placed at the top as in Fig. 15. It should be as clear and complete as possible. Subtitles, descriptions and footnotes should be added if necessary to insure clearness.

15b. Usually it is desirable to place a border line around the entire chart to give it a finished appearance (Fig. 15).

been adapted with slight changes from the original report, and are included here by kind permission.[1] For more detailed suggestions on graphic presentation the student is referred to specialized texts in that field.

16.6. Graphs on Nonarithmetic Scales.—Students of statistics are usually familiar with graphs made on common cross-section paper. They have seen them in popular magazines and newspapers, and they have made and used them in courses in mathe-

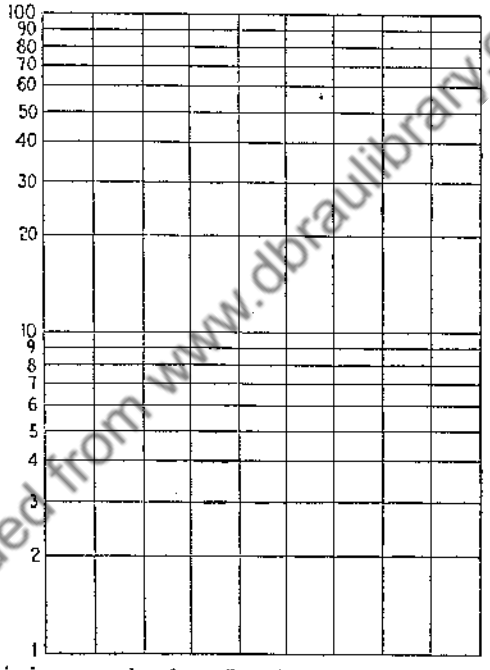


FIG. 16.1.—This is a sample of semilogarithmic paper. Note especially the vertical scale.

matics. Many students do not know, however, that useful charts can be made on other types of scales.

On common cross-section paper (also called "graph paper") a given distance in any direction always (in any one problem) represents some given quantity. One inch on the horizontal axis, for instance, may always represent one year, while 1 in. on the vertical axis may always represent $1000. We could, how-

[1] From J. R. RIGGLEMAN and I. N. FRISBEE, "Business Statistics," pp. 93–95, McGraw-Hill Book Company, Inc., New York, 1932.

over, produce scales on which equal distances did not always represent equal absolute magnitudes. The most common (and the most useful) scale of this kind is that in which we let given distances represent the logarithms of numbers rather than the numbers themselves. Figure 16.1 shows a chart grid in which the vertical lines are equidistant, as usual. But it is immediately noticeable that the horizontal lines are unequally spaced. Here the vertical distances are proportional to the logarithms of the numbers represented; that is, the distance from the figure 1 to

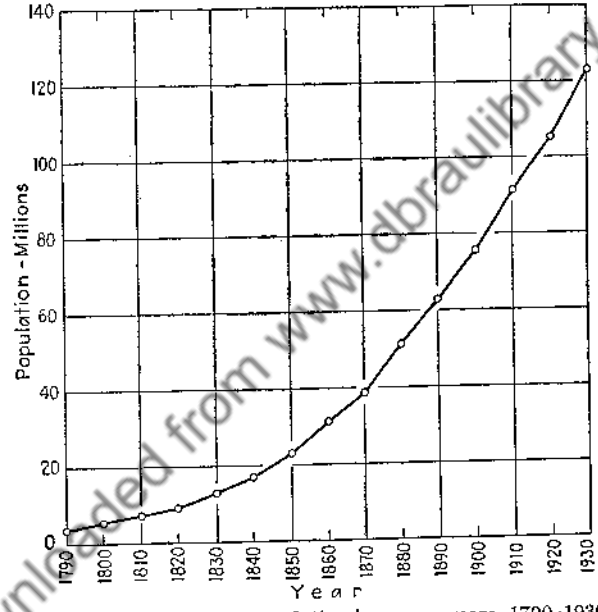


FIG. 16.2.—United States population in census years, 1790–1930.

the figure 2 is greater than the distance from the figure 8 to the figure 9. While the absolute difference between 1 and 2 is the same as that between 8 and 9, inspection of a table of logarithms will show that there is much more difference between the logarithm of one (.00000) and the logarithm of two (.30103) than there is between the logarithm of eight (.90309) and the logarithm of nine (.95424).

This type of graph paper (known as semilogarithmic paper) is useful in that all lines with the same slope show the same *percentage* rate of change on the vertical scale. Suppose, for exam-

ple, that we plot the data of Table 16.1 on ordinary graph paper. The result is shown in Fig. 16.2. It will be noted that the line showing the growth of population becomes steeper and steeper as we move toward the right. This happens because the absolute increase in population from one census to the next has become larger and larger as time has gone on. But an increase of 2 million people is large if the population at the beginning of the period is but 4 million. Such an increase would mean that we had added 50 per cent to the population. If our original population were

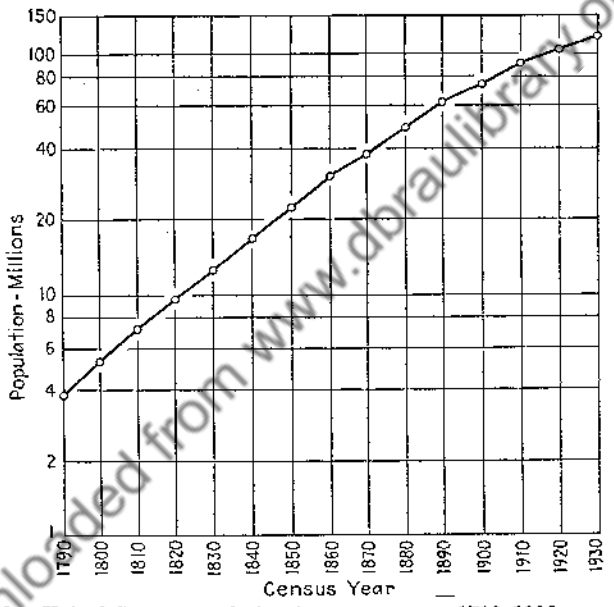


FIG. 16.3.—United States population in census years, 1790–1930, on semilogarithmic scales.

100 million, however, an increase of 2 million would be but a 2 per cent increase. If we wish to note the facts with regard to percentage rates of change in the population, we should chart the figures of Table 16.1 on semilogarithmic paper, as this has been done in Fig. 16.3. Here we note that while the line moves farther and farther from the base line (thus showing that the population has increased in absolute amount), nevertheless the slope of the line becomes more and more gentle, indicating that the *rate* of population increase has become less and less. The proportionate increase in population each decade is smaller than

it used to be. When as here, we have data plotted on semilogarithmic paper, equal slopes represent equal percentage rates of change, and the steeper the slope, the greater is the percentage rate of change. This fact makes it easy to compare the relative changes in various phenomena even though they may not be expressed in comparable units. When the diverse data have been plotted on semilogarithmic paper, we can compare the slopes of the lines and thus consider each variable on a relative basis.

