Statistical Methods for Research Workers

BY

R. A. FISHER, M.A.

Fellow of Gonville and Caius College, Cambridge
Chief Statistician, Rothamsted Experiment Station

OLIVER AND BOYD

EDINBURGH: TWEEDDALE COURT

LONDON: 33 PATERNOSTER ROW, E.C.

1925

Classics in the History of Psychology

An internet resource developed by
Christopher D. Green
York University, Toronto, Ontario
ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted March 2000

EDITORS' PREFACE

The increasing specialisation in biological inquiry has made it impossible for any one author to deal adequately with current advances in knowledge. It has become a matter of considerable difficulty for a research student to gain a correct idea of the present state of knowledge of a subject in which he himself is interested. To meet this situation the text-book is being supplemented by the monograph.

The aim of the present series is to provide authoritative accounts of what has been done in some of the diverse branches of biological investigation, and at the same time to give to those who have contributed notably to the development of a particular field of inquiry an opportunity of presenting the results of their researches, scattered throughout the scientific journals, in a more extended form, showing their relation to what has already been done and to problems that remain to be solved.

The present generation is witnessing " a return to practice of older days when animal physiology [p. vi] was not yet divorced from morphology." Conspicuous progress is now being seen in the field of general physiology, of experimental biology, and in the application of biological principles to economic problems. In this series, therefore, it is intended that biological research, both pure and applied, shall be represented.

F.A.E. Crew, Edinburgh
D. Ward Cutler, Rothamsted [p. vii]

AUTHOR'S PREFACE

For several years the author has been working in somewhat intimate co-operation with a number of biological research departments; the present book is in every sense the product of this circumstance. Daily contact with the statistical problems which present themselves to the laboratory worker has stimulated the purely mathematical researches upon which are based the methods here presented. Little experience is sufficient to show that the traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow, but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small sample problems on their merits does it seem possible to apply accurate tests to practical data. Such at least has been the aim of this book.

I owe more than I can say to Mr. W. S. Gosset, Mr. E. Somerfield, and Miss W. A. Mackenzie, who [p. viii] have read the proofs and made many valuable suggestions. Many small but none the less troublesome errors have been removed; I shall be grateful to readers who will notify me of any further errors and ambiguities they may detect.

ROTHAMSTED EXPERIMENTAL STATION,
February 1925

ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted March 2000

I

INTRODUCTORY

1. The Scope of Statistics

The science of statistics is essentially a branch of Applied Mathematics and may be regarded as mathematics applied to observational data. As in other mathematical studies the same formula is equally relevant to widely different groups of subject matter. Consequently the unity of the different applications has usually been overlooked, the more naturally because the development of the underlying mathematical theory has been much neglected. We shall therefore consider the subject matter of statistics under three different aspects, and then show in more mathematical language that the same types of problems arise in every case. Statistics may be regarded as (i.) the study of **populations**, (ii.) as the study of **variation**, (iii.) as the study of methods of the **reduction of data**.

The original meaning of the word "statistics" [p. 2] suggests that it was the study of populations of human beings living in political union. The methods developed, however, have nothing to do with the political unity of the group, and are not confined to populations of men or of social insects. Indeed, since no observational record can completely specify a human being, the populations studied are to some extent abstractions. If we have records of the stature of 10,000 recruits, it is rather the population of statures than the population of recruits that is open to study. Nevertheless, in a real sense, statistics is the study of populations, or aggregates of individuals, rather than of individuals. Scientific theories which involve the properties of large aggregates of individuals, and not necessarily the properties of the individuals themselves, such as the Kinetic Theory of Gases, the Theory of Natural Selection, or the chemical Theory of Mass Action, are essentially statistical arguments; and are liable to misinterpretation as soon as the statistical nature of the argument is lost sight of. Statistical. methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences. This particular dependence of social studies upon statistical methods has led to the painful misapprehension that statistics is to be regarded as a branch of economics, whereas in truth economists have much to learn from their scientific contemporaries, not only in general scientific method, but in particular in statistical practice.

The idea of a population is to be applied not only [p. 3] to living, or even material, individuals. If an observation, such as a simple measurement, be repeated a number of times, the aggregate of the results is a population of measurements. Such populations are the particular field of study of the **Theory of Errors**, one of the oldest and most fruitful lines of statistical investigation. Just as a single observation may be regarded as an individual, and its repetition as generating a population, so the entire result of an extensive experiment may be regarded as but one of a population of such experiments. The salutary habit of repeating important experiments, or of carrying out original observations in replicate, shows a tacit appreciation of the fact that the object of our study is not the individual result, but the population of possibilities of which we do our best to make our experiments representative. The calculation of means and probable errors shows a deliberate attempt to find out something about that population.

The conception of statistics as the study of variation is the natural outcome of viewing the subject as the study of populations; for a population of individuals in all respects identical is completely described by a description of any one individual, together with the number in the group, The populations which are the object of statistical study always display variation in one or more respects. To speak of statistics as the study of variation also serves to emphasise the contrast between the aims of modern statisticians and those of their predecessors. For, until comparatively recent times, the vast majority [p. 4] of workers in this field appear to have had no other aim than to ascertain aggregate, or average, values.

The variation itself was not an object of study, but was recognised rather as a troublesome circumstance which detracted from the value of the average. The error curve of the *mean* of a normal sample has been familiar for a century, but that of the *standard deviation* has scarcely been securely established for a decade. Yet, from the modern point of view, the study of the causes of variation of any variable phenomenon, from the yield of wheat to the intellect of man, should be begun by the examination and measurement of the variation which presents itself.

The study of variation leads immediately to the concept of a **frequency distribution**. Frequency distributions are of various kinds, according as the number of classes in which the population is distributed is finite or infinite, and also according as the intervals which separate the classes are finite or infinitesimal. In the simplest possible case, in which there are only two classes, such as male and female births, the distribution is simply specified by the proportion in which these occur, as for example by the statement that 51 per cent of the births are of males and 49 per cent of females. In other cases the variation may be discontinuous, but the number of classes indefinite, as with the number of children born to different married couples; the frequency distribution would then show the frequency with which 0, 1, 2 ... children were recorded, the number of classes being sufficient to include the largest family in the record. [p. 5] The variable quantity, such as the number of children, is called the **variate**, and the frequency distribution specifies how frequently the variate takes each of its possible values. In the third group of cases, the variate, such as human stature, may take any intermediate value within its range of variation; the variate is then said to vary continuously, and the frequency distribution may be expressed by stating, as a mathematical function of the variate, either (i.) the proportion of the population for which the variate is less than any given value, or (ii.) by the mathematical device of differentiating this function, the (infinitesimal) proportion of the population for which the variate falls within any infinitesimal element of its range.

The idea of a frequency distribution is applicable either to populations which are finite in number, or to infinite populations, but it is more usefully and more simply applied to the latter. A finite population can only be divided in certain limited ratios, and cannot in any case exhibit continuous variation. Moreover, in most cases only an infinite population can exhibit accurately, and in their true proportion, the whole of the possibilities arising from the causes actually at work, and which we wish to study. The actual observations can only be a sample of such possibilities. With an infinite population the frequency distribution specifies the fractions of the populations assigned to the several classes; we may have (i.) a finite number of fractions adding up to unity as in the Mendelian frequency distributions, or (ii.) an infinite series of finite fractions adding up to unity, or (iii.) a mathematical [p. 6] function expressing the fraction of the total in each of the infinitesimal elements in which the range of the variate may be divided. The last possibility may be represented by a frequency curve; the values of the variate are set out along a horizontal axis, the fraction of the total population, within any limits of the variate, being represented by the area of the curve standing on the corresponding length of the axis. It should be noted that the familiar concept of the frequency curve is only applicable to infinite populations with continuous variates.

The study of variation has led not merely to measurement of the amount of variation present, but to the study of the qualitative problems of the type, or form, of the variation. Especially important is the study of the simultaneous variation of two or more variates. This study, arising principally out of the work of Galton and Pearson, is generally known in English under the name of **Correlation**, but by some continental writers as **Covariation**.

The third aspect under which we shall regard the scope of statistics is introduced by the practical need to reduce the bulk of any given body of data. Any investigator who has carried out methodical and extensive observations will probably be familiar with the oppressive necessity of reducing his results to a more convenient bulk. No human mind is capable of grasping in its entirety the meaning of any considerable quantity of numerical data. We want to be able to express all the *relevant* information contained in the mass by means of comparatively few numerical [p. 7] values. This is a purely practical need which the science of statistics is able to some extent to meet. In some cases at any rate it *is* possible to give the whole of the relevant information by means of one or a few values. In all cases, perhaps, it is possible to reduce to a simple numerical form the main issues which the investigator has in view, in so far as the data are competent to throw light on such issues. The number of independent facts supplied by the data is usually far greater than the number of facts sought, and in consequence much of the information supplied by any body of actual data is irrelevant. It is the object of the statistical processes employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data.

2. General Method, Calculation of Statistics

The discrimination between the irrelevant and the relevant information is performed as follows. Even in the simplest cases the values (or sets of values) before us are interpreted as a random sample of a hypothetical infinite population of such values as might have arisen in the same circumstances. The distribution of this population will be capable of some kind of mathematical specification, involving a certain number, usually few, of **parameters**, or "constants" entering into the mathematical formula. These parameters are the characters of the population. If we could know the exact specification of the population, we should know all (and more than) any sample from [p. 8] the population could tell us. We cannot in fact know the specification exactly, but we can make estimates of the unknown parameters, which will be more or less inexact. These estimates, which are termed statistics, are of course calculated from the observations. If we can find a mathematical form for the population which adequately represents the data, and then calculate from the data the best possible estimates of the required parameters, then it would seem that there is little, or nothing, more that the data can tell us; we shall have extracted from it all the available relevant information.

The value of such estimates as we can make is enormously increased if we can calculate the magnitude and nature of the errors to which they are subject. If we can rely upon the specification adopted, this presents the purely mathematical problem of deducing from the nature of the population what will be the behaviour of each of the possible statistics which can be calculated. This type of problem, with which until recent years comparatively little progress had been made, is the basis of the tests of significance by which we can examine whether or not the data are in harmony with any suggested hypothesis. In particular, it is necessary to test the adequacy of the hypothetical specification of the population upon which the method of reduction was based.

The problems which arise in the reduction of data may thus conveniently be divided into three types:

(i.) Problems of **Specification**, which arise in the choice of the mathematical form of the population. [p. 9]

(ii.) Problems of **Estimation**, which involve the choice of method of calculating, from our sample, statistics fit to estimate the unknown parameters of the population.

(iii.) Problems of **Distribution**, which include the mathematical deduction of the exact nature of the distribution in random samples of our estimates of the parameters, and of other statistics designed to test the validity of our specification (tests of **Goodness of Fit**).

The statistical examination of a body of data is thus logically similar to the general alternation of inductive and deductive methods throughout the sciences. A hypothesis is conceived and defined with necessary exactitude; its consequences are deduced by a deductive argument; these consequences are compared with the available observations; if these are completely in accord with the deductions, the hypothesis may stand at any rate until fresh observations are available.

The deduction of inferences respecting samples, from assumptions respecting the populations from which they are drawn, shows us the position in Statistics of the **Theory of Probability**. For a given population we may calculate the probability with which any given sample will occur, and if we can solve the purely mathematical problem presented, we can calculate the probability of occurrence of any given statistic calculated from such a sample. The Problems of Distribution may in fact be regarded as applications and extensions of the theory of probability. [p. 10] Three of the distributions with which we shall be concerned, Bernoulli's binomial distribution, Laplace's normal distribution, and Poisson's series, were developed by writers on probability. For many years, extending over a century and a half, attempts were made to extend the domain of the idea of probability to the deduction of inferences respecting populations from assumptions (or observations) respecting samples. Such inferences are usually distinguished under the heading of **Inverse Probability**, and have at times gained wide acceptance. This is not the place to enter into the subtleties of a prolonged controversy; it will be sufficient in this general outline of the scope of Statistical Science to express my personal conviction, which I have sustained elsewhere, that the theory of inverse probability is founded upon an error, and must be wholly rejected. Inferences respecting populations, from which known samples have been drawn, cannot be expressed in terms of probability, except in the trivial case when the population is itself a sample of a super-population the specification of which is known with accuracy.

This is not to say that we cannot draw, from knowledge of a sample, inferences respecting the population from which the sample was drawn, but that the mathematical concept of probability is inadequate to express our mental confidence or diffidence in making such inferences, and that the mathematical quantity which appears to be appropriate for measuring our order of preference among different possible populations does not in fact obey the laws of probability. [p. 11] To distinguish it from probability, I have used the term "**Likelihood**" to designate this quantity; since both the words "likelihood" and "probability" are loosely used in common speech to cover both kinds of relationship.

3. The Qualifications of Satisfactory Statistics

The solutions of problems of distribution (which may be regarded as purely deductive problems in the theory of probability) not only enable us to make critical tests of the significance of statistical results, and of the adequacy of the hypothetical distribution upon which our methods of numerical deduction are based, but afford some guidance in the choice of appropriate statistics for purposes of estimation. Such statistics may be divided into classes according to the behaviour of their distributions in large samples.

If we calculate a statistic, such, for example, as the mean, from a very large sample, we are accustomed to ascribe to it great accuracy; and indeed it would usually, but not always, be true, that if a number of such statistics could be obtained and compared, the discrepancies between them would grow less and less, as the samples from which they are drawn are made larger and larger. In fact, as the samples are made larger without limit, the statistic will usually tend to some fixed value characteristic of the population, and, therefore, expressible in terms of the parameters of the population. If, therefore, such a statistic is to be used to estimate these parameters, there is only one parametric function to which it can properly be equated. [p. 12] If it be equated to some other parametric function, we shall be using a statistic which even from an infinite sample does not give the correct value; it tends indeed to a fixed value, but to a value which is erroneous from the point of view with which it was used. Such statistics are termed **Inconsistent** Statistics; except when the error is extremely minute, as in the use of Sheppard's corrections, inconsistent statistics should be regarded as outside the pale of decent usage.

Consistent statistics, on the other hand, all tend more and more nearly to give the correct values, as the sample is more and more increased; at any rate, if they tend to any fixed value it is not to an incorrect one. In the simplest cases, with which we shall be concerned, they not only tend to give the correct value, but the errors, for samples of a given size, tend to be distributed in a well-known distribution (of which more in Chap. III.) known as the. Normal Law of Frequency of Error, or more simply as the **normal distribution**. The liability to error may, in such cases, be expressed by calculating the mean value of the squares of these errors, a value which is known as the variance; and in the class of cases with which we are concerned, the variance falls off with increasing samples, in inverse proportion to the number in the sample.

Now, for the purpose of estimating any parameter, it is usually possible to invent any number of statistics which shall be consistent in the sense defined above, and each of which has in large samples a variance falling off inversely with the size of the sample. But [p. 13] for large samples of a fixed size, the variance of these different statistics will generally be different. Consequently a special importance belongs to a smaller group of statistics, the error distributions of which tend to the normal distribution, as the sample is increased, with the least possible variance. We may thus separate off from the general body of consistent statistics a group of especial value, and these are known as **efficient** statistics.

The reason for this term may be made apparent by an example. If from a large sample of (say) 1000 observations we calculate an efficient statistic, A, and a second consistent statistic, B, having twice the variance of A, then B will be a valid estimate of the required parameter, but one definitely inferior to A in its accuracy. Using the statistic B, a sample of 2000 values would be required to obtain as good an estimate as is obtained by using the statistic A from a sample of 1000 values. We may say, in this sense, that the statistic B makes use of 50 per cent of the relevant information available in the observations; or, briefly, that its **efficiency** is 50 per cent. The term "efficient" in its absolute sense is reserved for statistics the efficiency of which is 100 per cent.

Statistics having efficiency less than 100 per cent may be legitimately used for many purposes. It is conceivable, for example, that it might in some cases be laborious to increase the number of observations than to apply a more elaborate method of calculation the results. It may often happen that an inefficient statistic is accurate enough to answer the particular [p. 14] questions at issue. There is, however, one limitation to the legitimate use of inefficient statistics which should be noted in advance. If we are to make accurate tests of goodness of fit, the methods of fitting employed must not introduce **errors of fitting** comparable to the **errors of random sampling**; when this requirement is investigated, it appears that when tests of goodness of fit are required, the statistics employed in fitting must be not only consistent, but must be of 100 per cent efficiency. This is a very serious limitation to the use of inefficient statistics, since. in the examination of any body of data it is desirable to be able at any time to test the validity of one or more of the provisional assumptions which have been made.

Numerous examples of the calculation of statistics will be given in the following chapters, and in these illustrations of method efficient statistics have been chosen. The discovery of efficient statistics in new types of problem may require some mathematical investigation. The investigations of the author have led him to the conclusion that an efficient statistic can in all cases be found by the **Method of Maximum Likelihood**; that is, by choosing statistics so that the estimated population should be that for which the likelihood is greatest. In view of the mathematical difficulty of some of the problems which arise it is also useful to know that *approximations* to the maximum likelihood solution are also in most cases efficient statistics. A simple example of the application of the method of maximum likelihood to a genetical problem is given at the end of this chapter. [p. 15]

For practical purposes it is not generally necessary to press refinement of methods further than the stipulation that the statistics used should be efficient. With large samples it may be shown that all efficient statistics tend to equivalence, so that little inconvenience arises from diversity of practice. There is, however, one class of statistics, including some of the most frequently recurring examples, which is of theoretical interest for possessing the remarkable property that, even in small samples, a statistic of this class alone includes the whole of the relevant information which the observations contain. Such statistics are distinguished by the term **sufficient**, and, in the use of small samples, sufficient statistics, when they exist, are definitely superior to other efficient statistics. Examples of sufficient statistics are the **arithmetic mean** of samples from the normal distribution, or from the Poisson Series; it is the fact of providing sufficient statistics for these two important types of distribution which gives to the arithmetic mean its theoretical importance. The method of maximum likelihood leads to these sufficient statistics where they exist.

While diversity of practice within the limits of efficient statistics will not with large samples lead to inconsistencies, it is, of course, of importance in all cases to distinguish clearly the parameter of the population, of which it is desired to estimate the value, from the actual statistic employed as an estimate of its value; and to inform the reader by which of the considerable variety of processes which exist for the purpose the estimate was actually obtained. [p. 16]

4. Scope of this Book

The prime object of this book is to put into the hands of research workers, and especially of biologists, the means of applying statistical tests accurately to numerical data accumulated in their own laboratories or available in the literature. Such tests are the result of solutions of problems of distribution, most of which are but recent additions to our knowledge and have so far only appeared in specialised mathematical papers. The mathematical complexity of these problems has made it seem undesirable to do more than (i.) to indicate the kind of problem in question, (ii.) to give numerical -illustrations by which the whole process may be checked, (iii.) to provide numerical tables by means of which the tests may be made without the evaluation of complicated algebraical expressions .

It would have been impossible to give methods suitable for the great variety of kinds of tests which are required but for the unforeseen circumstances that each mathematical solution appears again and again in questions which at first sight appeared to be quite distinct. For example, Pearson's solution in 1900 of the distribution of c^2 is in reality equivalent to the distribution of the variance as estimated from normal samples, of which the solution was not given until 1908, and then quite tentatively, and without complete mathematical proof, by "Student." The same distribution was found by the author for the index of dispersion derived from small samples from a Poisson [p. 17] Series. What is even more remarkable is that, though Pearson's paper of 1900 contained a serious error, which vitiated most of the tests of goodness of fit made by this method until 1921, yet the correction of this error leaves the form of the distribution unchanged, and only requires that some few units should be deducted from one of the variables with which the table of c^2 is entered.

It is equally fortunate that the distribution of *t*, first established by "Student" in 1908, in his study of the probable error of the mean, should be applicable, not only to the case there treated, but to the more complex, but even more frequently needed problem of the comparison of two mean values. It further provides an exact solution of the sampling errors of the enormously wide class of statistics known as regression coefficients.