In Chap. XIV we discussed various types of curves and discovered that in some cases two variables are so related that, although one does not vary directly as the other, the one does vary as the logarithm of the other. Such curves we called "logarithmic curves," and such relationships were depicted in Fig. 14.7, page 440. Whenever two variables are so related that they can be described by a logarithmic curve, we can discover the relationship easily by using semilogarithmic paper. When a logarithmic curve is plotted on semilogarithmic paper, it becomes a straight line.

Let us take, for example, the hypothetical case of Table 16.3, in which the figures are supposed to represent the production of some commodity and the price of that commodity. If we plot

TABLE 16.3.—PRICE AND PRODUCTION OF COMMODITY *X*

Year	Price	Production
1928	35¢	82
1929	25	98
1930	65	52
1931	20	108
1932	16	116
1933	45	72
1934	58	60
1935	29	90

these data on arithmetic paper, we get Fig. 16.4. Comparison of this with Fig. 14.7, page 440, shows that these data fall along a curve somewhat like part of the logarithmic curves there portrayed. If we wish to test the data further in order to determine whether or not the relationship can be described by a logarithmic curve, we can plot the data from Table 16.3 on semilogarithmic

paper. This operation gives us the scatter diagram shown in Fig. 16.5. Here we see that the points fall along a straight line. If the data, when plotted on semilogarithmic paper, fall along a straight line or in a straight band, we know that a logarithmic

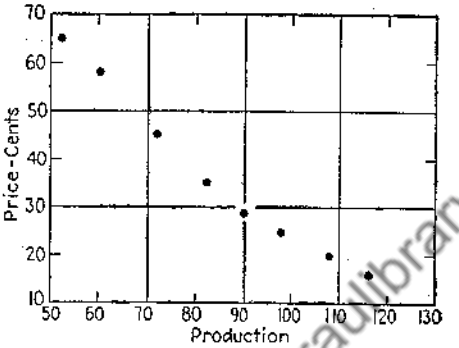


FIG. 16.4.—Relationship between the production and price of commodity *X*. Hypothetical case from page 493.

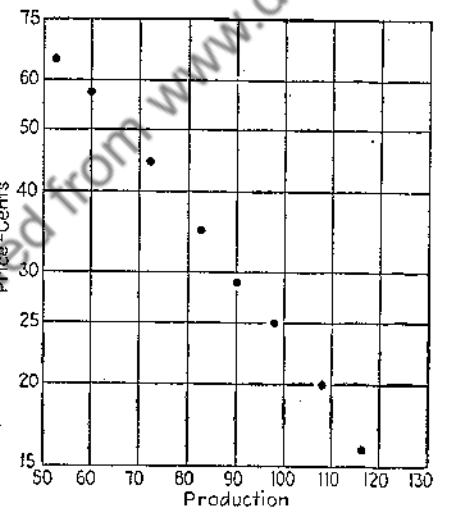


FIG. 16.5.—Relationship between the production and price of commodity *X*. Same data as those used in Fig. 16.4, but on semilogarithmic scales.

curve will describe the data, and we can then proceed to fit a logarithmic curve by the methods heretofore discussed (see pages 447*ff*.).

Semilogarithmic paper is the most useful of the cross-section papers that do not have the common arithmetic scales. There

are, however, many other such papers. For example, we may make both the vertical and the horizontal scales logarithmic, rather than merely the vertical scale. In this case we have what is called *logarithmic* paper as distinguished from the semilogarithmic paper which we have just described. Logarithmic scales are also called *ratio scales*, and we can differentiate the two types of paper just mentioned by saying that on semilogarithmic paper one of the scales is the usual arithmetic scale and the other is a ratio scale; while logarithmic paper has ratio scales both vertically and horizontally. On logarithmic paper we get a straight line if a given percentage change in one variable tends always to be accompanied by some given percentage change in the other variable. For example, if every 2 per cent change in X is associated with a 7 per cent change in Y, we shall get a straight line on logarithmic paper but a curve on arithmetic paper. Under such circumstances the log of Y varies as the log of X. This paper is particularly useful in some applied fields, as in the study of demand in the field of economics. A demand curve with constant elasticity yields a straight line on logarithmic paper.

Statisticians have also developed many other scales which can be used in analyzing diverse relationships. For example, if one plots data on *reciprocal* paper and finds a straight line, he knows that the relationship can be described by a reciprocal curve such as that used on page 452 to describe the relationship between potato production and price. If one is to do any extended amount of work in the field of curve fitting, it will pay him to acquaint himself thoroughly with the various graphic scales which are available.[1] By their use a considerable amount of time can be saved.

16.7. Types of Statistical Diagrams.—The student who wishes an extended discussion of diagrams must study one or more of the books which are entirely devoted to the subject. The limitations of space make it impossible to go into detail here. To be sure, the major considerations which should be kept in mind in the making of statistical diagrams have been covered in the admirable report on standards for graphic presentation which is quoted earlier in this chapter (see pages 487*ff.*). But the statistician

[1] For a description of such scales and their uses, see any good book on graphics, such as Karl Karsten, "Charts and Graphs," Prentice-Hall, Inc., New York, 1923.

who wishes suggestions as to other types of diagrams will profit most by studying the specialized literature in the field. We shall content ourselves here with but a few observations on the subject in addition to those which have already been made.

Diagrams may take the form of lines, bars, areas such as squares and circles, solids, or pictures of solids, etc. It is usually wise to rely, where possible, on diagrams in which comparisons are made in but one dimension. Thus, if we can depict our data by a comparison of the lengths of bars, we are likely to get a more rapid and a more accurate impression than if we attempted to show the same data by a comparison of the areas of circles or by pictures purporting to show three dimensions. The simpler the diagram, the better. Experiments seem to show[1] that comparisons of areas lead to much more error on the part of witnesses than do comparisons in one dimension. It does seem to be true, however, that people can make much more accurate comparisons of angles at the center of a circle than would be expected a priori.[2]

Rule 4 of the Joint Committee on Standards for Graphic Presentation suggests that when a diagram is so constructed that the zero line of the vertical scale would not be likely to appear, the omission of the zero line should be shown by a horizontal break in the diagram (see page 487). If no such break is made, many people will misinterpret the graph. Fluctuations which are really very small in relative size will appear large because of the omission of part of the graph. It might almost be said that in most cases these diagrams, even when properly made, are practically useless. Graphic presentation is intended to give a correct and quick impression; yet if the break in the diagram is omitted, the impression conveyed is likely to be false. But if the break is put into the diagram in order to guard against misinterpretation, it becomes necessary for the reader to inspect the scales carefully to see how much has been omitted. He then has to make mental calculations to determine whether or not the variations shown are important, and this takes so long that much of the advantage of the graph is lost. It is usually true that, if

[1] CROXTON and STEIN, Graphic Comparison by Bars, Squares, Circles, and Cubes, *Journal of the American Statistical Association*, Vol. 27, March, 1932, pp. 54–60.