In studying the exact theoretical distributions in a number of other problems, such as those presented by intraclass correlations, the goodness of fit of regression lines, the correlation ratio, and the multiple correlation coefficient, the author has been led repeatedly to a third distribution, which may be called the distribution of *z*, and which is intimately related to, and 'indeed a natural extension of, the distributions found by Pearson and "Student." It has thus been possible to classify the necessary distributions, covering a very great variety of cases, under these three main groups; and, what is equally important, to make some provision for the need of numerical values by means of a few tables only. [p. 18]

The book has been arranged so that the student may make acquaintance with these three main distributions in a logical order, and proceeding from more simple to more complex cases. Methods developed in later chapters are frequently seen to be generalisations of simpler methods developed previously. Studying the work methodically as a connected treatise, the student will, it is hoped, not miss the fundamental unity of treatment under which such very varied material has been brought together; and will prepare himself to deal competently and with exactitude with the many analogous problems, which cannot be individually exemplified. On the other hand, it is recognised that many will wish to use the book for laboratory reference, and not as a connected course of study. This use would seem desirable only if the reader will be at the pains to work through, in all numerical detail, one or more of the appropriate examples, so as to assure himself, not only that his data are appropriate for a parallel treatment, but that he has obtained some critical grasp of the meaning to be attached to the processes and results.

It is necessary to anticipate one criticism, namely, that in an elementary book, without mathematical proofs, and designed for readers without special mathematical training, so much has been included which from the teacher's point of view is advanced; and indeed much that has not previously appeared in print. By way *of* apology the author would like to put forward the following considerations.

(1) For non - mathematical readers, numerical [p. 19] tables are in any case necessary; accurate tables are no more difficult to use, though more laborious to calculate, than inaccurate tables embodying the current approximations.

(2) The process of calculating a probable error from one of the established formulæ gives no real insight into the random sampling distribution, and can only supply a test of significance by the aid of a table of deviations of the normal curve, and on the assumption that the distribution is in fact very nearly normal. Whether this procedure should, or should not, be used must be decided, not by the mathematical attainments of the investigator, but by discovering whether it will or will not give a sufficiently accurate answer. The fact that such a process has been used successfully by eminent mathematicians in analysing very extensive and important material does not imply that it is sufficiently accurate for the laboratory worker anxious to draw correct conclusions from a small group of perhaps preliminary observations.

(3) The exact distributions, with the use of which this book is chiefly concerned, have been in fact developed in response to the practical problems arising in biological and agricultural research; this is true not only of the author's own contribution to the subject, but from the beginning of the critical examination of statistical distributions in "Student's " paper of 1908.

The greater part of the book is occupied by numerical examples; and these perhaps could with advantage have been increased in number. In choosing them it has appeared to the author a hopeless task [p. 20] to attempt to exemplify the great variety of subject matter to which these processes may be usefully applied. There are no examples from astronomical statistics, in which important work has been done in recent years, few from social studies, and the biological applications are scattered unsystematically. The examples have rather been chosen each to exemplify a particular process, and seldom on account of the importance of the data used, or even of similar examinations of analogous data. By a study of the processes exemplified, the student should be able to ascertain to what questions, in his own material, such processes are able to give a definite answer; and, equally important, what further observations would be necessary to settle other outstanding questions. In conformity with the purpose of the examples the reader should remember that they do not pretend to be critical examinations of general scientific questions, which would require the examination of much more extended data, and of other evidence, but are solely concerned with the evidence of the particular batch of data presented.

5. Mathematical Tables

The tables of distributions supplied at the ends of several chapters form a part essential to the use of the book.

TABLES I. AND II.-The importance of the normal distribution has been recognised at least from the time of Laplace. (The formula has even been traced back to a little-known work by De Moivre of 1733) Numerous tables have given in one form or another the relation between the deviation, and the probability of a greater deviation. Important sources for these values are

> J. Burgess (1895), *Trans. Roy. Soc. Edin.*, XXXIX. pp. 257-321;
>
> J. W. L. Glaisher (1871), *Phil. Mag.*, Series IV. Vol. XLII. p. 436.

The very various forms in which this relation has been tabulated adds considerably to the labour of practical applications. The form which we have adopted for this, and for the other tables, has been used for the normal distribution by

> F. Galton and W. F. Sheppard (1907), *Biometrika,* V. p. 405;
>
> T. L. Kelley, *Statistical Method*, pp. 373-385;

both of which are valuable tables, on a more extensive scale than Table I. In Table II. we have given the normal deviations corresponding to very high odds. It should be remembered that even slight departures from the normal distribution will render these very small probabilities relatively very inaccurate, and that we seldom can be certain, in any particular case, that these high odds will be accurate. The table illustrates the general fact that the significance in the normal distribution of deviations exceeding four times the standard deviation is extremely pronounced.

TABLE III.; table of c^2. -- Tables of the value of *P* for different values of c^2 and *n'*, were given by

> K. Pearson (1900), *Phil. Mag.*, Series V. Vol. L. p. 175; [p. 22]

> W. P. Elderton (1902), *Biometrika*, I. pp. 155-163; the same relationship in a much modified form underlies

> K. Pearson (I922), *Tables of the incomplete* G-*function.*

Table III. gives the values of c^2 for different values of *P* and *n,* in a form designed for rapid laboratory use, and with a view to covering in sufficient detail the range of values actually occurring in practice. For higher values of *n* the test is supplemented by an easily calculated approximate test.

TABLE IV.; table of *t.* -- Tables of the same distribution as that of *t* have been given by

> "Student " (1908), *Biometrika*, VI. p. 19;

> "Student" (1917), *Biometrika*, XI. pp. 414-417.

"Student" gives the value of (1-½*P*) for different values of *z* (=*t*/[sqrt]*n* in our notation) and *n* (=*n*+1 in our notation). As in the case of the table of c^2, the very much extended application of this distribution has led to a reinterpretation of the meaning of *n* to cover a wider class of cases. Extended tables giving the values of *P* for different values of *t* are in preparation by the same author. For the purposes of the present book we require the values of *t* corresponding to given values of *P* and *n*.

TABLE V. A gives the values of the correlation coefficient for different levels of significance, according to the extent of the sample upon which the value is based. From this table the reader may see at a glance whether or not any correlation obtained may be regarded as significant, for samples up to 100 pairs of observations. [p. 23]

TABLE V. B gives the values of the well-known mathematical function, the hyperbolic tangent, which we have introduced in the calculation of sampling errors of the correlation coefficient. The function is simply related to the logarithmic and exponential functions, and may be found quite easily by such a convenient table of natural logarithms as is given in

> J. T. Bottomley, *Four-figure Mathematical Tables,*

while the hyperbolic tangent and its inverse appear in

> W. Hall, *Four-figure Tables and Constants.*

A table of natural logarithms is in other ways a necessary supplement in using this book, as in other laboratory calculations. Tables of the inverse hyperbolic tangent for correlational work have been previously given by

> R. A. Fisher (1921), *Metron.* Vol. I. No.4, pp. 26-27.

TABLE VI.; table of *z*. -- Tests involving the use of *z*, including as special cases the use of c^2 and of *t*, are so widespread, that it is probable that a more extended table of this function will be necessary. The exploration of this function is of such recent date, and the construction of a table of triple entry is such a laborious task, that all that can be offered at present is the small table corresponding to the important region, *P*= .05 It is probable, indeed, that if supplemented by a similar table for *P*=.01, all ordinary requirements would be met, although to avoid the labour of interpolation much larger tables for these two values would be needed.

At present I can only beg the reader's indulgence [p. 24] for the inadequacy of the present table, pleading in my defence that, on ground so recently won as is that occupied by the greater part of this book, the full facilities and conveniences which many workers can gradually accumulate cannot yet be expected.

6. The following example exhibits in a relatively simple case the application of the method of maximum likelihood to discover a statistic capable of giving an efficient estimate of an unknown parameter. Since this procedure belongs rather to the advanced mathematical treatment of theoretical statistics, it may be noted that to master it is not a necessary preliminary to understanding the practical methods developed in the rest of the book. Students, however, who wish to apply the fundamental principles mentioned in this introductory chapter to new types of data, may perhaps be glad of an example of the general procedure.

Ex. 1. *The derivation of an efficient statistic by means of the method of maximum likelihood.* -- Animals or plants heterozygous for two linked factors showing complete dominance are self fertilised ; if all four types are equally viable, how should the extent of linkage be estimated from the numerical proportions of the four types of offspring?

If the allelomorphs of the first factor are *A* and *a*, and of the second factor *B* and *b*, the four types of gametes *AB, Ab, aB* and *ab* will be produced by the males and females in proportions depending on the linkage of the factors, subject to the condition that the allelomorphs of each factor occur equally frequently. [p. 25] The proportions will the two sexes; suppose the proportions to be

	AB	Ab	aB	ab	
In the female . . .	$\frac{1}{2}p$	$\frac{1}{2}q$	$\frac{1}{2}q$	$\frac{1}{2}p$,	$p+q=1$;
In the male . . .	$\frac{1}{2}p'$	$\frac{1}{2}q'$	$\frac{1}{2}q'$	$\frac{1}{2}p'$,	$p'+q'=1$;

then, if the two dominant genes are derived from the same parent, *q*, *q'* will be the cross-over ratios, if from different parents the cross-over ratios will be *p*, *p*'.

By taking all possible combinations of the gametes, it appears that the four types of offspring will occur in the proportions

$A\text{–}B\text{–}$	$A\text{–}bb$	$aaB\text{–}$	$aabb$
$\frac{1}{4}(2+pp')$	$\frac{1}{4}(1-pp')$	$\frac{1}{4}(1-pp')$	$\frac{1}{4}pp'$.

The effect of linkage is wholly expressed by the quantity *pp'*, and from a sample of observations giving observed frequencies a, b, g, d, we require to obtain an estimate of the value of *pp'*. The rule for applying the method of maximum likelihood is to multiply each observed frequency by the logarithm of the corresponding theoretical frequency, and to find the value of the unknown quantity which makes the total of these products a maximum. Writing *x* for *pp'*,

$$\text{a} \log (2+x) + (\text{b}+\text{g}) \log (1-x) + \text{d} \log x$$

is to be made a maximum; by a well-known application of the differential calculus, this requires that

$$\frac{a}{2+x}+\frac{\delta}{x}=\frac{\beta+\gamma}{1-x},$$

which leads to the quadratic equation for *x*,

$$(a+b+g+d)x^2 - (a-2b-2g-d)x - 2d = 0, \text{ [p. 26]}$$

the positive solution of which is the most likely value for *pp'*, as judged from the data.

For two factors in Primula the following numbers were observed (de Winton and Bateson's data):

$$a=396, b=99, g=104, d=70;$$

the quadratic for *x* is

$$669x^2 + 80x - 140 = 0,$$

of which the positive solution is $x = .4016$. To obtain the cross-over values in the two sexes separately, using self-fertilisation only, it would of course be necessary to repeat the experiment with heterozygotes of the opposite composition.

The numbers expected, on the supposition that $pp' = 4016$, are :

Expectation . .	401·7	100·1	100·1	67·2
Observation . .	396	99	104	70

Classics in the History of Psychology

An internet resource developed by
Christopher D. Green
York University, Toronto, Ontario
ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted March 2000

II

DIAGRAMS

7. The preliminary examination of most data is facilitated by the use of diagrams. Diagrams prove nothing, but bring outstanding features readily to the eye; they are therefore no substitute for such critical tests as may be applied to the data, but are valuable in suggesting such tests, and in explaining the conclusions founded upon them.

8. Time Diagrams, Growth Rate and Relative Growth Rate

The type of diagram in most frequent use consists in plotting the values of a variable, such as the weight of an animal or of a sample of plants against its age, or the size of a population at successive intervals of time. Distinction should be drawn between those cases in which the same group of animals, as in a feeding experiment, is weighed at successive intervals of time, and the cases, more characteristic of plant physiology, in which the same individuals cannot be used twice, but a parallel sample is taken at each age. The same. distinction occurs in counts of micro-organisms [p. 28] between cases in which counts are made from samples of the same culture, or from samples of parallel cultures. If it is of importance to obtain the general form of the growth curve, the second method has the advantage that any deviation from the expected curve may be confirmed from independent evidence at the next measurement, whereas using the same material no such independent confirmation is obtainable. On the other hand, if interest centres on the growth rate, there is an advantage in using the same material, for only so are actual increases in weight measurable. Both aspects of the difficulty can be got over only by replicating the observations; by carrying out measurements on a number of animals under parallel treatment it is possible to test, from the individual weights, though not from the means, whether their growth curve corresponds with an assigned theoretical course of development, or differs significantly from it or from a series differently tested. Equally, if a number of plants from each sample are weighed individually, growth -rates may be obtained with known probable errors, and so may be used for critical comparisons. Care should of course be taken that each is strictly a random sample.

Fig. 1 represents the growth of a baby weighed to the nearest ounce at weekly intervals from birth. Table 1 indicates the calculation from these data of the absolute growth rate in ounces per day and the relative growth rate per day. The absolute growth rates, representing the average actual rates at which substance is added during each period, are found by [p. 29] [figure] [p. 30] subtracting from each value that previously recorded, and dividing by the length of the period. The relative growth rates measure the rate of increase not only per unit of time, but per unit of weight already attained; using the mathematical fact, that

$$\frac{1}{m}\frac{dm}{dt}=\frac{d}{dt}(\log_e m),$$

[p. 31]

it is seen that the true average value of the relative growth rate for any period is obtained from the natural logarithms of the successive weights, just as the actual rates of increase are from the weights themselves. Such relative rates of increase are conveniently multiplied by 100, and therefore expressed as the percentage rate of increase per day. If these percentage rates of increase had been calculated on the principle of simple interest, by dividing the actual increase by the weight at the beginning of the period, somewhat higher values would have been obtained; the reason for this is that the actual weight of the baby at any time during each period is usually somewhat higher than its weight at the beginning. The error introduced by the simple interest formula becomes exceedingly great when the percentage increases between successive weighings are large.

Fig. 1A shows the course of the increase in absolute weight ; the average slope of such a diagram shows the absolute rate of increase. In this diagram the points fall approximately on a straight line, showing that the absolute rate of increase was nearly constant at about 1.66 oz. per diem. Fig. 1B shows the course of the increase in the natural logarithm of the weight; the slope at any point shows the relative rate of increase, which, apart from the first week, falls off perceptibly with increasing age. The features of such curves are best brought out if the scales of the two axes are so chosen that the line makes approximately equal angles with the two axes; with nearly vertical, or nearly horizontal lines, changes in the slope are not so readily perceived. [p. 32]

A rapid and convenient way of displaying the line of increase of the logarithm is afforded by the use of graph paper in which the horizontal rulings are spaced on a logarithmic scale, with the actual values indicated in the margin. The horizontal scale can then be adjusted to give the line an appropriate slope. This method avoids the use of a logarithm table, which, however, will still be required if the values of the relative rate of increase are needed.

In making a rough examination of the agreement of the observations with any law of increase, it is desirable so to manipulate the variables that the law to be tested will be represented by a straight line. Thus Fig. 1A is suitable for a rough test of the law that the absolute rate of increase is constant ; if it were suggested that the relative rate of increase were constant, Fig. 1B would show clearly that this was not so. With other hypothetical growth curves other transformations may be used; for example, in the so-called "autocatalytic" curve the relative growth rate falls off in proportion to the actual weight attained at any time. If, therefore, the relative growth rate be plotted against the actual weight, the points should fall on a straight line if the "autocatalytic" curve fits the facts. For this purpose it is convenient to plot against each observed weight the mean of the two adjacent relative growth rates. To do this for the above data for the growth of an infant may be left as an exercise to the student; twelve points will be available for weights 114 to 254 ounces. The relative growth rates, even after averaging adjacent pairs, will be very irregular, [p. 33] so that no clear indications will be found from these data. If a straight line is found to fit the data, it should be produced to meet the horizontal axis to find the weight at which growth ceases.

9. Correlation Diagrams

Although most investigators make free use of diagrams in which an uncontrolled variable is plotted against the time, or against some controlled factor such as concentration of solution, or temperature, much more use might be made of correlation diagrams in which one uncontrolled factor is plotted against another. When this is done as a dot diagram, a number of dots are obtained each representing a single experiment, or pair of observations, and it is usually clear from such a diagram whether or not any close connexion exists between the variables. When the observations are few a dot diagram will often tell us whether or not it is worth while to accumulate observations of the same sort; the range and extent of our experience is visible at a &lance ; and associations may be revealed which are worth while following up.

If the observations are so numerous that the dots cannot be clearly distinguished, it is best to divide up the diagram into squares, recording the frequency in each; this semi-diagrammatic record is a correlation table.

Fig. 2 shows in a dot diagram the yields obtained from an experimental plot of wheat (dunged plot, Broadbalk field, Rothamsted) in years with different [p. 34] total rainfall. The plot was under uniform treatment during the whole period 1854-1888; the 35 pairs of observations, indicated by 35 dots, show well the association of high yield with low rainfall. Even when few observations are available a dot diagram may suggest associations hitherto unsuspected, or what is equally important, the absence of associations which would have been confidently predicted. Their value lies in giving a simple conspectus of the experience hitherto gathered, and in bringing to the mind suggestions [p. 35] which may be susceptible of more exact statistical examination.

Instead of making a dot diagram the device is sometimes adopted of arranging the values of one variate in order of magnitude, and plotting the values of a second variate in the same order. If the line so obtained shows any perceptible slope, or general trend, the variates are taken to be associated. Fig. 3 represents the line obtained far rainfall, when the years are arranged in order of wheat yield. Such diagrams are usually far less informative than the diagram, and often conceal features of importance brought out by the-former. In addition the dot diagram possesses the advantage that it is easily used [p. 36] as a correlation table if the number of dots is small, and easily transformed into one if the number of dots is large.

In the correlation table the values of both variates are divided into classes, and the class intervals should be equal for all values of the same variate. Thus we might divide the value for the yield of wheat throughout at intervals of one bushel, and the values of the rainfall at intervals of 1 inch. The diagram is thus divided into squares, and the number *of* observations falling into each square is counted and recorded. The correlation table is useful for three distinct purposes. It affords a valuable visual representation of the whole of the observations, which with a little experience is as easy to comprehend as a dot diagram; it serves as a compact record of extensive data, which, as far as the two variates are concerned, is complete. With more than two variates correlation tables may be given for every pair. This will not indeed enable the reader to reconstruct the original data in its entirety, but it is a fortunate fact that for the great majority of statistical purposes, a set of such twofold distributions provides complete information. Original data involving more than two variates is most conveniently recorded for reference on cards, each case being given a separate card with the several variates entered in corresponding positions upon them. The publication of such complete data presents difficulties, but it is not yet sufficiently realised how much of the essential information can be presented in a compact form by means of correlation tables. The third feature of value about [p. 37] the correlation table is that the data so presented form a convenient basis for the immediate application of methods of statistical reduction. The most important statistics which the data provide can be most readily calculated from the correlation table. An example of a correlation table is shown in Table 31, p. 140.·

10. Frequency Diagrams

When a large number of individuals are measured in respect of physical dimensions, weight, colour, density, etc., it is possible to describe with some accuracy the *population* of which our experience may be regarded as a sample. By this means it may be possible to distinguish it from other populations differing in their genetic origin, or in environmental circumstances. Thus local races may be very different as populations, although individuals may overlap in all characters; or, under experimental conditions, the aggregate may show environmental effects, on size, death-rate, etc., which cannot be detected in the individual. A visible representation of a large number of measurements of any one feature is afforded by a frequency diagram. The feature measured is used as abscissa, or measurement along the horizontal axis, and as ordinates are set off vertically the *frequencies*, corresponding to each range.