[2] See CROXTON and STRYKER, Bar Charts versus Circle Diagrams, *Journal of the American Statistical Association*, Vol. 22, December, 1927, pp. 473–482.

variations are so small that they will not show when the entire vertical scale is given, this is in itself an indication of the fact that the variations are relatively too small to worry about. Although every worker will be able to suggest exceptions to this general rule, it is nevertheless true in a large enough proportion of the cases so that a statistician who plans to use a broken diagram should feel that the burden of proof is on him. When the graph is one in which the comparison with zero is not important (as when we are comparing index numbers with 100, or when we are comparing one variable with another on a scatter diagram), this last observation obviously does not apply.

EXERCISES

1. Make a sheet of semilog paper. Note that the log of 2 is .30103, the log of 3 is .47712, the log of 10 is 1.00000, and so on. Thus, if we let the distance from 1 to 10 on the vertical axis be made up of 1000 units, the distance from 1 to 2 will contain 301 units, the distance from 1 to 3 will contain 477 units, and so forth. Progressing in this way, lay off the entire vertical scale from 1 to 100, using a decimally divided scale such as a meter stick or an engineers' triangular scale.

2. Make a sheet of reciprocal paper and plot on it the data of Table 14.1, page 448. Do the data fall in a straight band? If so, what of it?

3. If you were to make a table showing the population of the states in the United States, in what ways might you arrange the states in the stub of the table? Using what circumstances would you use each arrangement?

4. In Table 16.2, page 484, it is much easier to compare the maximum distance from the sun with the minimum distance than it is to compare the distance from the sun with the distance from the earth. This is so because these figures are in adjacent columns. Rearrange the table in such a way that it includes all the data given on page 484, but so that the easy comparison is between the distance to the earth and the distance to the sun. The two maxima should be compared together and the two minima should be compared together.

5. What general rules relative to the arrangement of columns in tables can you derive from Exercise 4 above?

6. Draw up the box headings for a table which is to show for the various states the number of each sex of the white race and of the black race who are native-born and who are foreign-born. Make the comparison between the sexes the easiest comparison. The names of the states themselves are to be in the stub.

CHAPTER XVII

COLLECTION AND ANALYSIS OF DATA

17.1. Definition of the Problem.—The first step in any statistical investigation is that of defining the problem to be studied. The statistician must delimit his field with care, deciding in advance exactly what it is that he wishes to discover. For example, do we wish to discover the relationship between potato production and the price of potatoes? Do we mean world production, United States production, Maine production, or production in Aroostook County? Do we really want total production or do we want commercial production (potatoes produced for sale)? Do we want figures on all potatoes, or do we want to limit our study to the early crop or the late crop? Do we wish to single out particular varieties, such as Cobblers or Green Mountains? Do we wish to make any allowance for imports or exports? And similarly with regard to the prices. Do we want the farm price or the price at the city? If the latter, which city? The price at what time of year? The price for what grade of potatoes? And is it to be the wholesale or the retail price?

We see that the problem is not at all clear if we state merely that we are to study the relationship of potato production to prices. We must amplify and clarify and specify until we can give a more detailed and accurate picture of the problem to be dealt with. We may conclude that we wish to find the relationship between the total commercial potato crop of the late-crop states (which would, of course, be listed) and the wholesale price of U.S. No. 1 Cobblers on Oct. 15 on the Boston market. It is very likely that analysis of the problem would indicate some further qualifications, but enough has been said to indicate the importance and nature of the definition of a problem before statistical work is commenced. Some preliminary statistical work may be necessary even to determine those definitions which will prove most useful.

17.2. Derived Data.—Having settled upon the problem to be studied, the statistician must next acquaint himself with other work which has been done in the same general field. If we continue to use the potato example which we mentioned in the preceding paragraph, this would mean that the statistician would find what he could about other price studies in general and about studies of potato prices in particular. He would try to discover what data were available that he could utilize in his own study, and also what methods other students had found helpful in similar analyses. In some cases, perhaps, he would find in one source or another all the numerical data he needed as a basis for his own study. Then it would not be necessary for him to collect any original data at all. He would, of course, wish to study the *derived data* (data from other studies) carefully to be sure that he understood exactly what it represented, and he would naturally be careful when using it to credit the source from which it came. The giving of credit to the sources of derived data is important not only in order that the user may escape the charge of plagiarism, but also so that other workers running across the new study may trace the data back to the original source to verify them and study their characteristics.

17.3. Sources of Derived Data.—It is impossible to give here a list of the sources of derived statistical data, since they are too numerous. The worker in any particular field soon learns where he should look for published data in that field. The United States Government publishes much valuable material of a statistical nature. The U.S. Census, the *Statistical Abstract*, and the publications of the Departments of Agriculture, Labor, and Commerce are especially helpful; the Treasury Department, the Department of the Interior, the Interstate Commerce Commission, the Shipping Board, and other governmental agencies also furnish current statistical material in their respective fields. In addition many agencies of state and local governments issue statistical material. Derived data may be found also in the technical journals of various fields, or may be published by trade associations or business houses.

The statistician who wishes to use derived material must make certain, of course, that the source is trustworthy, and he will usually wish to find how the data were collected and how they were analyzed. He will usually make use of derived data where

he can, of course, in order to save time and money which would be expended in gathering the data *de novo*.

17.4. Collecting Original Data.—Sometimes, of course, the statistician will discover that he needs data which have not been gathered. Under such circumstances it will be necessary for him to collect the data himself.

When we are to collect statistical data, the first problem is that of deciding exactly what data we need. If we collect more data than are necessary we usually add to the expense and time of the undertaking. If we collect too little we have to go back for more, which is even more costly. Before the collection of original data begins it is important that the problem be so carefully defined, and the methods to be used so completely visualized, that it is possible to determine just what additional data are necessary.

This does not seem like a very difficult problem. As a matter of fact it always turns out to be very troublesome, especially because before the data are gathered it is necessary to define the statistical units accurately and to determine the degree of accuracy which will be necessary. When the census takers set out to discover the number of farms in Texas, they must first define a farm. Is any patch of land on which crops are raised a farm? How will they classify the backyard vegetable garden of the suburbanite or the commercial greenhouse in the city? If a man owns two pieces of land, working one himself and letting the other to a tenant, has he one farm or two? Such questions must be foreseen and handled in advance. Suppose we wish to determine the number of cows on these farms. What *is* a cow? Some farmers would, of course, give the total number of cattle without distinction as to age or sex. If we wish to confine ourselves to females, we must make the point clear. And when does a calf stop being a a calf and become a heifer? When does the heifer change to a cow? These questions must be answered before the data are collected. One who patronizes city meat markets nowadays can well ask when lamb stops being lamb and becomes mutton.

Likewise, if we are getting figures on factory employment, how do we decide what is a factory? Do we want to include all employees—even those in clerical or executive positions? Do we want all employees or only permanent employees?

Should we count superannuated employees who are still on the pension roll but who are not actively at work?

It should now be clear that the statistical units must be carefully defined before the data are collected if the answers given by various informants are to be comparable and if the analyst is to know what his data mean when he studies them. There are cases in which units cannot be so defined as to make it possible to collect accurate data. Mayo-Smith points out[1] that an attempt was made in the U.S. census of 1890 to discover the number of persons of pure and mixed African blood by distinguishing between blacks, mulattoes, quadroons, and octoroons, but since there were no physical characteristics which could be used as a sure check, and since many of the persons in question were ignorant of their genealogy, the attempt was a failure.

As has been noted, it is necessary also at the beginning to determine the degree of accuracy on which we shall insist for our data. If we ask for time spent in harvesting wheat, do we want it to the nearest number of days or the nearest number of hours, or is stop-watch accuracy necessary? If we are collecting data on farm receipts, are we content to take rough estimates made by the farmer or must we gather information only from those farmers who keep farm accounts and who can give us data accurate to the cent? We have noted in an early chapter of this book (see pages 7*ff*.) that statistical data are seldom entirely and exactly accurate, and that no matter how we plan our study we cannot expect to reach the ultimate in accuracy. But if some particular degree of accuracy is necessary for our work, we should decide at the start what accuracy we need and how to get it.

17.5. Methods of Collecting Statistical Data.—When we have decided what additional original data we need and how accurate they must be, and when we have defined the units that are to be used in describing them, we must still decide how to go about the process of collecting them. We may decide to collect the data by *personal inquiry;* that is, the statistician may collect the data himself. This method has the great advantage that the man who is analyzing the data has collected them, and therefore under-

[1] In PALGRAVE, article on Statistics, "Dictionary of Political Economy," Vol. III, p. 468, 1926.

stands their limitations and peculiarities. In any large investigation, however, personal collection by one man may be out of the question on account of the time and expense which would be involved. Therefore it is common to use *enumerators* or *special agents* to collect the data. If enumerators are to be sent into the field to interview informants and to obtain statistical data from them, it is very important that the enumerators be carefully chosen and painstakingly instructed. They must understand the purpose for which the data are to be used, they must familiarize themselves with the definitions of statistical units which we have just mentioned, they must understand the necessity of reasonable accuracy, they must be tactful and resourceful. A lazy or dishonest enumerator may spoil an entire statistical study.

If one is to collect a large amount of data, he often finds it cheaper to send schedules directly to the informants to be filled out and returned by them rather than to send enumerators with the schedules. This method of using *informant's own schedules* has the advantage of saving time and money, but has the disadvantages that one usually gets far fewer returns, that different informants interpret even the simplest questions in different ways, and that the informant cannot easily ask questions about points which puzzle him.

17.6. The Schedule.—Whether the statistician collects his data in person or through enumerators or informant's schedules, he must make out some list of questions to be asked. This list of questions is called a *questionnaire* or *schedule*. We have already noted that the schedule must be complete, so that it will elicit information on all points which we need to cover without forcing people to give unnecessary information. It is also important, of course, that the questions be simple and not ambiguous. The definitions of statistical units should be clearly stated and explained, and the degree of accuracy expected in the answer should be made explicit. Statisticians find that they get more uniformly satisfactory results if they so frame their questions that they can be answered by "No," "Yes," or a number. Also, the questions should be tactfully phrased, and if there are any questions on subjects which may prove embarrassing to the informant it is usually wise to put them at the end of the schedule. There are certain topics on which informants seem to be prone to

give misinformation. Mayo-Smith illustrates this point as follows:[1]

> Classification of the population according to wealth or income fails because direct inquiries will not be answered truthfully, and outward marks fail. Even such a question as whether one is employer, employee, or working on one's own account was imperfectly answered in the English census of 1891, either because it was misunderstood or through perverseness. It is said that questions in regard to religious denominations excite suspicion in France; and questions in regard to mental or physical infirmity of members of the family excited considerable feeling in the United States in 1890.

It is sometimes said that women tend to understate their ages, and it is probable that an investigator would be given faulty information relative to past criminal records if he were to rely on the statements of the people suspected. Only by using the greatest tact can one hope to reduce such errors; and there is no sure way of completely eliminating them.

17.7. Choosing the Sample.—When the statistician has decided what questions to ask, he must still decide whom to question; that is, he must decide how many cases to study and how to select them. We have seen in earlier chapters that statistical conclusions can be applied only within the universe from which the sample was drawn. If we are investigating the rate of wages in a given city, and if we decide that wage data can be obtained most easily from labor unions, we must realize that our data hold true only for union labor. If there is any considerable number of nonunion workers in the city, our conclusions will not necessarily apply to them. One can never be entirely certain that he has a strictly random sample, and as a result he can never be entirely certain within what universe the conclusions are applicable. The careful statistician does try, however, to select his data in such a way that there will be no hidden causes influencing the nature of his results. Here again, as in so many other cases, it becomes evident that a statistician must be an expert within the field which he is studying as well as in the field of statistical method. Unless he knows his own field he has no way of knowing what factors are likely to influence the nature of his results.

[1] PALGRAVE, *op. cit.*, p. 468.

We have discovered also that the reliability of statistical conclusions depends on the number of cases studied. The statistician must decide in advance what accuracy is necessary in his conclusions, so that he may know approximately how many cases to study.

17.8. Aids in the Analysis of Statistical Data.—After the original data have been gathered, the statistician finds himself confronted with a vast mass of unrelated data scattered through many schedules. It is his duty to bring order out of the confusion. Sometimes he makes a start at this by transferring the data from questionnaires to large tables, the table having a column in which can be entered the answer to each question. These primary tables are not for publication, and are far more complicated than tables would ever be that were intended for presentation to the public. Or the statistician may copy the pertinent figures from each schedule onto a card, such as a 3- by 5-in. index card. This method makes it easy to sort and subsort the data. Thus if we are studying farm incomes on dairy farms, and if we make out a card for each farm, it is easy to sort the cards into piles which differ according to the number of cows on the farm, and to subsort each pile according to the breed of cattle and other specifications. In this way farms with particular characteristics can be easily grouped together for study.

This last procedure is made much easier if we can manage to do the sorting and subsorting by machine. This can be arranged if we punch holes in the cards to represent the numbers involved, and use one of the automatic card-sorters which sorts the cards by making electrical contacts through the holes. Such a method can be used economically only when there is a large number of schedules to be handled, but under such circumstances it makes for considerable saving in time and expense. Machines for punching the original cards, for sorting them, and for tabulating the results can be rented from the companies which manufacture them. The statistician who plans to use them will be wise to consult with the manufacturers when planning his study, so that the data may be gathered in a form which facilitates the use of the punched cards.

It goes without saying that the statistician uses adding and computing machinery wherever possible, and that to save time he relies when he can on tables of figures and the slide rule.

Tables giving the squares and cubes, square roots, and cube roots of numbers are common. Some tables give also the sums of the squares of the first N numbers or the sums of the squares of the first N odd numbers. We have observed the usefulness of tables showing the areas and ordinates of the normal curve. The time involved in statistical analysis can be reduced tremendously by judicious use of such aids.