Fig. 4 is a frequency diagram illustrating the distribution in stature of 1375 women (Pearson and Lee's data modified), The whole sample of women is divided up into successive height ranges of one inch. [p. 38] Equal areas on the diagram represent equal frequency; if the data be such that the ranges into which the individuals are subdivided are not equal, care should be taken to make the areas correspond to the observed frequencies, so that the area standing upon any interval of the base line shall represent the actual frequency observed in that interval.

The class containing the greatest number of observations is technically known as the modal class. In Fig. 4 the modal class indicated is the class whose central value is 63 inches. When, as is very frequently the case, the variate varies continuously, so that all intermediate values are possible, the choice of the grouping interval and limits is arbitrary and will make a perceptible difference to the appearance of the diagram. Usually, however, the possible limits of grouping will be governed by the smallest units in which the measurements are recorded. If, for example, measurements of height were made to the nearest [p. 39] quarter of an inch, so that all values between 66-7/8 inches and 67-1/8 Were recorded as 67 inches, all values between 67-1/8 and 67-3/8 were recorded as 67-1/4, then we have no choice but to take as our unit of grouping 1, 2, 3, 4, etc., quarters of an inch, and the limits of each group must fall on some odd number of eighths of an inch. For purposes of calculation the smaller grouping units are more accurate, but for diagrammatic purposes coarser grouping is often preferable. Fig. 4 indicates a unit of grouping suitable in relation to the total range for a large sample ; with smaller samples a coarser grouping is usually necessary in order that sufficient observations may fall in each class.

In all cases where the variation is continuous the frequency diagram should be in the form of a histogram, rectangular areas standing on each grouping interval showing the frequency of observations in that interval. The alternative practice of indicating the frequency by a single ordinate raised from the centre of the interval is sometimes preferred, as giving to the diagram a form more closely resembling a continuous curve. The advantage is illusory, for not only is the form of the curve thus indicated somewhat misleading, but the utmost care should always be taken to distinguish the infinitely large hypothetical population from which our sample of observations is drawn, from the actual sample of observations which we possess; the conception of a continuous frequency curve is applicable only to the former, and in illustrating the latter no attempt should be made to slur over this distinction. [p. 40]

This consideration should in no way prevent a frequency curve fitted to the data, from being super-imposed upon the histogram (as in Fig. 4); the contrast between the histogram representing the sample, and the continuous curve representing an estimate of the form of the hypothetical population, is well brought out in such diagrams, and the eye is aided in detecting any serious discrepancy between the observations and the hypothesis. No eye observation of such diagrams, however experienced, is really capable of discriminating whether or not the observations differ from expectation by more than we should expect from the circumstances of random sampling. Accurate methods of making such tests will be developed in later chapters.

With discontinuous variation, when, for example, the variate is confined to whole numbers, the above reason for insisting on the histogram form has little weight, for there are, strictly speaking, no ranges of variation within each class. On the other hand, there is no question of a frequency curve in such cases. Representation of such data by means of a histogram is usual and not inconvenient; it is especially appropriate if we regard the discontinuous variation as due to an underlying continuous variate, which can, however, express itself only to the nearest whole number.

It is, of course, possible to treat the values of the frequency like any other variable, by plotting the value of its logarithm, or its actual value on logarithmic paper, when it is desired to illustrate the agreement [p. 41] of the observations with any particular law of frequency. Fig. 5 shows in this way. the number of flowers (buttercups) having 5 to 10 petals (Pearson's data), plotted upon logarithmic paper, to facilitate comparison with the hypothesis that the frequency, for petals above five, falls off in geometric progression. Such illustrations are not, properly speaking, frequency diagrams, although the frequency is one of the variables [p. 42] employed, because they do not adhere to the convention that equal frequencies are represented by equal areas.

A useful form, similar to the above, is used to compare the death-rates, throughout life, of different populations. The logarithm of the number of survivors at any age is plotted against the age attained. Since the death-rate is the rate *of* decrease of the logarithm of the number of survivors, equal gradients on such curves represent equal death-rates. They therefore serve well to show the increase of death-rate with increasing age, and to compare populations with different death-rates. Such diagrams are less sensitive to small fluctuations than would be the corresponding frequency diagrams showing the distribution of the population according to age at death; they are therefore appropriate when such small fluctuations are due principally to errors of random sampling, which in the more sensitive type of diagram might obscure the larger features of the comparison. It should always be remembered that the choice of the appropriate methods of statistical treatment is quite independent of the choice of methods of diagrammatic representation.

Classics in the History of Psychology

An internet resource developed by
Christopher D. Green
York University, Toronto, Ontario
ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted March 2000

III

DISTRIBUTIONS

11. The idea of an infinite **population** distributed in a **frequency distribution** in respect of one or more characters is fundamental to all statistical work. From a limited experience, for example, of individuals of a species, or of the weather of a locality, we may obtain some idea of the infinite hypothetical population from which our sample is drawn, and so of the probable nature of future samples to which our conclusions are to be applied. If a second sample belies this expectation we infer that it is, in the language of statistics, drawn from a different population; that the treatment to which the second sample of organisms had been exposed did in fact make a material difference, or that the climate (or the methods of measuring it) had materially altered. Critical tests of this kind may be called tests of significance, and when such tests are available we may discover whether a second sample is or is not significantly different from the first.

A **statistic** is a value calculated from an observed sample with a view to characterising the population [p. 44] from which it is drawn. For example, the *mean* of a number of observations $x_1, x_2, \ldots x_n$, is given by the equation

$$\bar{x} = \frac{1}{n} S(x),$$

where S stands for summation over the whole sample, and n for the number of observations. Such statistics are of course variable from sample to sample, and the idea of a frequency distribution is applied with especial value to the variation of such statistics. If we know exactly how the original population was distributed it is theoretically possible, though often a matter of great mathematical difficulty, to calculate how any statistic derived from a sample of given size will be distributed. The utility of any particular statistic, and the nature of its distribution, both depend on the original distribution, and appropriate and exact methods have been worked out for only a few cases. The application of these cases is greatly extended by the fact that the distribution of many statistics tends to the **normal** form as the size of the sample is increased. For this reason it is customary to assume that such statistics are normally distributed, and to limit consideration of their variability to calculations of the standard error or probable error.

In the present chapter we shall give some account of three principal distributions -- (i.) the normal distribution, (ii.) the Poisson Series, (iii.) the binomial distribution. It is important to have a general knowledge of these three distributions, the mathematical formulæ by which they are represented, the experimental [p. 45] conditions upon which they occur, and the statistical methods of recognising their occurrence. On the latter topic we shall be led to some extent to anticipate methods developed more systematically in Chaps. IV. and V.

12. The Normal Distribution

A variate is said to be normally distributed when it takes all values from -[infinity], to +[infinity] , with frequencies given by a definite mathematical law, namely, that the logarithm of the frequency at any distance x from the centre of the distribution is less than the logarithm of the frequency at the centre by a quantity proportional to x^2. The distribution is therefore symmetrical, with the greatest frequency at the centre; although the variation is unlimited, the frequency falls off to exceedingly small values at any considerable distance from the centre, since a large negative logarithm corresponds to a very small number. Fig. 6B represents a normal curve of distribution. The frequency [p. 46] in any infinitesimal range *dx* may be written as

$$df = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\cdot\frac{(x-m)^2}{\sigma^2}} dx,$$

where *x-m* is the distance of the observation, *x*, from the centre of the distribution, *m*; and s, called the **standard deviation**, measures in the same units the extent to which the individual values are scattered. Geometrically s is the distance, on either side of the centre, of the steepest points, or points of inflexion of the curve (Fig. 4).

In practical applications we do not so often want to know the frequency at any distance from the centre as the total frequency beyond that distance; this is represented by the area of the tail of the curve cut off at any point. Tables of this total frequency, or probability integral, have been constructed from which, for any value of

$$\frac{x-m}{\sigma},$$

we can find what fraction of the total population has a larger deviation; or, in other words, what is the probability that a value so distributed, chosen at random, shall exceed a given deviation. Tables I. and II. have been constructed to show the deviations corresponding to different values of this probability. The rapidity with which the probability falls off as the deviation increases is well shown in these tables. A deviation exceeding the standard deviation occurs about once in three trials. Twice the standard deviation is exceeded only about once in 22 trials, thrice the standard deviation only once in 370 trials, while Table II. shows that to exceed the standard deviation sixfold would need [p. 47] nearly a thousand million trials. The value for which P =·.05, or 1 in 20, is 1.96 or nearly 2 ; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally regarded as significant. Using this criterion, we should be led to follow up a negative result only once in 22 trials, even if the statistics are the only guide available. Small effects would still escape notice if the data were insufficiently numerous to bring them out, hut no lowering of the standard of significance would meet this difficulty.

Some little confusion is sometimes introduced by the fact that in some cases we wish to know the probability that the deviation, known to be positive, shall exceed an observed value, whereas in other cases the probability required is that a deviation, which is equally frequently positive and negative, shall exceed an observed value; the latter probability is always half the former. For example,

Table I. shows that the normal deviate falls outside the range [plus or minus]1.598193 in 10 per cent of cases, and consequently that it exceeds +1.598193 in 5 per cent of cases.

The value of the deviation beyond which half the observations lie is called the **quartile** distance, and bears to the standard deviation the ratio .67449.·It is therefore a common practice to calculate the standard error and then, multiplying it by this factor, to obtain the **probable error**. The probable error is thus about two-thirds of the standard error, and as a test of significance a deviation of three times [p. 48] the probable error is effectively equivalent to one of twice the standard error. The common use of the probable error is its only recommendation ; when any critical test is required the deviation must be expressed in terms of the standard error in using the probability integral table.

13. Fitting the Normal Distribution

From a sample of *n* individuals of a normal population the mean and standard deviation of the population may be **estimated** by means of two easily calculated statistics. The best estimate of *m* is *x* where

$$\bar{x} = \frac{\text{I}}{n}\text{S}(x),$$

while for the best estimate of s, we calculate *s* from

$$s^2 = \frac{1}{n-1} S(x - \bar{x})^2 \ ;$$

these two statistics are calculated from the first two **moments** (*see* Appendix, p. 74) of the sample, and are specially related to the normal distribution, in that they summarise the whole of the information which the sample provides as to the distribution from which it was drawn, provided the latter was normal. Fitting by moments has also been widely applied to skew (asymmetrical) curves, and others which are not normal; but such curves have not the peculiar properties which make the first two moments especially appropriate, and where the curves differ widely from the normal form the above two statistics may be of little or no use.

Ex. 2. *Fitting a normal distribution to a large* [p. 49] *sample*. -- In calculating the statistics from a large sample it is not necessary to calculate individually the squares of the deviations from the mean of each measurement. The measurements are grouped together in equal intervals of the variate, and the whole of the calculation may be carried out rapidly as shown in Table 2, where the distribution of the stature of 1164 men is analysed.

The first column shows the central height in inches of each group, followed by the corresponding frequencies. A central group (68.5") is chosen as "working mean." To form the next column the frequencies are multiplied by 1, 2, 3, etc., according to their distance from the working mean; this process being repeated to form the fourth column, which is summed from top to bottom in a single operation; in the third column, however, the upper portion, representing negative deviations, is summed separately, and subtracted from the sum of the lower portion. The difference, in this case positive, shows that the whole sample of 1164 individuals has in all 167 inches more than if every individual were 68.5" in height. This balance divided by 1164 gives the amount by which the mean of the sample exceeds 68.5". The mean of the sample is therefore 68.6435". The sum of the fourth column is also divided by 1164, and gives an uncorrected estimate of the variance; two corrections are then applied -- one is for the fact that the working mean differs from the true mean, and consists in subtracting the square of the difference ; the second, which is Sheppard's correction for grouping, [p. 50] [table] [p. 51] allows for the fact that the process of grouping tends somewhat to exaggerate the variance, since in each group the values with deviations smaller than the central value will generally be more numerous than the values with deviations larger than the central value. Working in units of grouping, this correction is easily applied by subtracting a constant quantity 1/12 (=.0833) from the variance. From the variance so corrected the standard deviation is obtained by taking the square root. This process may be carried through as an exercise with the distribution of female statures given in the same table (p. 103).

Any interval may be used as a unit of grouping; and the whole calculation is carried through in such units, the final results being transformed into other units if required, just as we might wish to transform the mean and standard deviation from inches to centimetres by multiplying by the appropriate factor. It is advantageous that the units of grouping should be exact multiples of the units of measurement ; so that if the above sample had been measured to tenths of an inch, we might usefully have grouped them at intervals of 0.6" or 0.7".

Regarded as estimates of the mean and standard deviation of a normal population of which the above is regarded as a sample, the values found are affected by errors of random sampling; that is, we should not expect a second sample to give us exactly the same values. The values for different (large) samples of the same size would, however, be distributed very accurately in normal distributions, so the accuracy of [p. 52] any one such estimate may be satisfactorily expressed by its standard error. These standard errors may be calculated from the standard deviation of the population, and in treating large samples we take our estimate of this standard deviation as the basis of the calculation. The formulæ for the standard errors of random sampling of estimates of the mean and standard deviation of a large normal sample are (as given in Appendix, p. 75)

$$\frac{\sigma}{\sqrt{n}}, \quad \frac{\sigma}{\sqrt{2n}},$$

and their numerical values have been appended to the quantities to which they refer. From these values it is seen that our sample shows significant aberration from any population whose mean lay outside the limits 68.48"-68.80", and it is therefore likely that the mean of the population from which it was drawn lay between these limits; similarly it is likely that its standard deviation lay between 2.59" and 2·81".

It may be asked, Is nothing lost by grouping? Grouping in effect replaces the actual data by fictitious data placed arbitrarily at the central values of the groups; evidently a very coarse grouping might be very misleading. It has been shown that as regards obtaining estimates of the parameters of a normal population, the loss of information caused by grouping is less than 1 per cent, provided the group interval does not exceed one quarter of the standard deviation ; the grouping of the above sample in whole inches is thus somewhat too coarse; the loss in the estimation of the [p. 53] standard deviation is 2.28 per cent, or about 27 observations out of 1164; the loss in the estimation of the mean is half as great. With suitable group intervals, however, little is lost by grouping, and much labour is saved.

Another way of regarding the loss of information involved in grouping is to consider how near the values obtained for the mean and standard deviation will be to the values obtained without grouping. From this point of view we may calculate a standard error of grouping, not to be confused with the standard error of random sampling which measures the deviation of the sample values from the population value. In grouping units, the standard error due to grouping of both the mean and the standard deviation is

$$\frac{1}{\sqrt{12n}},$$

or in this case 0085". For sufficiently fine grouping this should not exceed one-tenth of the standard error of random sampling.

In the above analysis of a large sample the estimate of the variance employed was

$$\frac{1}{n}S(x-\bar{x})^2,$$

which differs from the formula given previously (p. 48) in that we have divided by *n* instead of by (*n*-1). In large samples the difference between these formulæ is small, and that using *n* may claim a theoretical advantage if we wish for an estimate to be used in conjunction with the estimate of the mean from the [p. 54] same sample, as in fitting a frequency curve to the data; otherwise it is best to use (*n*-1). In small samples the difference is still small compared to the probable error, but becomes important if a variance is estimated by averaging estimates from a number of small samples. Thus if a series of experiments are carried out each with six parallels and we have reason to believe that the variation is in all cases due to the operation of analogous causes, we may take the average of such quantities as

$$\frac{1}{n-1}S(x-\bar{x})^2 = \frac{1}{5}S(x-\bar{x})^2$$

to obtain an unbiassed estimate of the variance, whereas we should under-estimate it were we to divide by 6.

14. Test of Departure from Normality

It is sometimes necessary to test whether an observed sample does or does not depart significantly from normality. For this purpose the third, and sometimes the fourth moment, is calculated; from each of these it is possible to calculate a quantity, g, which is zero for a normal distribution, and is distributed normally about zero for large samples; the standard error being calculable from the size of the sample.

The quantity g_1, which is calculated from the third moment, is essentially a measure of asymmetry; it is equal to [plus or minus][sqrt]b_1, of Pearson's notation; g_2(=b_2-3), calculated from the fourth moment, measures a symmetrical type of departure from the normal form, [p. 55] by which the apex and the two tails of the curve are increased at the expense of the intermediate portion, or, when g_2, is negative, the top and tails are depleted and the shoulders filled out, making a relatively flat-topped curve. (See Fig. 6, p. 45·)

Ex. 3. *Use of higher moments to test normality*. -- Departures from normal form, unless very strongly marked, can only be detected in large samples; we give an example (Table 3) of the calculation for 65 values of the yearly rainfall at Rothamsted; the process of calculation is similar to that of finding the mean and standard deviation, but it is carried two stages further, in the calculation of the 3rd and 4th moments. The formulæ by which the two corrections are applied to the moments are gathered in an appendix, p. 74· For the moments we obtain

$$\mu_2 = 17{\cdot}89, \quad \mu_3 = +37{\cdot}33, \quad \mu_4 = 867{\cdot}17,$$

whence are calculated

$$\gamma_1 = \mu_3/\mu_2^{3/2} = +{\cdot}493, \quad \gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 = -{\cdot}290.$$

For samples from a normal distribution the standard errors of g_1 and g_2 are [sqrt]6/n and [sqrt]24/n, of which the numerical values are given. It will be seen that g_1, exceeds its standard error, but g_2, is quite insignificant; since g_1, is positive it appears that there may be some asymmetry of the distribution in the sense that moderately dry and very wet years are respectively more frequent than moderately wet and very dry years. [p. 56]

15. Discontinuous Distributions

Frequently a variable is not able to take all possible values, but is confined to a particular series of values, such as the whole numbers. This is obvious when the variable is a frequency, obtained by counting, such as the number of cells on a square of a hæmocytometer, [p. 57] or the number of colonies on a plate of culture medium. The normal distribution is the most important of the continuous distributions; but among discontinuous distributions the Poisson Series is of the first importance. If a number can take the values 0, 1, 2, . . ., *x*, . . ., and the frequency with which the values occur are given by the series

$$e^{-m}\left(1,\ m,\ \frac{m^2}{2!},\ \ldots,\ \frac{m^x}{x!},\ \ldots\right)$$

(where *x*! stands for "factorial *x*" =*x*(*x*-1)(*x*-2) ... 1), then the number is distributed in the Poisson Series. Whereas the normal curve has two unknown parameters, *m* and s, the Poisson Series has only one. This value may be estimated from a series of observations, by taking their mean, the mean being a statistic as appropriate to the Poisson Series as it is to the normal curve. It may be shown theoretically that if the probability of an event is exceedingly small, but a sufficiently large number of independent cases are taken to obtain a number of occurrences, then this number will be distributed in the Poisson Series. For example, the chance of a man being killed by horse-kick on any one day is exceedingly small, but if an army corps of men are exposed to this risk for a year, a certain number of them will often be killed in this way. The following data (Bortkewitch's data) were obtained from the records of ten army corps for twenty years: [p. 58]

TABLE 4

Deaths.	Frequency observed.	Expected.
0	109	108·67
1	65	66·29
2	22	20·22
3	3	4·11
4	1	·63
5	..	·08
6	..	·01

The average, *m*, is .61, and using this value the numbers calculated agree excellently with those observed.

The importance of the Poisson Series in biological research was first brought out in connexion with the accuracy of counting with a hæmocytometer. It was shown that when the technique of the counting process was effectively perfect, the number of cells on each square should be theoretically distributed in a Poisson Series; it was further shown that this distribution was, in favourable circumstances, actually realised in practice. Thus the table on page 59 (Student's data) shows the distribution of yeast cells in the 400 squares into which one square millimetre was divided.