Some workers avoid using the slide rule because of its inaccuracy. We have seen (see Chap. II), that no measurements are ever perfectly accurate except by chance. Accuracy beyond three or four significant figures is almost unheard of in data gathered in the fields of social or biological science. The ordinary 10-in. slide rule gives three- to four-place accuracy, and the 20-in. rule gives accuracy to at least one more place. Thus we see that the slide rule is usually accurate enough for the work we must do with it. It is at any rate a rapid and useful check on the accuracy of computations which have been carried out by longhand.

Many kinds of drafting equipment are useful in the graphic presentation of statistical results. Various kinds of drawing pens and lettering guides are especially useful. For complete descriptions of such instruments, the student is referred to the catalogues of companies manufacturing them. The statistician who has proper equipment with which to work finds that he can eliminate much of the routine and tedium of arithmetical manipulation, and that he can give more of his time to the more interesting and valuable work of planning statistical operations and interpreting statistical conclusions.

APPENDICES

APPENDIX I

Areas under the Normal Curve

Fractional parts of the total area (1.000) under the normal curve between the mean and a perpendicular erected at various numbers of standard deviations (x/σ) from the mean.[1] To illustrate the use of the table, 39.065 per cent of the total area under the curve will lie between the mean and a perpendicular erected at a distance of 1.23σ from the mean.

Each figure in the body of the table is preceded by a decimal point.

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	00000	00399	00798	01197	01595	01994	02392	02790	03188	03586
0.1	03983	04380	04776	05172	05567	05962	06356	06749	07142	07535
0.2	07926	08317	08706	09095	09483	09871	10257	10642	11026	11409
0.3	11791	12172	12552	12930	13307	13683	14058	14431	14803	15173
0.4	15554	15910	16276	16640	17003	17364	17724	18082	18439	18793
0.5	19146	19497	19847	20194	20450	20884	21226	21566	21904	22240
0.6	22575	22907	23237	23565	23891	24215	24537	24857	25175	25490
0.7	25804	26115	26424	26730	27035	27337	27637	27935	28230	28524
0.8	28814	29103	29389	29673	29955	30234	30511	30785	31057	31327
0.9	31594	31859	32121	32381	32639	32894	33147	33398	33646	33891
1.0	34134	34375	34614	34850	35083	35313	35543	35769	35993	36214
1.1	36433	36650	36864	37076	37286	37493	37698	37900	38100	38298
1.2	38493	38686	38877	39065	39251	39435	39617	39796	39973	40147
1.3	40320	40490	40658	40824	40988	41149	41308	41466	41621	41774
1.4	41924	42073	42220	42364	42507	42647	42786	42922	43056	43189
1.5	43319	43448	43574	43699	43822	43943	44062	44179	44295	44408
1.6	44520	44630	44738	44845	44950	45053	45154	45254	45352	45449
1.7	45543	45637	45728	45818	45907	45994	46080	46164	46246	46327
1.8	46407	46485	46562	46638	46712	46784	46856	46926	46995	47062
1.9	47128	47193	47257	47320	47381	47441	47500	47558	47615	47670
2.0	47725	47778	47831	47882	47932	47982	48030	48077	48124	48169
2.1	48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2.2	48610	48645	48679	48713	48745	48778	48809	48840	48870	48899
2.3	48928	48956	48983	49010	49036	49061	49086	49111	49134	49158
2.4	49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2.5	49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2.6	49534	49547	49560	49573	49585	49598	49609	49621	49632	49643
2.7	49653	49664	49674	49683	49693	49702	49711	49720	49728	49736
2.8	49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2.9	49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3.0	49865									
3.5	4997674									
4.0	4999683									
4.5	4999966									
5.0	4999997133									

[1] This table has been adapted, by permission, from F. C. Kent, "Elements of Statistics," McGraw-Hill Book Company, Inc., 1924.

APPENDIX II

Ordinates of the Normal Curve

Ordinates (heights) of the unit normal curve.[1] The height (y) at any number of standard deviations (x) from the mean is

$$y = 0.3989e^{-x^2/2}$$

To obtain answers in units of particular problems, multiply these ordinates by $N(Ci)/\sigma$ where N is the number of cases, Ci the class interval, and σ the standard deviation.

Each figure in the body of the table is preceded by a decimal point.

x/σ	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	39894	39892	39886	39876	39862	39844	39822	39797	39767	39733
0.1	39695	39654	39608	39559	39505	39448	39387	39322	39253	39181
0.2	39104	39024	38940	38853	38762	38667	38568	38466	38361	38251
0.3	38139	38023	37903	37780	37654	37524	27391	37255	37115	36973
0.4	36827	36678	36526	36371	36213	36053	35889	35723	35553	35381
0.5	35207	35029	34849	34667	34482	34294	34105	33912	33718	33521
0.6	33322	33121	32918	32713	32506	32297	32086	31874	31659	31443
0.7	31225	31006	30785	30563	30339	30114	29887	29658	29430	29200
0.8	28969	28737	28504	28269	28034	27798	27562	27324	27086	26848
0.9	26609	26369	26129	25888	25647	25406	25164	24923	24681	24439
1.0	24197	23955	23713	23471	23230	22988	22747	22506	22265	22025
1.1	21785	21546	21307	21069	20831	20594	20357	20121	19886	19652
1.2	19419	19186	18954	18724	18494	18265	18037	17810	17585	17360
1.3	17137	16915	16694	16474	16256	16038	15822	15608	15395	15183
1.4	14973	14764	14556	14350	14146	13943	13742	13542	13344	13147
1.5	12952	12758	12566	12376	12188	12001	11816	11632	11450	11270
1.6	11092	10915	10741	10567	10396	10226	10059	09893	09728	09566
1.7	09405	09246	09089	08933	08780	08628	08478	08329	08183	08038
1.8	07895	07754	07614	07477	07341	07206	07074	06943	06814	06687
1.9	06562	06438	06316	06195	06077	05959	05844	05730	05618	05508
2.0	05399	05292	05186	05082	04980	04879	04780	04682	04586	04491
2.1	04398	04307	04217	04128	04041	03955	03871	03788	03706	03626
2.2	03547	03470	03394	03319	03246	03174	03103	03034	02965	02898
2.3	02833	02768	02705	02643	02582	02522	02463	02406	02349	02294
2.4	02239	02186	02134	02083	02033	01984	01936	01888	01842	01797
2.5	01753	01709	01667	01625	01585	01545	01506	01468	01431	01394
2.6	01358	01323	01289	01256	01223	01191	01160	01130	01100	01071
2.7	01042	01014	00987	00961	00935	00909	00885	00861	00837	00814
2.8	00792	00770	00748	00727	00707	00687	00668	00649	00631	00613
2.9	00595	00578	00562	00545	00530	00514	00499	00485	00470	00457
3.0	00443									
3.5	0008727									
4.0	0001338									
4.5	0000160									
5.0	000001487									

[1] This table adapted, by permission, from Kent, "Elements of Statistics."

APPENDIX III

Chances of Differing from the Mean by Given Numbers of Standard Deviations

Chances that an item chosen at random from a normal distribution will lie as far from the mean as the number of standard deviations stated.[1]

Example: The chances are 0.0357 (that is, 357 chances out of 10,000) that an item chosen at random from a normal distribution will differ from the mean by as much as 2.1 standard deviations.

x/σ	Chances	x/σ	Chances
0.0	1.000	2.0	0.0454
0.1	0.920	2.1	0.0357
0.2	0.841	2.2	0.0277
0.3	0.764	2.3	0.0214
0.4	0.689	2.4	0.0163
0.5	0.617	2.5	0.0124
0.6	0.549	2.6	0.00932
0.7	0.484	2.7	0.00694
0.8	0.424	2.8	0.00512
0.9	0.368	2.9	0.00374
1.0	0.317	3.0	0.00270
1.1	0.271	3.5	0.000465
1.2	0.230	4.0	0.0000634
1.3	0.193	4.5	0.0000068
1.4	0.162	5.0	0.000000573
1.5	0.134		
1.6	0.109		
1.7	0.0891		
1.8	0.0719		
1.9	0.0574		

[1] The data in this table were computed from information in J. W. Glover's "Tables of Probability and Statistical Functions," Geo. Wahr, Ann Arbor, Michigan, 1930, except for the last four entries, which are based on H. O. Rugg, "Statistical Methods Applied to Education," Houghton Mifflin Company, Boston, 1917.

APPENDIX IV

Chances of Differing from the Mean in a Given Direction by More than Given Numbers of Standard Deviations

Chances that an item chosen at random from a normal distribution will be on the same side of the mean as the item chosen, and distant from the mean by as much as the stated number of standard deviations.[1] *Example:* The chances are 0.0179 (that is, 179 out of 10,000) that an item chosen at random from a normal distribution will be above the mean and removed from it by as much as 2.1 standard deviations. The chances that it will be this far below the mean are also 0.0179.

x/σ	Chances	x/σ	Chances
0.0	0.500	2.0	0.0227
0.1	0.460	2.1	0.0179
0.2	0.421	2.2	0.0139
0.3	0.382	2.3	0.0107
0.4	0.345	2.4	0.00820
0.5	0.309	2.5	0.00621
0.6	0.274	2.6	0.00547
0.7	0.242	2.7	0.00347
0.8	0.212	2.8	0.00256
0.9	0.184	2.9	0.00187
1.0	0.159	3.0	0.00135
1.1	0.136	3.5	0.000233
1.2	0.114	4.0	0.0000317
1.3	0.0968	4.5	0.0000034
1.4	0.0807	5.0	0.000000287
1.5	0.0668		
1.6	0.0548		
1.7	0.0446		
1.8	0.0359		
1.9	0.0287		

[1] Data from same sources as Appendix III.

APPENDIX V

VALUES OF $\sqrt{pq}$ WHEN $p + q = 1$

Each entry in the body of this table is preceded by a decimal point.

p or q	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	000	100	140	171	196	218	238	255	271	286
0.10	300	313	325	336	347	357	367	376	384	392
0.20	400	407	414	421	427	433	439	444	449	454
0.30	458	463	467	470	474	477	480	483	485	488
0.40	489	492	494	495	496	498	498	499	500	500
0.50	500	500	500	499	498	498	496	495	494	492
0.60	489	488	485	483	480	477	474	470	467	463
0.70	458	454	449	444	439	433	427	421	414	407
0.80	400	392	384	376	367	357	347	336	325	313
0.90	300	286	271	255	238	218	196	171	140	100

The table is to be used in connection with:

1. The determination of the standard deviation of the number of successes, as on page 88. The formula becomes

$$\sqrt{npq} = \sqrt{n}(\sqrt{pq})$$

The value of the latter is given in the table.

2. The determination of the standard error of a relative frequency, as on page 145. The formula becomes

$$\sqrt{\frac{pq}{N}} = \frac{\sqrt{pq}}{\sqrt{N}}$$

The numerator is given in the table.

Example: On page 247 it is necessary to evaluate the term $\sqrt{pq/n}$ when $p = 0.25$ and $n = 53{,}280$. From the table above $\sqrt{pq} = 0.433$ when $p = 0.25$, and $\sqrt{53{,}290} = 231$. Thus we have $0.433/231 = 0.00187$.

APPENDIX VI

VALUES OF r FOR VARIOUS VALUES OF z FROM 1 TO 3[1]

Each figure in the body of the table is preceded by a decimal point.
Example: When z has a value of 1.23, r has a value of 0.8426

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0000	0100	0200	0300	0400	0500	0599	0699	0798	0898
0.1	0997	1096	1194	1293	1391	1489	1586	1684	1781	1877
0.2	1974	2070	2165	2260	2355	2449	2543	2636	2729	2821
0.3	2913	3004	3095	3185	3275	3364	3452	3540	3627	3714
0.4	3800	3885	3969	4053	4136	4219	4301	4382	4462	4542
0.5	4621	4699	4777	4854	4930	5005	5080	5154	5227	5299
0.6	5370	5441	5511	5580	5649	5717	5784	5850	5915	5980
0.7	6044	6107	6169	6231	6291	6351	6411	6469	6528	6584
0.8	6640	6696	6751	6805	6858	6911	6963	7014	7064	7114
0.9	7163	7211	7259	7306	7352	7398	7443	7487	7531	7574
1.0	7616	7658	7699	7739	7779	7818	7857	7895	7932	7969
1.1	8005	8041	8076	8110	8144	8178	8210	8243	8275	8306
1.2	8337	8367	8397	8426	8455	8483	8511	8538	8565	8591
1.3	8617	8643	8668	8692	8717	8741	8764	8787	8810	8832
1.4	8854	8875	8896	8917	8937	8957	8977	8996	9015	9033
1.5	9051	9069	9087	9104	9121	9138	9154	9170	9186	9201
1.6	9217	9232	9246	9261	9275	9289	9302	9316	9329	9341
1.7	9354	9366	9379	9391	9402	9414	9425	9436	9447	9458
1.8	94681	94783	94884	94983	95080	95175	95268	95359	95449	95537
1.9	95624	95709	95792	95873	95953	96032	96109	96185	96259	96331
2.0	96404	96473	96541	96609	96675	96739	96803	96865	96926	96986
2.1	97045	97103	97159	97215	97269	97323	97375	97426	97477	97526
2.2	97574	97622	97668	97714	97759	97803	97846	97888	97929	97970
2.3	98010	98049	98087	98124	98161	98197	98233	98267	98301	98335
2.4	98367	98399	98431	98462	98492	98522	98551	98579	98607	98635
2.5	98661	98688	98714	98739	98764	98788	98812	98835	98858	98881
2.6	98903	98924	98945	98966	98987	99007	99026	99045	99064	99083
2.7	99101	99118	99136	99153	99170	99186	99202	99218	99233	99248
2.8	99263	99278	99292	99306	99320	99333	99346	99359	99372	99384
2.9	99396	99408	99420	99431	99443	99454	99464	99475	99485	99495
3.0	99505									
3.5	998178									
4.0	999329									
4.5	999753									
5.0	999909									
5.5	9999666									
6.0	9999877									
6.5	99999548									