The total number of cells counted is 1872, and the mean number is therefore 4.68. The expected frequencies calculated from this mean agree well with those observed. The methods of resting the agreement are explained in

When a number is the sum of several components each of which is independently distributed in a Poisson [p. 59] Series, then the total number is also so distributed. Thus the total count of 1872 cells may be regarded as a single sample of a series, for which *m* is not far from 1872. For such large values of *m* the distribution of numbers approximates closely to the normal form, in such a way that the variance is equal to *m*; we may therefore attach to the number counted, 1872, the standard error [plus of minus][sqrt]1872 = [plus or minus]43.26, to represent the standard error of random sampling of such a count. The density of cells in the original suspension is therefore estimated with a standard error of 2.31 per cent. If, for instance, a second sample differed by 7 per cent, the technique of sampling would be suspect. [p. 60]

16. Small Samples of a Poisson Series

Exactly the same principles as govern the accuracy of a hæmocytometer count would also govern a count of bacterial or fungal colonies in estimating the numbers of those organisms by the dilution method, if it could be assumed that the technique of dilution afforded a perfectly random distribution of organisms, and that these could develop on the plate without mutual interference. Agreement of the observations with the Poisson distribution thus affords in the dilution method of counting a test of the suitability of the technique and medium similar to the test afforded of the technique of hæmocytometer counts. The great practical difference between these cases is that from the hæmocytometer we can obtain a record of a large number of squares with only a few organisms on each, whereas in a bacterial count we may have only 5 parallel plates, bearing perhaps 200 colonies apiece. From a single sample of 5 it would be impossible to demonstrate that the distribution followed the Poisson Series; however, when a large number of such samples have been obtained under comparable conditions, it is possible to utilise the fact that for all Poisson Series the variance is numerically equal to the mean.

For each set of parallel plates with $x_1, x_2, \ldots, x_n$, colonies respectively, taking the mean x[bar], an index of dispersion may be calculated by the formula

$$\chi^2 = \frac{S(x-\bar{x})^2}{\bar{x}}.$$

[p. 61]

It has been shown that for true samples of a Poisson Series, χ^2 calculated in this way will be distributed in a known manner; Table III. (p. 98) shows the principal values of χ^2 for this distribution; entering the table take n equal to one less than the number of parallel plates. For small samples the permissible range of variation of χ^2 is wide; thus for five plates with n=4, χ^2 will be less than 1.064 in 10 per cent of cases, while the highest 10 per cent will exceed 7.779; a single sample of 5 thus gives us little information; but if we have 50 or 100 such samples, we are in a position to verify with accuracy if the expected distribution is obtained.

Ex. 4· *Test of agreement with a Poisson Series of a number of small samples*. -- From 100 counts of bacteria in sugar refinery products the following values were obtained (Table 6); there being 6 plates in each case, the values of χ^2 were taken from the χ^2 table for n =5.

It is evident that the observed series differs strongly from expectation; there is an enormous excess in the first class, and in the high values over 15; the relatively few values from 2 to 15 are not far from the expected proportions, as is shown in the last column by taking 43 per cent of the expected values. It is possible then that even in this case nearly half of the samples were satisfactory, but about 10 per cent were excessively variable, and in about 45 per cent of the cases the variability was abnormally depressed.

It is often desirable to test if the variability is of the right magnitude when we have not accumulated [p. 62] a large number of counts, all with the same number of parallel plates, but where a certain number of counts are available with various numbers of parallels. In this case we cannot indeed verify the theoretical distribution with any exactitude, but can test whether [p. 63] or not the general level of variability conforms with expectation. The sum of a number of independent values of χ^2 is itself distributed in the manner shown in the table of χ^2, provided we take for *n* the number S(*n*), calculated by adding the several values of *n* for the separate experiments. Thus for six sets of 4 plates each the total value of χ^2 was found to be 1385, the corresponding value of *n* is 6x3=18, and the χ^2 table shows that for *n*=18 the value 13.85 is exceeded in between 70 and 80 per cent of cases ; it is therefore not an abnormal value to obtain. In another case the following values were obtained:

TABLE 7

Number of Plates in Set.	Number of Sets.	S(*n*).	Total χ^2.
4	8	24	27·31
5	36	144	133·96
9	1	8	8·73
Total	..	176	170·00

We have therefore to test if χ^2=170 is an unreasonably small or great value for *n*=176· The χ^2 table has not been calculated beyond *n*=30, but for higher values we make use of the fact that the distribution of χ^2 becomes nearly normal. A good approximation is given by assuming that ([sqrt]2χ^2 - [sqrt]2*n*-1 is normally distributed about zero with unit standard deviation. If this quantity is materially greater than 2, the value of χ^2 is not in accordance with expectation. In the example before us [p. 64]

$$2\chi^2 = 340, \qquad \sqrt{2\chi^2} = 18\cdot44$$
$$2n - 1 = 351, \qquad \sqrt{2n-1} = 18\cdot73$$
$$\text{Difference} = -\cdot29$$

The set of 45 counts thus shows variability between parallel plates, very close to that to be expected theoretically. The internal evidence thus suggests that the technique was satisfactory.

17. Presence and Absence of Organisms in Samples

When the conditions of sampling justify the use of the Poisson Series, the number of samples containing 0, 1, 2, ... organisms is, as we have seen, connected by a calculable relation with the mean number of organisms in the sample. With motile organisms, or in other cases which do not allow of discrete colony formation, the mean number of organisms in the sample may be inferred from the proportion of fertile cultures, provided a single organism is capable of developing. If m is the mean number of organisms in the sample, the proportion of samples containing none, that is the proportion of sterile samples, is e^{-m}, from which relation we can calculate, as in the following table, the mean number of organisms corresponding to 10 per cent, 20 per cent, etc., fertile samples:

TABLE 8

Percentage of fertile samples	10	20	30	40	50	60	70	80	90
Mean number of organisms	·1054	·2232	·3567	·5108	·6932	·9163	1·2040	1·6095	2·3026

In connexion with the use of the above table it is worth noting that for a given number of samples [p. 65] tested the percentage is most accurately determined at 50 per cent, but for the minimum percentage error in the estimate of the number of organisms, nearly 60 per cent or 88 organisms per sample is most accurate. The Poisson Series also enables us to calculate what percentage of the fertile cultures obtained have been derived from a single organism, for the percentage of impure cultures, *i.e.* those derived from 2 or more organisms, can be calculated from the percentage of cultures which proved to be fertile. If e^{-m} are sterile, me^{-m} will be pure cultures, and the remainder impure. The following table gives representative values of the percentage of cultures which are fertile, and the percentage of fertile cultures which are impure:

TABLE 9

Mean number of organisms in sample	·1	·2	·3	·4	·5	·6	·7
Percentage fertile	9·52	18·13	25·92	32·97	39·35	45·12	50·34
Percentage of fertile cultures impure	4·92	9·67	14·25	18·67	22·92	27·02	30·95

If it is desired that the cultures should be pure with high probability, a sufficiently low concentration must be used to render at least nine-tenths of the samples sterile.

18. The Binomial Distribution

The binomial distribution is well known as the first example of a theoretical distribution to be established. It was found by Bernoulli, about the beginning of the eighteenth century, that if the probability of an event occurring were p and the probability of it not occurring were $q(=1-p)$, then if a random sample of n trials [p. 66] were taken, the frequencies with which the event occurred 0, 1, 2,..., n times were given by the expansion of the binomial

$$(q+p)^n.$$

This rule is a particular case of a more general theorem dealing with cases in which not only a simple alternative is considered, but in which the event may happen in s ways with probabilities p_1, p_2 ..., p_s; then it can be shown that the chance of a random sample of n giving a_1, of the first kind, a_2, of the second, ..., a_s of the last is

$$\frac{n!}{a_1!a_2! \ldots a_s!} p_1^{a_1} p_2^{a_2} \ldots p_s^{a_s},$$

which is the general term in the multinomial expansion of

$$(p_1+p_2+\ldots+p_s)^n.$$

Ex. 5· *Binomial distribution given by dice records.* -- In throwing a true die the chance of scoring more than 4 is 1/3, and if 12 dice are thrown together the number of dice scoring 5 or 6 should be distributed with frequencies given by the terms in the expansion of

$$(2/3 + 1/3)^{12}$$

If, however, one or more of the dice were not true, but if all retained the same bias throughout the experiment, the frequencies should be given approximately by

$$(q+p)^{12},$$

where p is a fraction to be determined from the data. [p. 67] The following frequencies were observed (Weldon's data) in an experiment of 26,306 throws.

TABLE 10

Number of Dice with 5 or 6.	Observed Frequency.	Expected True Dice.	Expected Biassed Dice.	Measure of Divergence $\frac{x^2}{m}$.	
				True Dice.	Biassed Dice.
0	185	202·75	187·38	1·554	·030
1	1149	1216·50	1146·51	3·745	·005
2	3265	3345·37	3215·24	1·931	·770
3	5475	5575·61	5464·70	1·815	·019
4	6114	6272·56	6269·35	4·008	3·849
5	5194	5018·05	5114·65	6·169	1·231
6	3067	2927·20	3042·54	6·677	·197
7	1331	1254·51	1329·73	4·664	·001
8	403	392·04	423·76	·306	1·017
9	105	87·12	96·03	3·670	·838
10	14	13·07	14·69	·066	·032
11	4	1·19	1·36	6·143	4·688
12	..	·05	·06	..	..
	26306	26306·02	26306·00	40·748	12·677
				$n=11$	$n=10$

It is apparent that the observations are not compatible with the assumption that the dice were. unbiassed. With true dice we should expect more cases than have been observed of 0, 1, 2, 3, 4, and less cases than have been observed of 5, 6, ..., 11 dice scoring more than four. The same conclusion is more clearly brought out in the fifth column, which shows the values of the measure of divergence

$$\frac{x^2}{m},$$

where *m* is the expected value and *x* the difference [p. 68] between the expected and observed values. The aggregate of these values is χ^2, which measures the deviation of the whole series from the expected series of frequencies, and the actual chance in this case of χ^2 exceeding 40.75 if the dice had been true is .00003.

The total number of times in which a die showed 5 or 6 was 106,602, out of 315,672 trials, whereas the expected number with true dice is 105,224; from the former number, the value of *p* can be calculated, and proves to be .337,698,6, and hence the expectations of the fourth column were obtained. These values are much more close to the observed series, and indeed fit them satisfactorily, showing that the conditions of the experiment were really such as to give a binomial series.

The standard deviation of the binomial series is [sqrt]*pqn*.

Thus with true dice and 315,672 trials the expected number of dice scoring more than 4 is 105,224 with standard error 264.9; the observed number exceeds expectation by 2378, or 5.20 times its standard error; this is the most sensitive test of the bias, and it may be legitimately applied, since for such large samples the binomial distribution closely approaches the normal. From the table of the probability integral it appears that a normal deviation only exceeds 5.2 times its standard error once in 5 million times.

The reason why this last test gives so much higher odds than the test for goodness of fit, is that the latter is testing for discrepancies of any kind, such, for example, as copying errors would introduce. The actual discrepancy is almost wholly due to a single item, namely, the value of p, and when that point [p. 69] is tested separately its significance is more clearly brought out.

Ex. 6. *Comparison of sex ratio in human families with the binomial distribution.* -- Biological data are rarely so extensive as this experiment with dice; Geissler's data on the sex ratio in German families will serve as an example. It is well known that male births are slightly more numerous than female births, so that if a family of 8 is regarded as a random sample of eight from the general population, the number of boys in such families should be distributed in the binomial

$$(q+p)^8,$$

where p is the proportion of boys. If, however, families differ not only by chance, but by a tendency on the part of some parents to produce males or females, then the distribution of the number of boys should show an excess of unequally divided families, and a deficiency of equally or nearly equally divided families. The data in Table 11 show that there is evidently such an excess of very unequally divided families.

The observed series differs from expectation markedly in two respects: one is the excess of unequally divided families; the other is the irregularity of the central values, showing an apparent bias in favour of even values. No biological reason is suggested for the latter discrepancy, which therefore detracts from the value of the data. The excess of the extreme types of family may be treated in more detail by [p. 70] comparing the observed with the expected variance. The expected variance, *npq*, calculated from the data is 1.998,28, while that calculated from the data is 2.067,42, showing an excess of .06914, or 3.46 per cent. The standard error of the variance is

$$\sqrt{\frac{\mu_4 - \mu_2^2}{N}},$$

where N is the number of samples, and m_2 and m_4, are the second and fourth moments of the theoretical distribution, namely,

$$\mu_2 = npq,$$
$$\mu_4 = 3n^2p^2q^2 + npq(1 - 6pq),$$

so that

$$\mu_4 - \mu_2^2 = 2n^2p^2q^2 + npq(1 - 6pq).$$

The approximate values of these two terms are 8 and -1 giving +7, the actual value being 6.98966. Hence the standard error of the variance is .01141; the discrepancy is over six times its standard error. [p. 71]

One possible cause of the excessive variation lies in the occurrence of multiple births, for it is known that children of the same birth tend to be of the same sex. The multiple births are not separated in these data, but an idea of the magnitude of this effect may be obtained from other data for the German Empire. These show about 12 twin births per thousand, of which 5/8 are of like sex and 3/8 of unlike, so that one-quarter of the twin births, 3 per thousand, may be regarded as "identical" in respect of sex. Six children per thousand would therefore probably belong to such "identical" twin births, the additional effect of triplets, etc., being small. Now with a population of identical twins it is easy to see that the theoretical variance is doubled; consequently, to raise the variance by 3.46 per cent we require that 3.46 per cent of the children should be "identical" twins; this is more than five times the general average, and although it is probable that the proportion of twins is higher in families of 8 than in the general population, we cannot reasonably ascribe more than a fraction of the excess variance to multiple births.

19. Small Samples of the Binomial Series

With small samples, such as ordinarily occur in experimental work, agreement with the binomial series cannot be tested with such precision from a single sample. It is, however, possible to verify that the variation is approximately what it should be, by calculating an index of dispersion similar to that used for the Poisson Series. [p. 72]

Ex. 7· *The accuracy of estimates of infestation.* -- The proportion of barley ears infected with goutfly may be ascertained by examining 100 ears, and counting the infected specimens; if this is done repeatedly, the numbers obtained, if the material is homogeneous, should be distributed in the binomial

$$(q+p)^{100},$$

where p is the proportion infested, and q the proportion free from infestation. The following are the data from 10 such observations made on the same plot (J. G. H. Frew's data):

16, 18, 11, 18, 21, 10, 20, 18, 17, 21. Mean 17.0·

Is the variability of these numbers ascribable to random sampling; *i.e.* Is the material apparently homogeneous? Such data differs from that to which the Poisson Series is appropriate, in that a fixed total of 100 is in each case divided into two classes, infected and not infected, so that in taking the variability of the infected series we are equally testing the variability of the series of numbers not infected. The modified form of χ^2, the index of dispersion, appropriate to the binomial is

$$\chi^2 = \frac{S(x-\bar{x})^2}{n\bar{p}\bar{q}} = \frac{S(x-\bar{x})^2}{\bar{x}\bar{q}},$$

differing from the form appropriate to the Poisson Series in containing the divisor *q*[bar], or in this case, .83.· The value of χ^2 is 9.22, which, as the χ^2, table shows, is a perfectly reasonable value for *n*=9, one less than the number of values available. [p. 73]

Such a test of the single sample is, of course, far from conclusive, since χ^2 may vary within wide limits. If, however, a number of such small samples are available, though drawn from plots of very different infestation, we can test, as with the Poisson Series, if the general trend of variability accords with the binomial distribution. Thus from 20 such plots the total χ^2 is 193.64, while S(*n*) is 180. Testing as before (p. 63), we find

$$\begin{aligned} \sqrt{387{\cdot}28} &= 19{\cdot}68 \\ \sqrt{359} &= 18{\cdot}95 \\ \text{Difference} \quad & +{\cdot}73 \end{aligned}$$

The difference being less than one, we conclude that the variance shows no sign of departure from that of the binomial distribution. The difference between the method appropriate for this case, in which the samples are small (10), but each value is derived from a considerable number (100) of observations, and that appropriate for the sex distribution in families of 8, where we had many families, each of only 8 observations, lies in the omission of the term

$$npq(1\text{-}6pq)$$

in calculating the standard error of the variance. When n is 100 this term is very small compared to $2n^2p^2q^2$, and in general the χ^2 method is highly accurate if the number in all the observational categories is as high as 10. [p. 74]

APPENDIX OF TECHNICAL NOTATION AND FORMULÆ

A. *Definition of moments of sample.*

The following statistics are known as the first four moments of the variate x; the first moment is the mean

the second and higher moments are the mean values of the second and higher powers of the deviations from the mean

B. *Moments of theoretical distribution in terms of parameters.*

[p. 75]

C. *Variance of moments derived from samples of* N.

D. *Corrections in calculating moments.*

(*a*) Correction for mean, if v' is the moment about the working mean, and v the corresponding value corrected to the true mean:

$$v_2 = v'_2 - v'^2_1,$$

$$v_3 = v'_3 - 3v'_1 v'_2 + 2v'^3_1,$$

$$v_4 = v'_4 - 4v'_1 v'_3 + 6v'^2_1 v'_2 - 3v'^4_1.$$

(*b*) Correction for grouping, if v is the estimate uncorrected for grouping, and m the corresponding estimate corrected:

$$m_1 = v_2 - 1/12,$$

$$m_2 = v_3,$$

$$m_3 = v4 - 1/2m_2 - 1/80.$$ [p. 76]

TABLE I

Table of x

The deviation in the normal distribution in terms of the standard deviation.

	·01.	·02.	·03.	·04.	·05.	·06.	·07.	·08.	·09.	·10.
·00	2·575829	2·326348	2·170090	2·053749	1·959964	1·880794	1·811911	1·750686	1·695398	1·644854
·10	1·598193	1·554774	1·514102	1·475791	1·439521	1·405072	1·372204	1·340755	1·310579	1·281552
·20	1·253565	1·226528	1·200359	1·174987	1·150349	1·126391	1·103063	1·080319	1·058122	1·036433
·30	1·015222	·994458	·974114	·954165	·934589	·915365	·896473	·877896	·859617	·841621
·40	·823894	·806421	·789192	·772193	·755415	·738847	·722479	·706303	·690309	·674490
·50	·658838	·643345	·628006	·612813	·597760	·582841	·568051	·553385	·538836	·524401
·60	·510073	·495850	·481727	·467699	·453762	·439913	·426148	·412463	·398855	·385320
·70	·371856	·358459	·345125	·331853	·318639	·305481	·292375	·279319	·266311	·253347
·80	·240426	·227545	·214702	·201893	·189118	·176374	·163658	·150969	·138304	·125661
·90	·113039	·100434	·087845	·075270	·062707	·050154	·037608	·025069	·012533	0

The value of P for each entry is found by adding the column heading to the value in the left-hand margin. The corresponding value of x is the deviation such that the probability of an observation falling outside the range from $-x$ to $+x$ is P. For example, P = ·03 for $x = 2{\cdot}170090$; so that 3 per cent of normally distributed values will have positive or negative deviations exceeding the standard deviation in the ratio 2·170090.

TABLE II

Values of x for Small Values of P

P . .	·001	·000,1	·000,01	·000,001	·000,000,1	·000,000,01	·000,000,001
x . .	3·29053	3·89059	4·41717	4·89164	5·32672	5·73073	6·10941

Classics in the History of Psychology

An internet resource developed by
Christopher D. Green
York University, Toronto, Ontario
ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted April 2000

IV

TESTS OF GOODNESS OF FIT, INDEPENDENCE AND HOMOGENEITY; WITH TABLE OF χ^2

20. The χ^2 Distribution

In the last chapter some use has been made of the χ^2 distribution as a means of testing the agreement between observation and hypothesis; in the present chapter we shall deal more generally with the very wide class of problems which may be solved by means of the same distribution.

The common factor underlying all such tests is the comparison of the numbers actually observed to fall into any number of classes with the numbers which upon some hypothesis are expected. If *m* is the number expected, and *m+x* the number observed in any class, we calculate

$$\chi^2 = S\left(\frac{x^2}{m}\right),$$

the summation extending over all the classes. This formula gives the value of χ^2, and it is clear that the more closely the observed numbers agree with those expected the smaller will χ^2 be; in order to utilise the table it is necessary to know also the value of *n* with which the table is to be entered. The rule for finding [p. 78] *n* is that *n* is equal to the number-of degrees of freedom in which the observed series may differ from the hypothetical; in other words, it is equal to the number of classes the frequencies in which may be filled up arbitrarily. Several examples will be given to illustrate this rule.