For greater accuracy, or for values beyond the table:

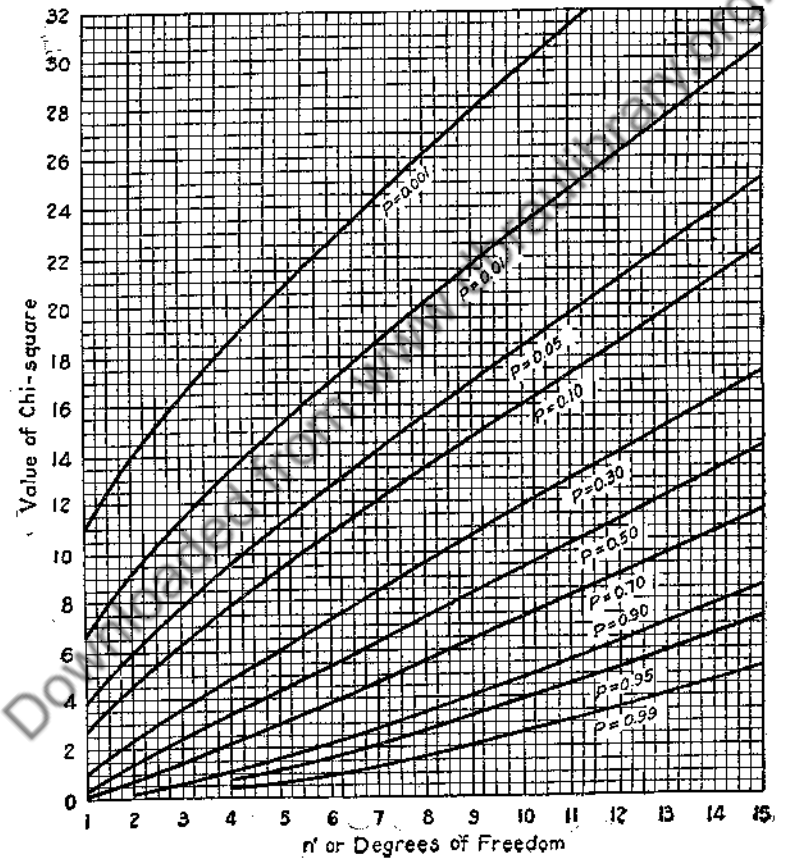
$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \qquad z = 1.15129254 \log_{10} \frac{1 + r}{1 - r}$$

$$\sigma_z = \frac{1}{\sqrt{n - 3}}$$

[1] Taken by permission, with minor additions, from R. A. Fisher, "Statistical Methods for Research Workers," Oliver & Boyd, Edinburgh, 1938.

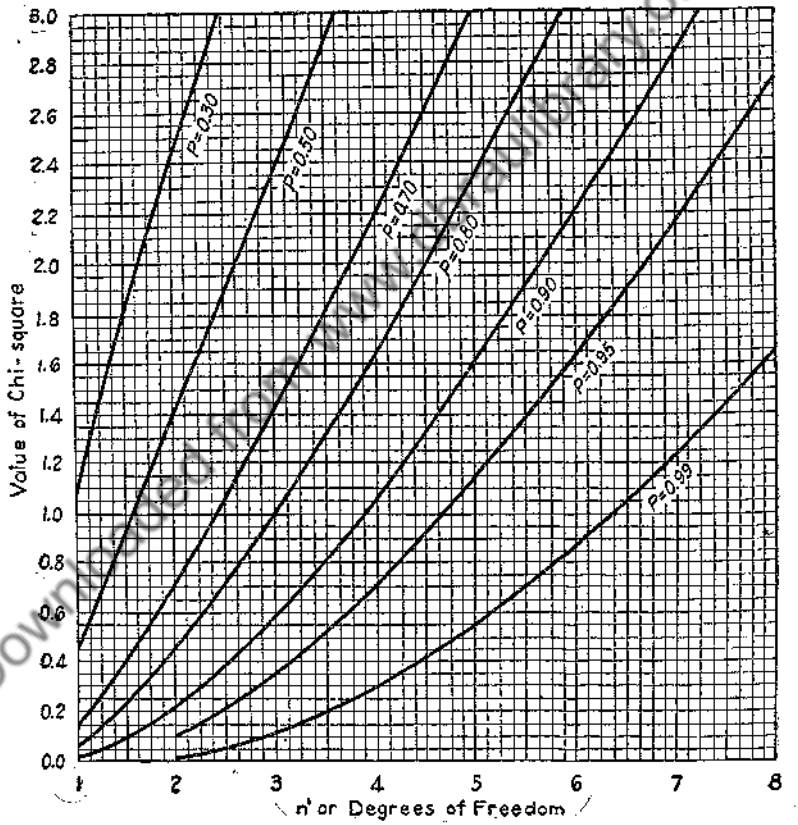
APPENDIX VII

VALUES OF P FOR VALUES OF CHI SQUARE BETWEEN 0 AND 32 AND VALUES OF n' BETWEEN 1 AND 15



APPENDIX VIII

VALUES OF P FOR VALUES OF CHI SQUARE BETWEEN 0 AND 3 AND VALUES OF n' BETWEEN 1 AND 8



APPENDIX IX

SELECTED BOOKS ON STATISTICAL METHOD

The student will find the following books on statistical method helpful. The list is not complete, but it includes the texts which the author has found most helpful with beginning students.

BARLOW: "Tables of Squares, Cubes, Square Roots, Cube Roots, and Reciprocals," Spon and Chamberlain, New York, 1919. (Covers all numbers from 1 to 10,000.)

A. L. BOWLEY: "Elements of Statistics," P. S. King, London, 1920. (The second part is mathematical and advanced; the first part, elementary.)

B. H. CAMP: "Mathematical Part of Elementary Statistics," D. C. Heath and Company, Boston, 1931. (Especially good for students with more than average mathematical preparation.)

R. E. CHADDOCK: "Principles and Methods of Statistics," Houghton Mifflin Company, Boston, 1925.

A. L. CRELLE: "Calculating Tables," Nutt, Berlin, 1898. (Helpful in multiplication when computing machines are not available.)

F. E. CROXTON and D. J. COWDEN: "Practical Business Statistics," Prentice-Hall, Inc., New York, 1934. (Specializes particularly in business analysis.)

C. B. DAVENPORT and M. P. EKLAS: "Statistical Methods in Biology, Medicine and Psychology," John Wiley & Sons, Inc., New York, 1936. (An up-to-date, compact summary of statistical methods, generally useful but specialized as the title implies.)

D. H. DAVENPORT and F. V. SCOTT: "An Index to Business Indices," Business Publications, Inc., Chicago, 1937. (Describes the commoner index numbers in the business field.)