For any value of *n,* which must be a whole number, the form of distribution of χ^2 was established by Pearson in 1900; it is therefore possible to calculate in what proportion of cases any value of χ^2 will be exceeded. This proportion is represented by P, which is therefore the probability that χ^2 shall exceed any specified value. To every value of χ^2 there thus corresponds a certain value of P; as χ^2 is increased from o to infinity, P diminishes from 1 to 0. Equally, to any value of P in this range there corresponds a certain value of χ^2. Algebraically the relation between these two quantities is a complex one, so that it is necessary to have a table of corresponding values, if the χ^2 test is to be available for practical use.

An important table of this sort was prepared by Elderton, and is known as Elderton's Table of Goodness of Fit. Elderton gave the values of P to six decimal places corresponding to each integral value of χ^2 from 1 to 30, and thence by tens to 70. In place of n, the quantity n' ($=n+1$) was used, since it was then believed that this could be equated to the number of frequency classes. Values of n' from 3 to 30 were given, these corresponding to values of n from 2 to 29. A table for $n'=2$, or $n=1$, was subsequently supplied by Yule. Owing to copyright restrictions [p. 79] we have not reprinted Elderton's table, but have given a new table (Table III. p. 98) in a form which experience has shown to be more convenient. Instead of giving the values of P corresponding to an arbitrary series of values of χ^2, we have given the values of χ^2 corresponding to specially selected values of P. We have thus been able in a compact form to cover those parts of the distributions which have hitherto not been available, namely, the values of χ^2 less than unity, which frequently occur for small values of n, and the values exceeding 30, which for larger values of n become of importance.

It is of interest to note that the measure of dispersion, ϕ^2, introduced by the German economist, Lexis, is, if accurately calculated, equivalent to χ^2/n of our notation. In the many references in English to the method of Lexis, it has not, I believe, been noted that the discovery of the distribution of χ^2 in reality completed the method of Lexis. If it were desired to use Lexis' notation, our table could be transformed into a table of ϕ^2 merely by dividing each entry by *n*.

In preparing this table we have borne in mind that in practice we do not want to know the exact value of P for any observed χ^2, but, in the first place, whether or not the observed value is open to suspicion. If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05, and consider that higher values of χ^2 indicate a real discrepancy. [p. 80]

To compare values of χ^2, or of P, by means of a "probable error" is merely to substitute an inexact (normal) distribution for the exact distribution given by the χ^2 table.

The term Goodness of Fit has caused some to fall into the fallacy of believing that the higher the value of P the more satisfactorily is the hypothesis verified. Values over .999 have sometimes been reported which, if the hypothesis were true, would only occur once in a thousand trials. Generally such cases have proved to be due to the use of inaccurate formulæ, but occasionally small values of χ^2 beyond the expected range do occur, as in Ex. 4 with the colony numbers obtained in the plating method of bacterial counting. In these cases the hypothesis considered is as definitely disproved as if P had been .001.

When a large number of values of χ^2 are available for testing, it may be possible to reveal discrepancies which are too small to show up in a single value ; we may then compare the observed distribution of χ^2 with that expected. This may be done immediately by simply distributing the observed values of χ^2 among the classes bounded by values given in the χ^2 table, as in Ex. 4, p. 61. The expected frequencies in these classes are easily written down, and, if necessary, the χ^2 test may be used to test the agreement of the observed with the expected frequencies.

It is useful to remember that the sum of any number of quantities, χ^2, is distributed in the χ^2 distribution, with *n* equal to the sum of the values of *n* corresponding to the values of χ^2 used. Such a test is sensitive, [p. 81] and will often bring to light discrepancies which are hidden or appear obscurely in the separate values.

The table we give has values of *n* up to 30; beyond this point it will be found sufficient to assume that [sqrt]2χ^2 is distributed normally with unit standard deviation about a mean [sqrt]2n-1, The values of P obtained by applying this rule to the values of χ^2 given for *n*=30, may be.

worked out as an exercise. The errors are small for n=30, and become progressively smaller for higher values of n.

Ex. 8. *Comparison with expectation of Mendelian class frequencies.* -- In a cross involving two Mendelian factors we expect by interbreeding the hybrid (F_1) generation to obtain four classes in the ratio 9:3:3:1; the hypothesis in this case is that the two factors segregate independently, and that the four classes of offspring are equally viable. Are the following observations on *Primula* (de Winton and Bateson) in accordance with this hypothesis?

TABLE 12

	Flat Leaves.		Crimped Leaves.		Total.
	Normal Eye.	Primrose Queen Eye.	Lee's Eye.	Primrose Queen Eye.	
Observed $(m+x)$	328	122	77	33	560
Expected (m)	315	105	105	35	560
x^2/m	·537	2·752	7·467	·114	10·870

The expected values are calculated from the observed total, so that the four classes must agree in their sum, and if three classes are filled in arbitrarily the fourth is therefore determinate, hence n=3, [p. 82] χ^2=10.87, the chance of exceeding which value is between .01 and .02; if we take P=.05 as the limit of significant deviation, we shall say that in this case the deviations from expectation are clearly significant.

Let us consider a second hypothesis in relation to the same data, differing from the first in that we suppose that the plants with crimped leaves are to some extent less viable than those with flat leaves. Such a hypothesis could of course be tested by means of additional data; we are only here concerned with the question whether or no it accords with the values before us. The hypothesis tells us nothing of what degree of relative viability to expect; we therefore take the totals of flat and crimped leaves observed, and divide each class in the ratio 3:1.

TABLE 13

	Flat Leaves.		Crimped Leaves.		χ^2.
	Normal Eye.	Primrose Queen Eye.	Lee's Eye.	Primrose Queen Eye.	
Observed	328	122	77	33	..
Expected	337·5	112·5	82·5	27·5	..
x^2/m	·267	·804	·367	1·109	2·547

The value of *n* is now 2, since only two entries can be made arbitrarily; the value of χ^2, however, is so much reduced that P exceeds .2, and the departure from expectation is no longer significant. The significant part of the original discrepancy lay in the proportion of flat to crimped leaves.

It was formerly believed that in entering the χ^2 [p. 83] table *n* was always to be equated to one less than the number of frequency classes; this view led to many discrepancies, and has since been disproved with the establishment of the rule stated above. On the old view, any complication of the hypothesis such as that which in the above instance admitted differential viability, was bound to give us apparent improvement in the agreement between observation and hypothesis. When the change in *n* is allowed for this bias disappears, and if the value of P, rightly calculated; is many fold increased, as in this instance, the increase may safely be ascribed to an improvement in the hypothesis, and not to a mere increase of available constants.

Ex. 9. *Comparison with expectation of the Poisson Series and Binomial Series*. -- In Table 5, p. 59, we give the observed and expected frequencies in the case of a Poisson Series. In applying the χ^2 test to such a series it is desirable that the number expected should in no group be less than 5, since the calculated distribution of χ^2 is not very closely realised for very small classes. We therefore pool the numbers for 0 and 1 cells, and also those for 10 and more, and obtain the following comparison:

TABLE 14

	0 and 1	2	3	4	5	6	7	8	9	10 and more	Total.
Observed	20	43	53	86	70	54	37	18	10	9	400
Expected	21·08	40·65	63·41	74·19	69·44	54·16	36·21	21·18	11·02	8·66	400
x^2/m	·055	·136	1·709	1·880	·005	·005	·017	·477	·093	·013	4·390

[p.84]

Using 10 frequency classes we have χ^2=4.390; in ascertaining the value of n we have to remember that the expected frequencies have been calculated, not only from the total number of values observed (400), but also from the observed mean; there remain, therefore, 8 degrees of freedom and n=8. For this value the χ^2 table shows that P is between .8 and .9, showing a close but not an unreasonably close, agreement with expectation.

Similarly in Table 10, p. 67, we have given the value of χ^2 based upon 12 classes for the two hypotheses of "true dice" and "biassed dice"; with "true dice" the expected values are calculated from the total number of observations alone, and n=11, but in allowing for bias we have brought also the means into agreement so that n is reduced to 10. In the first case χ^2 is far outside the range of the table showing a highly significant departure from expectation; in the second it appears that P lies between .2 and .3, so that the value of χ^2 is within the expected range.

21. Tests of Independence, Contingency Tables

A special and important class of cases where the agreement between expectation and observation may be tested comprises the tests of **independence**. If the same group of individuals is classified in two (or more) different ways, as persons may be classified as inoculated and not inoculated, and also as attacked and not attacked by a disease, then we may require to know if the two classifications are independent. [p. 85]

Ex. 10: *Test of independence in a 2x2 classification*. -- In the simplest case, when each classification comprises only two classes, we have a fourfold table, as in the following example (from Greenwood and Yule's data) for Typhoid:

TABLE 15

OBSERVED

	Attacked.	Not Attacked.	Total.
Inoculated . .	56	6,759	6,815
Not Inoculated .	272	11,396	11,668
Total .	328	18,155	18,483

TABLE 16

EXPECTED

	Attacked.	Not Attacked.	Total.
Inoculated . .	120·93	6,694·07	6,815
Not Inoculated .	207·07	11,460·93	11,668
Total .	328	18,155	18,483

In testing independence we must compare the observed values with values calculated so that the four frequencies are *in proportion*; since we wish to test independence only, and not any hypothesis as to the total numbers attacked, or inoculated, the "expected" values are calculated from the marginal totals observed, so that the numbers expected agree with the numbers [p. 86] observed in the margins; only one value need be calculated, *e.g.*

$$\frac{328 \times 6815}{18483} = 120{\cdot}93;$$

the others are written down at once by subtraction from the margins. It is thus obvious that the observed values can differ from those expected in only 1 degree of freedom, so that in testing independence in a four; fold table, n =1. Since χ^2=56.234 the observations are clearly opposed to the hypothesis of independence. Without calculating the expected values, χ^2 may, for fourfold tables, be directly calculated by the formula

$$\chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)},$$

where *a, b, c,* and *d* are the four observed numbers.

When only one of the classifications is of two classes, the calculation of χ^2 may be simplified to some extent, if it is not desired to calculate the expected numbers. If *a, a'* represent any pair of observed frequencies, and *n, n'* the corresponding totals, we calculate from each pair

$$\frac{1}{a+a'}(an'-a'n)^2,$$

and the sum of these quantities divided by *nn'* will be χ^2.

Ex. 11. *Test of independence in a 2xn classification*. --

From the pigmentation survey of Scottish children (Tocher's data) the following are the numbers of boys and girls from the same district (No. 1) whose hair colour falls into each of five classes: [p. 87]

TABLE 17

HAIR COLOUR

	Fair.	Red.	Medium.	Dark.	Jet Black.	Total.
Boys . . .	592	119	849	504	36	2100
Girls . . .	544	97	677	451	14	1783
Total .	1136	216	1526	955	50	3883
	6·642	·333	5·555	2·460	24·204	39·194

The quantities calculated from each pair of observations are given below in millions. Thus

$$\frac{1}{1136}(544 \times 2100 - 592 \times 1783)^2 = 6{,}642{,}000$$

approximately; the total of 39 millions odd divided by 2100 and by 1783 gives χ^2=10.468· In this table 4 values could be filled in arbitrarily without conflicting with the marginal totals, so that *n*=4. The value of P is between .02 and .05, so that sex difference in the classification by hair colours is probably significant as judged by this district alone. The calculation of χ^2 from "expected" values, though somewhat more laborious, would have in this case the advantage of showing in which classes the boys, and in which classes the girls, were in excess. It appears from the numbers in the lowest line that the principle discrepancy is in the "Jet Black" class.

Ex. 12. *Test of independence in a* 4 x4 *classification*. -- As an example of a more complex contingency table we may take the results of a series of back-crosses [p. 88] in mice, involving the two Brown, Self-Piebald (Wachter's data):

TABLE 18

	Black Self.	Black Piebald.	Brown Self.	Brown Piebald.	Total.
Coupling—					
F_1 Males .	88 (85·37)	82 (75·24)	75 (70·93)	60 (73·46)	305
F_1 Females	38 (34·43)	34 (30·34)	30 (28·60)	21 (29·63)	123
Repulsion—					
F_1 Males .	115 (117·00)	93 (103·11)	80 (97·21)	130 (100·68)	418
F_1 Females	96 (100·20)	88 (88·31)	95 (83·26)	79 (86·23)	358
Total .	337	297	280	290	1204

The back-crosses were made in four ways, according as the male or female parents were heterozygous (F_1) in the two factors, and according to whether the two dominant genes were received both from one (Coupling) or one from each parent (Repulsion).

The simple Mendelian ratios may be disturbed by differential viability, by linkage, or by linked lethals. Linkage is not suspected in these data, and if the only disturbance were due to differential viability the four classes in each experiment should appear in the same ratio; to test if the data show significant departures we may apply the χ^2 test to the whole 4x4 table. The values expected on the hypothesis that the proportions are independent of the matings used, or that the four series are homogeneous, are given above in brackets. The contributions to χ^2 made by each cell are given on page 89.

The value of χ^2 is therefore 21.832; the value of *n* is 9, for we could fill up a block of three rows and [p. 89]

TABLE 19

·081	·607	·234	2·466	3·388
·370	·442	·069	2·514	3·395
·034	·991	3·047	8·539	12·611
·176	·001	1·655	·606	2·438
·661	2·041	5·005	14·125	21·832

three columns and still adjust the remaining entries to check with the margins. In general for a contingency table of *r* rows and *c* columns $n=(r-1)(c-1)$. For $n=9$, the value of χ^2 shows that P is less than .01, and therefore the departures from proportionality are not fortuitous; it is apparent that the discrepancy is due to the exceptional number of Brown Piebalds in the F_1 males repulsion series.

It should be noted that the methods employed in this chapter are not designed to measure the *degree* of association between one classification and another, but solely to test whether the observed departures from independence are or are not of a magnitude ascribable to chance. The same degree of variation may be significant for a large sample but insignificant for a small one; if it is insignificant we have no reason on the data present to suspect any degree of association at all, and it is useless to attempt to measure it. If, on the other hand, it is significant the value of χ^2 indicates the fact, but does not measure the degree of association. Provided the deviation is clearly significant, it is of no practical importance whether P is .01 or ·.000,001, and it is for this reason that we have not tabulated the value of χ^2 beyond .01. To measure [p. 90] the degree of association it is necessary to have some hypothesis as to the nature of the departure from independence to be measured. With Mendelian frequencies, for example, the cross-over percentage may be used to measure the degree of association of two factors, and the significance of evidence for linkage may be tested by comparing the difference between the cross-over percentage and 50 per cent (the value for unlinked factors), with its standard error. Such a comparison, if accurately carried out, must agree absolutely with the conclusion drawn from the χ^2 test. To take a second example, the values in a four-fold table may be sometimes regarded as due to the partition of a normally correlated pair of variates, according as the values are above or below arbitrarily chosen dividing-lines ; as if a group of stature measurements of fathers and sons were divided between those above and those below 68 inches. In this case the departure from independence may be properly measured by the correlation in stature between father and son; this quantity can be estimated from the observed frequencies, and a comparison between the value obtained and its standard error, if accurately carried out, will agree with the χ^2 test as to the significance of the association; the significance will become more and more pronounced as the sample is increased in size, but the correlation obtained will tend to a fixed value. The χ^2 test does not attempt to measure the degree of association, but as a test of significance it is independent of all additional hypotheses as to the nature of the association. [p. 91]

Tests of **homogeneity** are mathematically identical with tests of independence; the last example may equally be regarded in either light. In Chapter III. the tests of agreement with the Binomial Series were essentially tests of homogeneity; the ten samples of 100 ears of barley (Ex. 7, p. 72) might have been represented as a 2x10 table. The χ^2 index of dispersion would then be equivalent to the χ^2 obtained from the contingency table. The method of this chapter is more general, and is applicable to cases in which the successive samples are not all of the same size.

Ex. 13· *Homogeneity of different families in respect of ratio black: red*. -- The following data show in 33 families of *Gammarus* (Huxley's data) the numbers with black and red eyes respectively:

TABLE 20

Black	79	120	24	117	62	79	66	45	61	64	208	154	31	158	21	105	28
Red	14	31	6	29	17	20	12	11	14	13	52	45	4	45	4	28	7
Total	93	151	30	146	79	99	78	56	75	77	260	199	35	203	25	133	35
Black	58	81	25	95	47	67	30	70	139	179	129	44	24	19	45	91	2565
Red	19	27	8	29	16	21	11	28	57	62	44	17	9	8	23	41	772
Total	77	108	33	124	63	88	41	98	196	241	173	61	33	27	68	132	3337

The totals 2565 black and 772 red are distinctly not in the ratio 3:1, which is ascribed to linkage. The question before us is whether or not all the families indicate the same ratio between black and red, or whether the discrepancy is due to a few families only. For the whole table χ^2=35.620, n=32. This is [p. 92] beyond the range of the table, so we apply the method explained on p. 63:

$$\sqrt{2\chi^2} = 8{\cdot}44\,;$$
$$\sqrt{2n-1} = 7{\cdot}94\,;$$
$$\text{Difference} = +{\cdot}50 \pm 1.$$

The series is therefore not significantly heterogeneous; effectively all the families agree and confirm each other in indicating the black-red ratio observed in the total.

Exactly the same procedure would be adopted if the black and red numbers represented two samples distributed according to some character or characters each into 33 classes. The question "Are these samples of the same population?" is in effect identical with the question "Is the proportion of black to red the same in each family?" To recognise this identity is important, since it has been very widely disregarded.

Ex. 14· *Agreement with expectation of normal variance*. -- Closely akin to tests of homogeneity is the use of the χ^2 distribution to test whether or not an observed series of values, normally or nearly normally distributed, agrees in its variance with expectation. If x_1, x_2,..., are a sample of a normal population, the standard deviation of which population is s, then

$$\frac{1}{\sigma^2}S(x-\bar{x})^2$$

is distributed in random samples as is χ^2, taking *n* one less than the number of the sample. J. W. Bispham gives three series of experimental values of the partial correlation coefficient, which he assumes should be [p. 93] distributed so that $1/s^2=29$, but which theoretically should have $1/s^2=28$. Th, values of $S(x-x[\text{bar}])^2$ for the three samples of 1000, 200, 100 respectively are, as judged from the grouped data,

35.0278, 7.1071, 3.6169,

whence the values of χ^2 on the two theories are

TABLE 21

	Exp. 1.	2.	3.	Total.	$\sqrt{2\chi^2}$.	Difference.
29 $S(x-\bar{x})^2$.	1015·81	217·71	104·89	1338·41	51·74	+·82
28 $S(x-\bar{x})^2$.	980·78	210·20	101·27	1292·25	50·82	−·10
Expectation (n)	999	199	99	1297	50·92	

It will be seen that the true formula for the variance gives slightly the better agreement. That the difference is not significant may be seen from the last two columns. About 6000 observations would be needed to discriminate experimentally, with any certainty, between the two formulæ.

22. Partition of χ^2 into its Components

Just as values of χ^2 may be aggregated together to make a more comprehensive test, so in some cases it is possible to separate the contributions to χ^2 made by the individual degrees of freedom, and so to test the separate components of a discrepancy.

Ex. I5· *Partition of observed discrepancies from Mendelian expectation*. -- The following table (de Winton and Bateson's data) gives the distribution of sixteen families of primula in the eight classes obtained from a back-cross with the triple recessive: [p. 94]

TABLE 22

Type.	Family Number.																Total.
	54.	55.	58.	59.	107.	110.	119.	121.	122.	127.	129.	131.	132	133.	135.	178.	
Ch G W .	5	18	17	2	12	17	9	10	24	9	3	16	20	9	11	10	192
Ch G w .	10	13	11	12	20	16	10	7	23	3	6	24	18	2	13	12	200
Ch g W .	4	10	17	3	14	10	6	8	19	5	5	23	18	10	7	12	171
Ch g w .	9	17	11	11	13	13	9	8	9	6	3	12	18	1	9	12	161
ch G W .	13	22	20	10	5	5	16	2	30	3	8	21	19	4	9	12	199
ch G w .	14	16	18	9	12	6	14	3	16	5	7	13	14	4	13	10	174
ch g W .	10	11	12	6	7	3	18	2	11	5	4	14	23	4	6	13	149
ch g w .	7	12	16	6	10	8	10	4	23	5	4	22	23	7	8	16	181
Total	72	119	122	59	93	78	92	44	155	41	40	145	153	41	76	97	1427
χ^2 .	9·78	7·86	5·48	13·00	12·55	19·23	10·09	12·36	18·06	4·86	4·80	9·21	3·18	14·22	5·05	2·05	151·78

[p. 95]

The theoretical expectation is that the eight classes should appear in equal numbers, corresponding to the hypothesis that in each factor the allelomorphs occur with equal frequency, and that the three factors are unlinked. This expectation is fairly realised in the totals of the sixteen families, but the individual families are somewhat irregular. The values of χ^2 obtained by comparing each family with expectation are given in the lowest line. These values each correspond to seven degrees of freedom, and it appears that in 5 cases out of 16, P is less than .1, and of these 2 are less than .02. This confirms the impression of irregularity, and the total value of χ^2 (not to be confused with χ^2 derived from the totals), which corresponds to 112 degrees of freedom, is 151.78. Now

$$\sqrt{223} = 14{\cdot}93\ ;$$
$$\sqrt{303{\cdot}56} = 17{\cdot}42\ ;$$
$$\text{Difference} = +2{\cdot}49\ ;$$

so that, judged by the total χ^2, the evidence for departures from expectation in individual families, is clear.

Each family is free to differ from expectation in seven independent ways. To carry the analysis further, we must separate the contribution to χ^2 of each of these seven degrees of freedom. Mathematically the subdivision may be carried out in more than one way, but the only way which appears to be of biological interest is that which separates the parts due to inequality of the allelomorphs of the three factors, and the three possible linkage connexions. If we separate [p. 95] the frequencies into positive and negative values according to the following seven ways,

TABLE 23

	Ch.	G.	W.	G W.	Ch W.	Ch G.	Ch G W.
Ch G W .	+	+	+	+	+	+	+
Ch G w .	+	+	−	−	−	+	−
Ch g W .	+	−	+	−	+	−	−
Ch g w .	+	−	−	+	−	−	+
ch G W .	−	+	+	+	−	−	−
ch G w .	−	+	−	−	+	−	+
ch g W .	−	−	+	−	−	+	+
ch g w .	−	−	−	+	+	+	−

then it will be seen that all seven subdivisions are wholly independent, since any two of them agree in four signs and disagree in four. The first three degrees of freedom represent the inequalities in the allelomorphs of the three factors Ch, G, and W; the next are the degrees of freedom involved in an enquiry into the linkage of the three pairs of factors, while the seventh degree of freedom has no simple biological meaning but is necessary to complete the analysis. If we take in the first family, for example, the difference between the numbers of the W and w plants, namely 8, then the contribution of this degree of freedom to χ^2 is found by squaring the difference and dividing by the number in the family, *e.g.* 82/72=889. In this way the contribution of each of the 112 degrees of freedom in the sixteen families is found separately, as shown in the following table: [p. 97]

TABLE 24

Family.	Ch.	G.	W.	G W.	Ch W.	Ch G.	Ch G W.	Total.
54	3·556	2·000	·889	·222	2·000	·889	·222	9·778
55	·076	3·034	·076	3·034	·412	1·017	·210	7·859
58	·820	·820	·820	·295	1·607	·820	·295	5·477
59	·153	·831	4·898	·017	6·119	·831	·153	13·002
107	6·720	·269	3·108	1·817	·097	·269	·269	12·549
110	14·821	1·282	·821	·821	·205	1·282	0	19·232
119	6·261	·391	·391	·174	2·130	·043	·696	10·086
121	11·000	0	0	·364	·818	·091	·091	12·364
122	·161	6·200	1·090	1·865	·523	·316	7·903	18·058
127	·610	·024	·220	·610	1·195	·220	1·976	4·855
129	·900	1·600	0	·400	·100	·900	·900	4·800
131	·172	·062	·062	·062	·062	·338	8·448	9·206
132	·163	·791	·320	·320	·059	1·471	·059	3·183
133	·220	·220	4·122	·024	8·805	·220	·610	14·221
135	·211	3·368	1·316	·053	·053	0	·053	5·054
178	·258	·835	·093	·093	·010	·258	·505	2·052
Total	46·102	21·727	18·226	10·171	24·195	8·965	22·390	151·776

Looking at the total values of χ^2 for each column, since *n* is 16 for these, we see that all except the first have values of P between .05 and .95, while the contribution of the first degree of freedom is very clearly significant. It appears then that the greater part, if not the whole, of the discrepancy is ascribable to the behaviour of the Sinensis-Stellata factor, and its behaviour strongly suggests close linkage with a recessive lethal gene of one of the familiar types. In four families, 107-121, the only high contribution is in the first column. If these four families are excluded χ^2=97.545, and this exceeds the expectation for *n*=84 by only just over the standard error; the total discrepancy cannot therefore be regarded as significant. There does, however, appear to be an excess of very large entries, and it is noticeable of the seven largest, [p. 98-99]

TABLE III

TABLE OF χ^2

n.	P = ·99.	·98.	·95.	·90.	·80.	·70.	·50.	·30	·20.	·10.	·05.	·02.	·01.
1	·000157	·000628	·00393	·0158	·0642	·148	·455	1·074	1·642	2·706	3·841	5·412	6·635
2	·0201	·0404	·103	·211	·446	·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210
3	·115	·185	·352	·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·341
4	·297	·429	·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277
5	·554	·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086
6	·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892

For larger values of *n*, the expression $\sqrt{2\chi^2}-\sqrt{2n-1}$ may be used as a normal deviate with unit standard error.

[p. 100]

six appear in pairs belonging to the same family. The distribution of the remaining 12 families according to the value of P is as follows:

TABLE 25

P . .	1·0	·9	·8	·7	·5	·3	·2	·1	·05	·02	·01	0	Total
Families .	1	1	0	4	1	2	0	1	1	1	0		12

from which it would appear that there is some slight evidence of an excess of families with high values of χ^2. This effect, like other non-significant effects, is only worth further discussion in connexion with some plausible hypothesis capable of explaining it.

N.B. -- Table of χ^2, p. 98.

Classics in the History of Psychology

An internet resource developed by
Christopher D. Green
York University, Toronto, Ontario
ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted April 2000

V

TESTS OF SIGNIFICANCE OF *MEANS*, DIFFERENCES OF MEANS, AND REGRESSION COEFFICIENTS

23. The Standard Error of the Mean

The fundamental proposition upon which the statistical treatment of mean values is based is that -- If a quantity be normally distributed with standard deviation s, then the mean of a random sample of *n* such quantities is normally distributed with standard deviation s/[sqrt]*n*.

The utility of this proposition is somewhat increased by the fact that even if the original distribution were not exactly normal, that of the mean usually tends to normality, as the size of the sample is increased; the method is therefore applied widely and legitimately to cases in which we have not sufficient evidence to assert that the original distribution was normal, but in which we have reason to think that it does not belong to the exceptional class of distributions for which the distribution of the mean does not tend to normality.

If, therefore, we know the standard deviation of a population, we can calculate the standard deviation of [p. 102] the mean of a random sample of any size, and so test whether or not it differs significantly from any fixed value. If the difference is many times greater than the standard error, it is certainly significant, and it is a convenient convention to take twice the standard error as the limit of significance ; this is roughly equivalent to the corresponding limit P=.05, already used for the c^2 distribution. The deviations in the normal distribution corresponding to a number of values of P are given in the lowest line of the table of *t* at the end of this chapter (p. 137)· More detailed information has been given in Table I.

Ex. 16. *Significance of mean of a large sample.* -- We may consider from this point of view Weldon's die-casting experiment (Ex. 5, p. 66). The variable quantity is the number of dice scoring "5" or "6" in a throw of 12 dice. In the experiment this number varies from zero to eleven, with an observed mean of 4.0524; the expected mean, on the hypothesis that the dice were true, is 4, so that the deviation observed is .0524· If now we estimate the variance of the whole sample of 26,306 values as explained on p. 50, but without using Sheppard's correction (for the data are not grouped), we find

$$s^2 = 2.69825,$$

$$\text{whence } s^2/n = .0001025,$$

$$\text{and } s/\sqrt{n} = .01013.$$

The standard error of the mean is therefore about .01, and the observed deviation is nearly 5.2 times as great; thus by a slightly different path we arrive [p. 103] at the same conclusion as that of p. 68. The difference between the two methods is that our treatment of the mean does not depend upon the hypothesis that the distribution is of the binomial form, but on the other hand we do assume the correctness of the value of s derived from the observations. This assumption breaks down for small samples, and the principal purpose of this chapter is to show how accurate allowance can be made in these tests of significance for the errors in our estimates of the standard deviation.

To return to the cruder theory, we may often, as in the above example, wish to compare the observed mean with the value appropriate to a hypothesis which we wish to test; but equally or more often we wish to compare two experimental values and to test their agreement. In such cases we require the standard error of the difference between two quantities whose standard errors are known; to find this we make use of the proposition that the variance of the difference of two independent variates is equal to the sum of their variances. Thus, if the standard deviations are s_1, s_2, the variances are s_1^2, s_2^2; consequently the variance of the difference is $s_1^2+s_2^2$, and the standard error of the difference is [sqrt]$s_1^2+s_2^2$.

Ex. 17· *Standard error of difference of means from large samples*. -- In Table 2 is given the distribution in stature of a group of men, and also of a group of women; the means are 68.64 and 63.85 inches, giving a difference of 4.79 inches. The variance obtained for the men was 7.2964 square inches; this is the value obtained by dividing the sum of the squares of [p. 104] the deviations by 1164 ; if we had divided by 1163, to make the method comparable to that appropriate to small samples, we should have found 7.3027. Dividing this by 1164, we find the variance of the mean is .006274. Similarly the variance for the women is .63125, which divided by 1456 gives the variance of the mean of the women as .004335. To find the variance of the difference between the means, we must add together these two contributions, and find in all .010609; the standard error of the difference between the means is therefore .1030 inches. The sex difference in stature may therefore be expressed as

4.79 [plus or minus] .103 inches.

It is manifest that this difference is significant, the value found being over 46 times its standard error. In this case we can not only assert a significant difference, but place its value with some confidence at between 4½ and 5 inches. It should be noted that we have treated the two samples as *independent,* as though they had been given by different authorities; as a matter of fact, in many cases brothers and sisters appeared in the two groups; since brothers and sisters tend to be alike in stature, we have overestimated the probable error of our estimate of the sex difference. Whenever possible, advantage should be taken of such facts in designing experiments. In the common phrase, sisters provide a better "control" for their brothers than do unrelated women. The sex difference could therefore be more accurately estimated from the comparison of each brother with his own sister. In [p. 105] the following example (Pearson and Lee's data), taken from a correlation table of stature of brothers and sisters, the material is nearly of this form; it differs from it in that in some instances the same individual has been compared with more than one sister, or brother.

Ex. I8. *Standard error of mean of differences*. -- The following table gives the distribution of the excess in stature of a brother over his sister in 1401 pairs

TABLE 26

Stature difference in inches	−5	−4	−3	−2	−1	0	1	2	3	4	5
Frequency . .	·25	1·5	1·25	4·5	11·25	27·25	71·75	122·75	171·75	209·75	220·5

Stature difference in inches	6	7	8	9	10	11	12	13	14	15	16	Total
Frequency . .	205·5	148·75	95·75	57	26	11·25	8·5	2·75	1	1	·75	1401

Treating this distribution as before we obtain: mean=4.895, variance=6.4074, variance of mean=.004573, standard error of mean =.0676 ; showing that we may estimate the mean sex difference as 4¾ to to 5 inches.

In the above examples, which are typical of the use of the standard error applied to mean values, we have assumed that the variance of the population is known with exactitude. It was pointed out by "Student" in 1908, that with small samples, such as are of necessity usual in field and laboratory experiments, the variance of the population can only be roughly estimated from the sample, and that the errors of estimation seriously affect the use of the standard error. [p. 106]

If *x* (for example the mean of a sample) is a value with normal distribution and s is its true standard error, then the probability that x/s exceeds any specified value may be obtained from the appropriate table of the normal distribution; but if we do not know s, but in its place have *s*, an estimate of the value of s,the distribution required will be that of *x/s*, and this is not normal. The true value has been divided by a factor, *s*/s, which introduces an error. We have seen in the last chapter that the distribution in random samples of s^2/s^2 is that of c^2/n, when *n* is equal to the number of degrees of freedom, of the group (or groups) of which s^2 is the mean square deviation. Consequently the distribution of *s*/s calculable, and if its variation is completely independent of that of *x*/s (as in the cases to which this method is applicable), then the true distribution of *x/s* can be calculated, and accurate allowance made for its departure from normality. The only modification required in these cases depends solely on the number *n*, representing the number of degrees of freedom available for the estimation of s. The necessary distributions were given by "Student" in 1908; fuller tables have since been given by the same author, and at the end of this chapter (p. 137) we give the distributions in a similar form to that used for our table of c^2.

24. The Significance of the Mean of a Unique Sample

If x_1, x_2, ..., x_n, is a sample of *n'* values of a variate, *x*, and if this sample constitutes the whole of [p. 107] the information available on the point in question, then we may test whether the mean of *x* differs significantly from zero, by calculating the statistics

$$\bar{x} = \frac{1}{n'} S(x),$$

$$\frac{s^2}{n'} = \frac{1}{n'(n'-1)} S(x-\bar{x})^2,$$

$$t = \frac{\bar{x}\sqrt{n'}}{s},$$

$$n = n' - 1.$$

The distribution of t for random samples of a normal population distributed about zero as mean, is given in the table of t for each value of n. The successive columns show, for each value of n, the values of t for which P, the probability of falling outside the range [plus or minus]t, takes the values .9,...,.01, at the head of the columns. Thus the last column shows that, when n=10, just I per cent of such random samples will give values of t exceeding +3.169, or less than -3.169. If it is proposed to consider the chance of exceeding the given values of t, in a positive (or negative) direction only, then the values of P should be halved. It will be seen from the table that for any degree of certainty we require higher values of t, the smaller the value of n. The bottom line of the table, corresponding to infinite values of n, gives the values of a normally distributed variate, in terms of its standard deviation, for the same values of P.

Ex. 19. *Significance of mean of a small sample*. -- The following figures (Cushny and Peebles' data) [p. 108] which I quote from Student's paper show the result of an experiment with ten patients, on the effect of the optical isomers of hyoscyamine hydrobromide in producing sleep.

TABLE 27

ADDITIONAL HOURS OF SLEEP GAINED BY THE USE OF HYOSCYAMINE HYDROBROMIDE

Patient.	1 (Dextro –).	2 (Laevo –).	Difference (2 – 1).
1	+0·7	+1·9	+1·2
2	–1·6	+0·8	+2·4
3	–0·2	+1·1	+1·3
4	–1·2	+0·1	+1·3
5	–0·1	–0·1	0·0
6	+3·4	+4·4	+1·0
7	+3·7	+5·5	+1·8
8	+0·8	+1·6	+0·8
9	0·0	+4·6	+4·6
10	+2·0	+3·4	+1·4
Mean ($\bar{x}$)	+·75	+2·33	+1·58

The last column gives a controlled comparison of the efficacy of the two drugs as soporifics, for the same patients were used to test each; from the series of differences we find

$$\bar{x} = +1{\cdot}58,$$

$$\frac{s^2}{10} = {\cdot}1513,$$

$$s/\sqrt{10} = {\cdot}3890,$$

$$t = 4{\cdot}06$$

For n=9, only one value in a hundred will exceed *3250* by chance, so that the difference between the results is clearly significant. By the methods of the [p. 109] previous chapters we should, in this case, have been led to the same conclusion with almost equal certainty; for if the two drugs had been equally effective, positive and negative signs would occur in the last column with equal frequency. Of the 9 values other than zero, however, all are positive, and it appears from the binomial distribution,

$$(\tfrac{1}{2}+\tfrac{1}{2})^9,$$

that all will be of the same sign, by chance, only twice in 512 trials. The method of the present chapter differs from that in taking account of the actual values and not merely of their signs, and is consequently the more reliable method when the actual values are available.

To test whether two samples belong to the same population, or differ significantly in their means. If x'_1, $x'_2, \ldots, x'_{n_1+1}$, and If x_1, $x_2, \ldots, x_{n_2+1}$ be two samples, the significance of the difference between their means may be tested by calculating the following statistics.

$$\bar{x} = \frac{1}{n_1+1} S(x), \quad \bar{x}' = \frac{1}{n_2+1} S(x'),$$

$$s^2\left(\frac{1}{n_1+1}+\frac{1}{n_2+1}\right) = \frac{(n_1+n_2+2)}{(n_1+1)(n_2+1)(n_1+n_2)}\{S(x-\bar{x})^2+S(x'-\bar{x}')^2\},$$

$$t = \frac{\bar{x}-\bar{x}'}{s}\sqrt{\frac{(n_1+1)(n_2+1)}{n_1+n_2+2}},$$

$$n = n_1+n_2.$$

The means are calculated as usual; the standard [p. 110] deviation is estimated by pooling the sums of squares from the two samples and dividing by the total number of the degrees of freedom contributed by them; if a were the true standard deviation, the variance of the first mean would be $s^2/(n_1+1)$, of the second mean $s^2/(n_2+1)$, and therefore of the difference $s^2\{1/(n_1+1)+1/(n_2+1)\}$; t is therefore found by dividing $\bar{x}-\bar{x}'$ by its standard error as estimated, and the error of the estimation is allowed for by entering the table with n equal to the number of degrees of freedom available for estimating s; that is $n=n_1+n_2$. It is thus possible to extend Student's treatment of the error of a -an to the comparison of the means of two samples.

Ex. 20. *Significance of difference of means of small samples*. -- Let us suppose that the above figures (Table 27) had been obtained using different patients for the two drugs; the experiment would have been less well controlled, and we should expect to obtain less certain results from the same number of observations, for it is *a priori* probable, and the above figures suggest, that personal variations in response to the drugs will be to some extent correlated.

Taking, then, the figures to represent two different sets of patients, we have

$$\bar{x} - \bar{x}' = +1{\cdot}58,$$
$$s^2(\tfrac{1}{10} + \tfrac{1}{10}) = {\cdot}7213,$$
$$t = +1{\cdot}861,$$
$$n = 18.$$

The value of P is, therefore, between .1 and .05, and [p. 111] cannot be regarded as significant. This example shows clearly the value of design in small scale experiments, and that the efficacy of such design is capable of statistical measurement.

The use of Student's distribution enables us to appreciate the value of observing a sufficient number of parallel cases; their value lies, not only in the fact that the probable error of a mean decreases inversely as the square root of the number of parallels, but in the fact that the accuracy of our estimate of the probable error increases simultaneously. The need for duplicate experiments is sufficiently widely realised; it is not so widely understood that in some cases, when it is desired to place a high degree of confidence (say P =.01) on the results, triplicate experiments will enable us to detect with confidence differences as small as one-seventh of those which, with a duplicate experiment, would justify the same degree of confidence.