G. R. DAVIES and W. F. CROWDER: "Methods of Statistical Analysis in the Social Science," John Wiley & Sons, Inc., New York, 1933.

G. R. DAVIES and D. YODER: "Business Statistics," John Wiley & Sons, Inc., New York, 1941. (A good summary of sources of business data, and worth-while up-to-date material on the application of analysis of variance to simple statistical distributions.)

H. T. DAVIS and W. F. C. NELSON: "Elements of Statistics," Principia Press, Bloomington, Indiana, 1935. (Specialized in economic statistics, with good sections on curve fitting and the use of tables.)

E. E. DAY: "Statistical Analysis," The Macmillan Company, New York, 1930.

J. W. DUNLAP and A. K. KURTZ: "Handbook of Statistical Nomographs, Tables, and Formulas," World Book Company, Yonkers-on-Hudson,

New York, 1932. (A useful collection of nomographs and tables, and an invaluable collection of formulas, including several important corrections in published formulas.)

W. P. ELDERTON: "Frequency Curves and Correlation," C. and E. Layton, London, 1906. (Perhaps the best reference on the Pearson-type curves.)

MORDECAI EZEKIEL: "Methods of Correlation Analysis," John Wiley & Sons, Inc., New York, 1941. (Confined to correlation, but unusually lucid and complete.)

I. FISHER: "The Making of Index Numbers," Houghton Mifflin Company, Boston, 1922.

R. A. FISHER: "Statistical Methods for Research Workers," Oliver & Boyd, Edinburgh, 1938. (Advanced, and often difficult reading. But up to date and original.)

H. E. GARRETT: "Statistics in Psychology and Education," Longmans, Green & Company, New York, 1926. (Especially readable. Specialized, as title implies.)

G. I. GAVETT: "A First Course in Statistical Method," McGraw-Hill Book Company, Inc., New York, 1937. (Has useful sections reviewing briefly and clearly the elementary mathematics underlying statistical procedures.)

J. W. GLOVER: "Tables of Applied Mathematics in Finance, Insurance, and Statistics," George Wahr, Ann Arbor, Michigan, 1923. (Valuable tables, especially in description of the normal curve and in interpolation aids. Includes seven-place logarithms.)

K. J. HOLZINGER: "Statistical Methods for Students in Education," Ginn and Company, Boston, 1928. (Specialized, as the title implies.)

K. J. HOLZINGER: "Statistical Tables for Students in Education and Psychology," University of Chicago Press, Chicago, 1928. (Time-saving tables of general use in calculation.)

D. C. JONES: "A First Course in Statistics," G. Bell & Sons, London, 1921.

K. G. KARSTEN: "Charts and Graphs," Prentice-Hall, Inc., New York, 1923. (Thorough and painstaking. Much valuable illustrative material.)

T. L. KELLEY: "Statistical Method," The Macmillan Company, New York, 1923. (An advanced treatment. Dependable.)

W. I. KING: "Elements of Statistical Method," The Macmillan Company, New York, 1912. (A very simple presentation, now somewhat out of date.)

J. LIPKA: "Graphical and Mechanical Computation," John Wiley & Sons, Inc., New York, 1918. (Advanced.)

F. C. MILLS: "Statistical Methods Applied to Economics and Business," Henry Holt and Company, New York, 1938.

F. C. MILLS and D. H. DAVENPORT: "Manual of Problems and Tables in Statistics," Henry Holt and Company, New York, 1925.

J. R. MINER: "Tables of $\sqrt{1 - r^2}$ and $1 - r$; for Use in Partial Correlation and Trigonometry," Johns Hopkins Press, Baltimore, 1922.

W. C. MITCHELL: Making and Using Index Numbers, Part I, *U. S. Bureau of Labor Statistics Bulletin* 284, 1921.

R. PEARL: "Medical Biometry and Statistics," W. B. Saunders Company, Philadelphia, 1933. (Specialized for students of medicine and vital statistics.)

K. PEARSON: "Tables for Statisticians and Biometricians," Cambridge University Press, 1914. (Invaluable as a time saver.)

C. H. RICHARDSON: "An Introduction to Statistical Analysis," Harcourt, Brace and Company, New York, 1934.

H. L. RIETZ: "Mathematical Statistics," Open Court Publishing Company, Chicago, 1927. (Advanced, specialized, and highly mathematical, but very valuable.)

H. L. RIETZ *et al.*: "Handbook of Mathematical Statistics," Houghton Mifflin Company, Boston, 1924. (For advanced students only. Requires good grounding in mathematics. Includes essays on several important topics in the field of statistics.)

J. R. RIGGLEMAN and I. N. FRISBEE: "Business Statistics," McGraw-Hill Book Company, Inc., New York, 1932. (Useful material describing mechanical aids in statistical work and the preparation of statistical reports.)

H. O. RUGG: "Statistical Methods Applied to Education," Houghton Mifflin Company, Boston, 1917.

M. SASULY: "Trend Analysis of Statistics," Brookings Institution, Washington, D. C., 1934. (An invaluable book for its specialized field.)

H. SECRIST: "Introduction to Statistical Methods," The Macmillan Company, New York, 1929.

J. G. SMITH: "Elementary Statistics, An Introduction to the Principles of Scientific Methods," Henry Holt and Company, Inc., New York, 1934. (The sections on scientific method are particularly valuable.)

G. W. SNEDECOR: "Calculation and Interpretation of Analysis of Variance and Covariance," Collegiate Press, Ames, Iowa, 1934. (A very important and authoritative small book in a new but highly significant field.)

H. SORENSON: "Statistics for Students of Psychology and Education," McGraw-Hill Book Company, Inc., New York, 1936. (An authoritative text which covers the general field of statistics but gives especial emphasis to problems arising in psychology and education.)

H. M. WALKER: "Mathematics Essential for Elementary Statistics," Henry Holt and Company, Inc., New York, 1934. (Brief, simple, and self-teaching.)

H. M. WALKER: "Studies in the History of Statistical Method," Williams & Wilkins Company, Baltimore, 1929. (Interesting, instructive, and authoritative.)

H. A. WALLACE and G. W. SNEDECOR: "Correlation and Machine Calculation," Iowa State College Bookstore, Ames, Iowa, 1931. (Very useful simple directions for actual steps in calculation.)

G. C. WHIPPLE: "Vital Statistics," John Wiley & Sons, Inc., New York, 1919. (Contains many statistical problems and methods from this specialized field.)

E. T. WHITAKER and G. ROBINSON: "The Calculus of Observations," Blackie & Son, Ltd., London, 1924. (Advanced but important discussion of least-squares theory.)

G. U. YULE and M. G. KENDALL: "An Introduction to the Theory of Statistics," Charles Griffin & Company, Ltd., London, 1937. (While the treatment is somewhat advanced for beginners, nevertheless this is one of the most complete, authoritative, and useful treatments in the English language.)

AUTHOR INDEX

SUBJECT INDEX

T

U