The confidence to be placed in a result depends not only on the actual value of the mean value obtained, but equally on the agreement between parallel experiments. Thus, if in an agricultural experiment a first trial shows an apparent advantage of 8 tons to the acre, and a duplicate experiment shows an advantage of 9 tons, we have $n=1$, $t=17$, and the results would justify some confidence that a real effect had been observed; but if the second experiment had shown an apparent advantage of 18 tons, although the mean is now higher, we should place not more but less confidence in the conclusion that the treatment was [p. 112] beneficial, for t has fallen to 2.6, a value which for $n=1$ is often exceeded by chance. The apparent paradox may be explained by pointing out that the difference of 10 tons between the experiments indicates the existence of uncontrolled circumstances so influential that in both cases the apparent benefit may be due to chance, whereas in the former case the relatively close agreement of the results suggests that the uncontrolled factors are not so very influential. Much of the advantage of further replication lies in the fact that with duplicates our estimate of the importance of the uncontrolled factors is so extremely hazardous.

In cases in which each observation of one series corresponds in some respects to a particular observation of the second series, it is always legitimate to take the differences and test them as in Ex. 18, 19 as a single sample; but it is not always desirable to do so. A more precise comparison is obtainable by this method only if the corresponding values of the two series are positively correlated, and only if they are correlated to a sufficient extent to counterbalance the loss of precision due to basing our estimate of variance upon fewer degrees of freedom. An example will make this plain.

Ex. 21. *Significance of change in bacterial numbers*. -- The following table shows the mean number of bacterial colonies per plate obtained by four slightly different methods from soil samples taken at 4 P.M. and 8 P.M. respectively (H. G. Thornton's data): [p. 113]

TABLE 28

Method.	4 P.M.	8 P.M.	Difference.
A	29·75	39·20	+9·45
B	27·50	40·60	+13·10
C	30·25	36·20	+5·95
D	27·80	42·40	+14·60
Mean	28·825	39·60	+10·775

From the series of differences we have *x*[bar]=+10.775, $\frac{1}{4}s^2$=3.756, *t*=5.560, *n*=3, whence the table shows that P is between .01 and .02. If, On the contrary, we use the method of Ex. 20, and treat the two separate series, we find *x*[bar]-x'[bar]=+10.775, $\frac{1}{2}s^2$=2.188, *t* =7.285, *n*=6; this is not only a larger value of *n* but a larger value of *t*, which is now far beyond the range of the table, showing that P is extremely small. In this case the differential effects of the different methods are either negligible, or have acted quite differently in the two series, so that precision was lost in comparing each value with its counterpart in the other series. In cases like this it sometimes occurs that one method shows no significant difference, while the other brings it out; if either method indicates a definitely significant difference, its testimony cannot be ignored, even if the other method fails to show the effect. When no correspondence exists between the members of one series and those of the other, the second method only is available. [p. 114]

25. Regression Coefficients

The methods of this chapter are applicable not only to mean values, in the strict sense of the word, but to the very wide class of statistics known as regression coefficients. The idea of regression is usually introduced in connection with the theory of correlation, but it is in reality a more general, and, in some respects, a simpler idea, and the regression co-efficients are of interest and scientific importance in many classes of data where the correlation coefficient, if used at all, is an artificial concept of no real utility. The following qualitative examples are intended to familiarise the student with the concept of regression, and to prepare the way for the accurate treatment of numerical examples.

It is a commonplace that the height of a child depends on his age, although knowing his age, we cannot accurately calculate his height. At each age the heights are scattered over a considerable range in a frequency distribution characteristic of that age; any feature of this distribution, such as the mean, will be a continuous function of age.

The function which represents the mean height at any age is termed the regression function of height on age; it is represented graphically by a regression curve, or regression line. In relation to such a regression line age is termed the **independent** variate, and *height* the **dependent** variate.

The two variates bear very different relations to the regression line. If errors occur in the heights, this [p. 115] will not influence the regression of height on age, provided that at all ages positive and negative errors are equally frequent, so that they balance in the averages. On the contrary, errors in age will in general alter the regression of height on age, so that from a record with ages subject to error, or classified in broad age-groups, we should not obtain the true physical relationship between mean height and age. A second difference should be noted: the regression function does not depend on the frequency distribution of the independent variate, so that a true regression line may be obtained even when the age groups are arbitrarily selected, as when an investigation deals with children of "school age." On the other hand a selection of the dependent variate will change the regression line altogether.

It is clear from the above instances that the regression of height on age is quite different from the regression of age on height; and that one may have a definite physical meaning in cases in which the other has only the conventional meaning given to it by mathematical definition. In certain cases both regressions are of equal standing; thus, if we express in terms of the height of the father the average adult height of sons of fathers of a given height, observation shows that each additional inch of the fathers' height corresponds to about half an inch in the mean height of the sons. Equally, if we take the mean height of the fathers of sons of a given height, we find that each additional inch of the sons' height corresponds to half an inch in the mean height of the fathers. No selection [p. 116] has been exercised in the heights either of fathers or of sons; each variate is distributed normally, and the aggregate of pairs of values forms a normal correlation surface. Both regression lines are straight, and it is consequently possible to express the facts of regression in the simple rules stated above.

When the regression line with which we are concerned is straight, or, in other words, when the regression function is linear, the specification of regression is much simplified, for in addition to the general means we have only to state the ratio which the increment of the mean of the dependent variate bears to the corresponding increment of the independent variate. Such ratios are termed regression coefficients. The regression function takes the form

$$Y = a+b(x-\bar{x}),$$

where b is the regression coefficient of y on x, and Y is the mean value of y for each value of x. The physical dimensions of the regression coefficient depend on those of the variates; thus, over an age range in which growth is uniform we might express the regression of height on age in inches per annum, in fact as an average growth rate, while the regression of father's height on son's height is half an inch per inch, or simply ½. Regression coefficients may, of course, be positive or negative.

Curved regression lines are of common occurrence ; in such cases we may have to use such a regression function as

$$Y = a+bx+cx^2+dx^3,$$ [p. 117]

in which all four coefficients of the regression function may, by an extended use of the term, be called regression coefficients. More elaborate functions of *x* may be used, but their practical employment offers difficulties in cases where we lack theoretical guidance in choosing the form of the regression function, and at present the simple power series (or, polynomial in *x*) is alone in frequent use. By far the most important case in statistical practice is the straight regression line.

26. Sampling Errors of Regression Coefficients

The straight regression line with formula

$$Y = a+b(x-\bar{x})$$

is fitted by calculating from the data, the two statistics

$$a=\bar{y}, \quad b=\frac{S\{y(x-\bar{x})\}}{S\{(x-\bar{x})^2\}};$$

these are estimates, derived from the data, of the two constants necessary to specify the straight line; the true regression formula, which we should obtain from an infinity of observations, may be represented by

a+b(*x*-*x*[bar]),

and the differences *a*-a, *b*-b, are the errors of random sampling of our statistics. If s^2 represent the variance of *y* for any value of *x* about a mean given by the above formula, then the variance of *a*, the mean of *n*' observations, will be s^2/n', while that of *b*, which is [p. 118] merely a *weighted* mean of the values of *y* observed, will be

$$\sigma^2 \frac{S(x-\bar{x})^2}{\{S(x-\bar{x})^2\}^2}=\frac{\sigma^2}{S(x-\bar{x})^2}.$$

In order to test the significance of the difference between *b*, and any hypothetical value, b, to which it is to be compared, we must estimate the value of s^2; the best estimate for the purpose is

$$s^2=\frac{1}{n'-2}S(y-Y)^2,$$

found by summing the squares of the deviations of *y* from its calculated value Y, and dividing by (*n*'-2). The reason that the divisor is (*n*'-2) is that from the *n*' values of *y* two statistics have already been calculated which enter into the formula for Y, consequently the group of differences, *y*-Y, represent in reality only *n*'-2 degrees of freedom.

When *n*' is small, the estimate of s^2 obtained above is somewhat uncertain, and in comparing the difference *b*-b with its standard error, in order to test its significance we shall have to use Student's method, with *n*=*n*'-2. When *n*' is large this distribution tends to the normal distribution. The value of *t* with which the table must be entered is

$$\frac{(b-\beta)\sqrt{S(x-\bar{x})^2}}{s}.$$

Similarly, to test the significance of the difference between *a* and any hypothetical value a, the table is entered with [p. 119]

$$t=\frac{(a-\alpha)\sqrt{n'}}{s}, \quad n=n'-2;$$

this test for the *significance* of *a* will be more sensitive than the method previously explained, if the variation in *y* is to any considerable extent expressible in terms of that of *x,* for the value of *s* obtained from the regression line will then be smaller than that obtained from the original group of observations. On the other hand, one degree of freedom is always lost, so that if *b* is small, no greater precision is obtained.

Ex. 22. *Effect of nitrogenous fertilisers in maintaining* *yield*. -- The yields of dressed grain in bushels per acre shown in Table 29 were obtained from two plots on Broadbalk wheat field during thirty years; the only difference in manurial treatment was that "9*a*" received nitrate of soda, while "7*b*" received an equivalent quantity of nitrogen as sulphate of ammonia. In the course of the experiment plot "9*a*" appears to be gaining in yield on plot "7*b*." Is this apparent gain significant?

A great part of the variation in yield from year to year is evidently similar in the two plots; in consequence, the series of differences will give the clearer result. In one respect the above data are especially simple, for the thirty values of the independent variate form a series with equal intervals between the successive values, with only one value of the dependent variate corresponding to each. In such cases the work is simplified by using the formula

$$S(x-\bar{x})^2 = \frac{1}{12} n' (n'^2-1),$$ [p. 120]

TABLE 29

Harvest Year.	9 *a*.	7 *b*.	9 *a* − 7 *b*.
1855	29·62	33·00	−3·38
1856	32·38	36·91	−4·53
1857	43·75	44·84	−1·09
1858	37·56	38·94	−1·38
1859	30·00	34·66	−4·66
1860	32·62	27·72	+4·90
1861	33·75	34·94	−1·19
1862	43·44	35·88	+7·56
1863	55·56	53·66	+1·90
1864	51·06	45·78	+5·28
1865	44·06	40·22	+3·84
1866	32·50	29·91	+2·59
1867	29·13	22·16	+6·97
1868	47·81	39·19	+8·62
1869	39·00	28·25	+10·75
1870	45·50	41·37	+4·13
1871	34·44	22·31	+12·13
1872	40·69	29·06	+11·63
1873	35·81	22·75	+13·06
1874	38·19	39·56	−1·37
1875	30·50	26·63	+3·87
1876	33·31	25·50	+7·81
1877	40·12	19·12	+21·00
1878	37·19	32·19	+5·00
1879	21·94	17·25	+4·69
1880	34·06	34·31	−·25
1881	35·44	26·13	+9·31
1882	31·81	34·75	−2·94
1883	43·38	36·31	+7·07
1884	40·44	37·75	+2·69
Mean .	37·50	33·03	+4·47

$$S(x-\bar{x})^2 = \frac{n'(n'^2-1)}{12} = 2247·5$$

$$b = ·2668$$

$$S(y-\bar{y})^2 = 1020·56$$
$$b^2S(x-\bar{x})^2 = 159·99$$
$$S(y-Y)^2 = 860·57$$

$$s^2 = 30·73$$

$$s^2/S(x-\bar{x})^2 = ·013675 = (·1169)^2$$

$$t = 2·282$$

$$n = 28$$

where n' is the number of terms, or 30 in this case. To evaluate 6 it is necessary to calculate

$$S\{y(x-\bar{x})\};$$

this may be done in several ways. We may multiply [p. 121] the successive values of *y* by -29, -27,... +27, +29, add, and divide by 2. This is the direct method suggested by the formula. The same result is obtained by multiplying by 1, 2, ..., 30 and subtracting 15½

$$\left(=\frac{n'+1}{2}\right)$$

times the sum of values of *y*; the latter method may be conveniently carried out by successive addition. Starting from the bottom of the column, the successive sums 2.69, 9.76, 6.82, ... are written down, each being found by adding a new value of *y* to the total already accumulated; the sum of the new column, less 15½ times the sum of the previous column, will be the value required. In this case we find the value 599.615, and dividing by 2247.5, the value of *b* is found to be .2668. The yield of plot "9*a*" thus appears to have gained on that of "7*b*" at a rate somewhat over a quarter of a bushel per annum.

To estimate the standard error of 6, we require the value of

$$S(y-Y)^2;$$

knowing the value of *b*, it is easy to calculate the thirty values of Y from the formula

$$Y = \bar{y} + (x-\bar{x})b;$$

for the first value, $x-\bar{x}=-14.5$ and the remaining values may be found in succession by adding *b* each time. By subtracting each value of Y from the corresponding *y*, squaring, and adding, the required quantity may be calculated directly. This method is laborious, and it is preferable in practice to utilise the algebraical fact that [p. 122]

$$\begin{aligned} S(y - Y)^2 &= S(y - \bar{y})^2 - b^2 S(x - \bar{x})^2 \\ &= S(y^2) - n'\bar{y}^2 - b^2 S(x - \bar{x})^2. \end{aligned}$$

The work then consists in squaring the values of *y* and adding, then subtracting the two quantities which can be directly calculated from the mean value of *y* and the value of *b*. In using this shortened method it should be noted that small errors in *y*[bar] and *b* may introduce considerable errors in the result, so that it is necessary to be sure that these are calculated accurately to as many significant figures as are needed in the quantities to be subtracted. Errors of arithmetic which would have little effect in the first method, may altogether vitiate the results if the second method is used. The subsequent work in calculating the standard error of *b* may best be followed in the scheme given beside the table of data ; the estimated standard error is .1169, so that in testing the hypothesis that b=0 that is that plot "9*a*" has not been gaining on plot "7*b*," we divide *b* by this quantity and find *t*=2.282. Since *s* was found from 28 degrees of freedom *n*=28, and the table of *t* shows that P is between .02 and .05.·

The result must be judged significant, though barely so; in view of the data we cannot ignore the possibility that on this field, and in conjunction with the other manures used, nitrate of soda has conserved the fertility better than sulphate of ammonia ; these data do not, however, demonstrate the point beyond possibility of doubt.

The standard error of *y*[bar], calculated from the above data, is 1.012, so that there can be no doubt that the [p. 123] difference in mean yields is significant; if we had tested the significance of the mean, without regard to the order of the values, that is calculating s^2 by dividing 1020.56 by 29, the standard error would have been 1.083. The value of *b* was therefore high enough to have reduced the standard error. This suggests the possibility that if we had fitted a more complex regression line to the data the probable errors would be further reduced to an extent which would put the significance of *b* beyond doubt. We shall deal later with the fitting of curved regression lines to this type of data.

Just as the method of comparison of means is applicable when the samples are of different sizes, by obtaining an estimate of the error by combining the sums of squares obtained from the two different samples, so we may compare regression coefficients when the series of values of the independent variate are not identical; or if they are identical we can ignore the fact in comparing the regression coefficients.

Ex. 23. *Comparison of relative growth rate of two cultures of an alga*. -- Table 30 shows the logarithm (to the base 10) of the volumes occupied by algal cells on successive days, in parallel cultures, each taken over a period during which the relative growth rate was approximately constant. In culture A nine values are available, and in culture B eight (Dr. M. Bristol-Roach's data).

The method of finding S*y*(*x*-*x*[bar]) by summation is shown in the second pair of columns: the original values are added up from the bottom, giving successive [p. 124] totals from 6.087 to 43.426; the final value should, of course, tally with the total below the original values. From the sum of the column of totals is subtracted the sum of the original values multiplied by 5 for A and by 4½ for B. The differences are S*y*(*x*-*x*[bar]); these must be divided by the respective values of S(*x*-*x*bar])^2,

TABLE 30

	Log Values.		Summation Values.			
	A.	B.	A.	B.		
	3·592	3·538	43·426	38·358	$S(y-Y)^2$, A	·05089
	3·823	3·828	39·834	34·820	,, B	·07563
	4·174	4·349	36·011	30·992		
	4·534	4·833	31·837	26·643	ns^2	·12652
	4·956	4·911	27·303	21·810	s^2	·009732
	5·163	5·297	22·347	16·899	$s^2/60$	·0001622
	5·495	5·566	17·184	11·602	$s^2/42$	·0002317
	5·602	6·036	11·689	6·036		
	6·087	..	6·087	..		·0003939
Total	43·426	38·358	235·718	187·160	Standard error	·01985
Mean	4·8251	4·7947	217·130	172·611	$b'-b$	·0366
			$Sy(x-\bar{x})$ 18·588	14·549	t	1·844
			b ·3098	·3464	n	13

namely, 60 and 42, to give the values of b, measuring the relative growth rates of the two cultures. To test if the difference is *significant* we calculate in the two cases $S(y^2)$, and subtract successively the product of the mean with the total, and the product of b with $Sy(x-\bar{x})$; this process leaves the two values of $S(y-Y)^2$, which are added as shown in the table, and the sum divided by n, to give s^2. The value of n is found by adding the 7 degrees of freedom from series A to the 6 degrees from series B, and is therefore 13. [p. 125] Estimates of the variance of the two regression coefficients are obtained by dividing s^2 by 60 and 42, and that of the variance of their difference is the sum of these. Taking the square root we find the standard error to be .01985, and t=1.844· The difference between the regression coefficients, though relatively large, cannot be regarded as significant. There is not sufficient evidence to assert culture B was growing more rapidly than culture A.

27. The Fitting of Curved Regression Lines

Little progress has been made with the theory of the fitting of curved regression lines, save in the limited but most important case when the variability of the independent variate is the same for all values of the dependent variate, and is normal for each such value. When this is the case a technique has been fully worked out for fitting by successive stages any line of the form

$$Y = a+bx+cx^2+dx^3+..\,;$$

we shall give details of the case where the successive values of x are at equal intervals.

As it stands the above form would be inconvenient in practice, in that the fitting could not be carried through in successive stages. What is required is to obtain successively the mean of y, an equation linear in x, an equation quadratic in x, and so on, each equation being obtained from the last by adding, a new term being calculated by carrying a single process of [p. 126] computation through a new stage. In order to do this we take

$$Y = A + Bx_1 + Cx_2 + Dx_3 + ...,$$

where x_1, x_2, x_3, shall be functions of x of the 1st, 2nd, and 3rd degrees, out of which the regression formula may be built. It may be shown that the functions required for this purpose may be expressed in terms of the moments of the x distribution, as follows:

$$\xi_1 = x - \bar{x},$$

$$\xi_2 = \xi_1^2 - \mu_2 \qquad = \xi_1^2 - \frac{n'^2 - 1}{12},$$

$$\xi_3 = \xi_1^3 - \frac{\mu_4}{\mu_2}\xi_1 \qquad = \xi_1^3 - \frac{3n'^2 - 7}{20}\xi_1,$$

$$\xi_4 = \xi_1^4 - \frac{\mu_6 - \mu_2\mu_4}{\mu_4 - \mu_2^2}\xi_1^2 + \frac{\mu_2\mu_6 - \mu_4^2}{\mu_4 - \mu_2^2}$$

$$= \xi_1^4 - \frac{3n'^2 - 13}{14}\xi_1^2 + \frac{3(n'^2 - 1)(n'^2 - 9)}{560},$$

$$\xi_5 = \xi_1^5 - \frac{\mu_2\mu_8 - \mu_4\mu_6}{\mu_2\mu_6 - \mu_4^2}\xi_1^3 + \frac{\mu_4\mu_8 - \mu_6^2}{\mu_2\mu_6 - \mu_4^2}\xi_1$$

$$= \xi_1^5 - \frac{5(n'^2 - 7)}{18}\xi_1^3 + \frac{15n'^4 - 230n'^2 + 407}{1008}\xi_1,$$

where the values of the moment functions have been expressed in terms of n', the number of observations, as far as is needed for fitting curves up to the 5th degree. The values of x are taken to increase by unity.

Algebraically the process of fitting may now be represented by the equations

$$A = \bar{y} = \frac{1}{n'}S(y),$$

[p. 127]

$$B = \frac{12}{n'(n'^2-1)} S(y\xi_1),$$

$$C = \frac{180}{n'(n'^2-1)(n'^2-4)} S(y\xi_2),$$

and, in general, the coefficient of the term of the *r*th degree is

$$\frac{(2r)!(2r+1)!}{(r!)^4 n'(n'^2-1) \ldots (n'^2-r^2)} S(y\xi_r).$$

As each term is fitted the regression line approaches more nearly to the observed values, and the sum of the squares of the deviation

$$S(y-Y)^2$$

is diminished. It is desirable to be able to calculate this quantity, without evaluating the actual values of Y at each point of the series; this can be done by subtracting from $S(y^2)$ the successive quantities

$$n'A^2, \quad \frac{n'(n'^2-1)}{12}B^2, \quad \frac{n'(n'^2-1)(n'^2-4)}{180}C^2,$$

and so on. These quantities represent the reduction which the sum of the squares of the residuals suffers each time the regression curve is fitted to a higher degree; and enable its value to be calculated at any stage by a mere extension of the process already used in the preceding examples. To obtain an estimate, s^2, of the residual variance, we divide by *n*, the number of degrees of freedom left after fitting, which is found from *n'* by subtracting from it the number of constants in the regression formula. Thus, if a straight line has been fitted, *n*=*n'*-2; while if a curve of the fifth degree has been fitted, *n*=*n'*-6. [p. 128]

28. The Arithmetical Procedure of Fitting

The main arithmetical labour of fitting curved regression lines to data of this type may be reduced to a repetition of the process of summation illustrated in Ex. 23. We shall assume that the values of *y* are written down in a column in order of increasing values of *x,* and that at each stage the summation is commenced at the top of the column (not at the bottom, as in that example). The sums of the successive columns will be denoted by S_1, S_2, ... When these values have been obtained, each is divided by an appropriate divisor, which depends only on *n',* giving us a new series of quantities *a, b, c,...* according to the following equations

$$a = \frac{1}{n'} S_1 = \frac{1}{n'} S(y) = \bar{y},$$

$$b = \frac{1.2}{n'(n'+1)} S_2,$$

$$c = \frac{1.2.3.}{n'(n'+1)(n'+2)} S_3,$$

and so on.

From these a new series of quantities a', b', c',... are obtained by equations independent n', of which we give below the first six, which are enough to carry the process of fitting up to the 5th degree:

$$
\begin{aligned}
a' &= a, \\
b' &= a - b, \\
c' &= a - 3b + 2c, \\
d' &= a - 6b + 10c - 5d, \\
e' &= a - 10b + 30c - 35d + 14e, \\
f' &= a - 15b + 70c - 140d + 126e - 42f.
\end{aligned}
$$

These new quantities are proportional to the required coefficients of the regression equation, and need only be divided by a second group of divisors to give the actual values. The equations are

$$A = a', \qquad B = \frac{6}{n' - 1} b',$$

$$C = \frac{30}{(n' - 1)(n' - 2)} c', \qquad D = \frac{140}{(n' - 1)(n' - 2)(n' - 3)} d',$$

$$E = \frac{630}{(n' - 1)(n'^2 - 2) \dots (n' - 4)} e', \quad F = \frac{2772}{(n' - 1) \dots (n' - 5)} f',$$

the numerical part of the factor being

$$\frac{(2r + 1)!}{(r!)^2}$$

for the term of degree r.

If an equation of degree *r* has been fitted, the estimate of the standard errors of the coefficients are all based upon the same value of s^2, *i.e.*

$$s^2 = \frac{1}{n'-r-1}\left\{S(y^2) - n'A^2 - \frac{n'(n'^2-1)}{12}B^2 - \ldots\right\},$$

from which the estimated standard error of any coefficient, such as that of x_p, is obtained by dividing by

$$S(\xi_p^2) = \frac{(p!)^4}{(2p)!(2p+1)!}n'(n'^2-1)\ \ldots\ (n'^2-p^2)$$

and taking out the square root. The number of degrees of freedom upon which the estimate is based is $(n'-r-1)$, and this must be equated to n in using the table of t.

A suitable example of use of this method may be obtained by fitting the values of Ex. 22 (p. 120) with a curve of the second or third degree. [p. 130]

29. Regression with several Independent Variates

It frequently happens that the data enable us to express the average value of the dependent variate *y*, in terms of a number of different independent variates $x_1, x_2, \ldots x_p$. For example, the rainfall at any point within a district may be recorded at a number of stations for which the longitude, latitude, and altitude are all known. If all of these three variates influence the rainfall, it may be required to ascertain the average effect of each separately. In speaking of longitude, latitude, and altitude as independent variates, all that is implied is that it is in terms of them that the average rainfall is to be expressed; it is not implied that these variates vary independently, in the sense that they are uncorrelated. On the contrary, it may well happen that the more southerly stations lie on the whole more to the west than do the more northerly stations, so that for the stations available longitude measured to the west may be negatively correlated with latitude measure to the north. If, then, rainfall increased to the west but was independent of latitude, we should obtain merely, by comparing the rainfall recorded at different latitudes, a fictitious regression indicating a falling off of rain with increasing latitude. What we require is an equation taking account of all three variates at each station, and agreeing as nearly as possible with the values recorded; this is called a partial regression equation, and its coefficients are known as partial regression coefficients. [p. 131]

To simplify the algebra we shall suppose that y, x_1, x_2, x_3, are all measured from their mean values, and that we are seeking a formula of the form

$$Y = b_1, x_1 + b_2 x_2 + b_3 x_3.$$

If S stands for summation over all the sets of observations we construct the three equations

$$b_1 S(x_1^2) + b_2 S(x_1 x_2) + b_3 S(x_1 x_3) = S(x_1 y),$$
$$b_1 S(x_1 x_2) + b_2 S(x_2^2) + b_3 S(x_2 x_3) = S(x_2 y),$$
$$b_1 S(x_1 x_3) + b_2 S(x_2 x_3) + b_3 S(x_3^2) = S(x_3 y),$$

of which the nine coefficients are obtained from the data either by direct multiplication and addition, or, if the data are numerous, by constructing correlation tables for each of the six pairs of variates. The three simultaneous equations for b_1, b_2, and b_3, are solved in the ordinary way; first b_3 is eliminated from the first and third, and from the second and third equations, leaving two equations for b_1 and b_2; eliminating b_2 from these, b_1 is found, and thence by substitution, b_2 and b_3.

It frequently happens that, for the same set of values of the independent variates, it is desired to examine the regressions for more than one set of values of the dependent variates; for example, if for the same set of rainfall stations we had data for several different months or years. In such cases it is preferable to avoid solving the simultaneous equations afresh on each occasion, but to obtain a simpler formula which may be applied to each new case.

This may be done by solving once and for all the [p. 132] three sets, each consisting of three simultaneous equations:

$$b_1 S(x_1^2) + b_2 S(x_1 x_2) + b_3 S(x_1 x_3) = 1, \quad 0, \quad 0,$$
$$b_1 S(x_1 x_2) + b_2 S(x_2^2) + b_3 S(x_2 x_3) = 0, \quad 1, \quad 0,$$
$$b_1 S(x_1 x_3) + b_2 S(x_2 x_3) + b_3 S(x_3^2) = 0, \quad 0, \quad 1;$$

the three solutions of these three sets of equations may be written

$$b_1 = c_{11}, \quad c_{12}, \quad c_{13},$$
$$b_2 = c_{12}, \quad c_{22}, \quad c_{23},$$
$$b_3 = c_{13}, \quad c_{23}, \quad c_{33}.$$

Once the six values of *c* are known, then the partial regression coefficients may be obtained in any particular case merely by calculating $S(x_1 y)$, $S(x_2 y)$, $S(x_3 y)$ and substituting in the formulæ,

$$b_1 = c_{11} S(x_1 y) + c_{12} S(x_2 y) + c_{13} S(x_3 y),$$
$$b_2 = c_{12} S(x_1 y) + c_{22} S(x_2 y) + c_{23} S(x_3 y),$$
$$b_3 = c_{13} S(x_1 y) + c_{23} S(x_2 y) + c_{33} S(x_3 y).$$

The method of partial regression is of very wide application. It is worth noting that the different independent variates may be related in any way; for example, if we desired to express the rainfall as a linear function of the latitude and longitude, and as a quadratic function of the altitude, the square of the altitude would be introduced as a fourth independent variate, without in any way disturbing the process outlined above, save that

$S(x_3x_4)$, $S(x_3{}^3)$ would be calculated directly from the distribution of altitude.

In estimating the sampling errors of partial [p. 133] regression coefficients we require to know how nearly our calculated value, Y, has reproduced the observed values of y; as in previous cases, the sum of the squares of $(y-Y)$ may be calculated by differences, for, with three variates,

$$S(y-Y)^2 = S(y^2) - b_1S(x_1y) - b_2S(x_2y) - b_3S(x_3y).$$

If we had n' observations, and p independent variates, we should therefore find

$$s^2 = \frac{1}{n'-p-1}S(y-Y)^2,$$

and to test if b_1, differed significantly from any hypothetical value, b_1, we should calculate

$$t = \frac{b_1 - \beta_1}{s\sqrt{c_{11}}},$$

entering the table of t with $n=n'-p-1$.

In the practical use of a number of variates it is convenient to use cards, on each of which is entered the values of the several variates which may be required. By sorting these cards in suitable grouping units with respect to any two variates the corresponding correlation table may be constructed with little risk of error, and thence the necessary sums of squares and products obtained.

Ex. 24. *Dependence of rainfall on position and altitude*. -- The situations of 57 rainfall stations in Hertfordshire have a mean longitude 12'.4 W., a mean latitude 51° 48'.5 N., and a mean altitude 302 feet. Taking as units 2 minutes of longitude, one [p. 134] minute of latitude, and 20 feet of altitude, the following values of the sums of squares and products of deviations from the mean were obtained:

$$S(x_1^2) = 1934{\cdot}1, \qquad S(x_2x_3) = +119{\cdot}6,$$
$$S(x_2^2) = 2889{\cdot}5, \qquad S(x_3x_1) = +924{\cdot}1,$$
$$S(x_3^2) = 1750{\cdot}8, \qquad S(x_1x_2) = -772{\cdot}2.$$

To find the multipliers suitable for any particular set of weather data from these stations, first solve the equations

$$1934.1\, c_{11} - 772.2\, c_{12} + 924.1\, c_{13} = 1$$
$$-772.2\, c_{11} + 2889.5\, c_{12} + 119.6\, c_{13} = 0$$
$$+924.1\, c_{11} + 119.6\, c_{13[sic]} + 1750.8\, c_{13} = 0;$$

using the last equation to eliminate c_{13} from the first two, we have

$$2532.3\, c_{11} - 1462.5\, c_{12} = 1.7508$$
$$=1462.5\, c_{11} + 5044.6\, c_{12} = 0;$$

from these eliminate c_{12}, obtaining

$$10{,}635.5\, c_{11} = 8.8321;$$

whence

$$c_{11} = .00083043,\ c_{12} = .00024075,\ c_{13} = -.00045476$$

the last two being obtained successively by substitution.

Since the corresponding equations for c_{12}, c_{22}, c_{23} differ only in changes in the right-hand number, we can at once write down

$$-1462.5\ c_{12} + 5044.6\ c_{22} = 1.7508;$$

whence, substituting for c_{12} the value already obtained,

$$c_{22} = .00041686,\ c_{23} = -.00015554;$$ [p. 135]

finally, to obtain c_{33} we have only to substitute in the equation

$$924.1c_{13} + 119.6c_{23} + 1750.8c_{33} = 1,$$

giving

$$c_{33} = .00082182.$$

It is usually worth while, to facilitate the detection of small errors by checking, to retain as above one more decimal place than the data warrant.

The partial regression of any particular weather data on these three variates can now be found with little labour. In January 1922 the mean rainfall recorded at these stations was 3.87 inches, and the sums of products of deviations with those of the three independent variates were (taking 0.1 inch as the unit for rain)

$$S(x_1y) = +1137.4,\ S(x_2y) = -592.9,\ S(x_3y) = +891.8;$$

multiplying these first by c_{11}, c_{12}, c_{13} and adding, we have for the partial regression on longitude

$$b_2 = .39624;$$

similarly using the multipliers c_{12}, c_{22}, c_{23} we obtain for the partial regression on latitude

$$b_2 = -11204;$$

and finally, by using c_{13}, c_{23}, c_{33},

$$b_3 = .30787$$

gives the partial regression on altitude.

Remembering now the units employed, it appears that in the month in question rainfall increased by .0198 of an inch for each minute of longitude westwards, [p. 136] is decreased by .0112 of an inch for each minute of latitude northwards, and increased by .00154 of an inch for each foot of altitude.

Let us calculate to what extent the regression on altitude is affected by sampling errors. For the 57 recorded deviations of the rainfall from its mean value, in the units previously used

$$S(y^2) = 1786.6;$$

whence, knowing the values of b_1, b_2, and b_3, we obtain by differences

$$S(y-Y)^2 = 994.9.$$

To find s^2, we must divide this by the number of degrees of freedom remaining after fitting a formula involving three variates -- that is, by 53 -- so that

$$s^2 = 8.772;$$

multiplying this by c_{33}, and taking the square root,

$$s\sqrt{c_{33}} = .12421.$$

Since n is as high as 53 we shall not be far wrong in taking the regression of rainfall on altitude to be in working units .308, with a standard error .124; or in inches of rain per 100 feet as .154, with a standard error .062.

TABLE IV.—TABLE OF t

n.	P = ·9.	·8.	·7.	·6.	·5.	·4.	·3.	·2.	·1.	·05.	·02.	·01.
1	·158	·325	·510	·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841
4	·134	·271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4·604
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169
11	·129	·260	·396	·540	·697	·876	1·088	1·363	1·796	2·201	2·718	3·106
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055
13	·128	·259	·394	·538	·694	·870	1·079	1·350	1·771	2·160	2·650	3·012
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977
15	·128	·258	·393	·536	·691	·866	1·074	1·341	1·753	2·131	2·602	2·947
16	·128	·258	·392	·535	·690	·865	1·071	1·337	1·746	2·120	2·583	2·921
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807
24	·127	·256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787
26	·127	·256	·390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750
∞	·12566	·25335	·38532	·52440	·67449	·84162	1·03643	1·28155	1·64485	1·95996	2·32634	2·57582

Classics in the History of Psychology

An internet resource developed by
Christopher D. Green
York University, Toronto, Ontario
ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted April 2000

VI

THE CORRELATION COEFFICIENT

30. No quantity is more characteristic of modern statistical work than the correlation coefficient, and no method has been applied successfully to such various data as the method of correlation. Observational data in particular, in cases where we can observe the occurrence of various possible contributory causes of a phenomenon, but cannot control them, has been given by its means an altogether new importance. In experimental work proper its position is much less central; it will be found useful in the exploratory stages of an enquiry, as when two factors which had been thought independent appear to be associated in their occurrence; but it is seldom, with controlled experimental conditions, that it is desired to express our conclusion in the form of a correlation coefficient.

One of the earliest and most striking successes of the method of correlation was in the biometrical study of inheritance. At a time when nothing was known of the mechanism of inheritance, or of the structure of the germinal material, it was possible by this method to demonstrate the existence of inheritance, and to [p. 139] "measure its intensity"; and this in an organism in which experimental breeding could not be practised, namely, Man. By comparison of the results obtained from the physical measurements in man with those obtained from other organisms, it was established that man's nature is not less governed by heredity than that of the rest of the animate world. The scope of the analogy was further widened by demonstrating that correlation coefficients of the same magnitude were obtained for the mental and moral qualities in man as for the physical measurements.

These results are still of fundamental importance, for not only is inheritance in man still incapable of experimental study, and existing methods of mental testing are still unable to analyse the mental disposition, but even with organisms suitable for experiment and measurement, it is only in the most favourable cases that the several factors causing fluctuating variability can be resolved, and their effects studied, by Mendelian methods. Such fluctuating variability, with an approximately normal distribution, is characteristic of the majority of the useful qualities of domestic plants and animals; and although there is strong reason to think that inheritance in such cases is ultimately Mendelian, the biometrical method of study is at present alone capable of holding out hopes of immediate progress.

We give in Table 31 an example of a correlation table. It consists of a record in compact form of the stature of 1376 fathers and daughters. (Pearson and Lee's data.) The measurements are grouped in [p. 140-141] [table] [p. 142] inches, and those whose measurement was recorded as an integral number of inches have been split; thus a father recorded as of 67 inches would appear as 1/2 under 66.5 and 1/2 under 67.5. Similarly with the daughters; in consequence, when both measurements are whole numbers the case appears in four quarters. This gives the table a confusing appearance, since the majority of entries are fractional, although they represent frequencies. It is preferable, if bias in measurement can be avoided, to group the observations in such a way that each possible observation lies wholly within one group.

TABLE 31

Height of Daughters in Inches.	Height of Fathers in Inches.																		
	58·5	59·5	60·5	61·5	62·5	63·5	64·5	65·5	66·5	67·5	68·5	69·5	70·5	71·5	72·5	73·5	74·5	75·5	Total.
52·5	..	..	..	..	·25	·25	..	..	..	..	..	..	..	..	..	..	..	..	·5
53·5	..	..	..	..	·25	·25	..	..	..	..	..	..	..	..	..	..	..	..	·5
54·5	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
55·5	..	..	..	..	..	..	..	1	..	..	..	..	..	..	..	..	..	..	1
56·5	·25	·25	..	·25	1·25	·5	..	1	·5	·5	..	..	..	..	..	..	..	..	4·5
57·5	·25	·25	·5	1·5	4·5	1	1·5	1·5	2·5	..	·5	·5	..	..	..	..	..	..	14·5
58·5	·25	·75	·5	·75	·75	1	1·75	1·25	5	2·75	·5	·25	..	..	..	..	..	..	15·5
59·5	·5	1	2	..	6	4·75	5	6·25	11·75	3·5	3·5	2	1·75	·5	..	..	..	..	48·5
60·5	·75	·75	..	2·5	8	6·25	12·5	18·25	20·25	11	9	4·75	2·5	1·25	1·25	..	..	..	99
61·5	..	·5	1·75	2	9·75	11·5	13	23·75	23·75	20·25	16·5	10·25	4·25	3	1·25	..	..	..	141·5
62·5	..	1	2·25	2	4·5	12	22·75	26	33	28·25	24·75	14·25	13·75	4·75	·75	·5	..	..	190·5
63·5	..	..	·25	2	6	8·25	11	27·25	35·75	37·25	31·5	26·25	16·25	7·75	1·5	·75	·25	..	212
64·5	..	..	·25	2·5	1·75	3·25	9·25	23	18·75	28·5	33	34·25	24·5	11·75	5·5	1	·25	1	198·5
65·5	..	..	..	·5	1	·5	11	12·25	9·25	19·75	30	26·5	22·25	15	4·75	3·75	2	1	159·5
66·5	..	..	..	·5	·5	1·5	3·25	7·25	8·75	16	26·25	26·75	20·5	18·5	7·75	4·25	·25	·5	142·5
67·5	..	..	..	..	..	..	1	5·75	7	4	14·25	13·25	12	11·25	4·5	3·75	·75	..	77·5
68·5	..	..	..	..	·25	·25	·25	·25	1·5	3	5·5	4·25	5·75	5·25	3·75	2·5	1·5	2	36
69·5	..	..	..	..	·25	·25	·25	·25	·25	·25	1	2·5	6·5	2·25	2·75	2	1	..	19·5
70·5	..	..	..	..	..	..	..	..	..	..	1·75	·25	4·5	·75	1·25	·75	·25	..	9·5
71·5	..	..	..	..	..	..	..	..	..	..	·5	..	·5	·5	1·5	·75	·25	..	4
72·5	..	..	..	..	..	..	..	..	..	..	1	..	..	..	..	..	..	..	1
Total	2	4·5	7·5	14·5	45	51·5	92·5	155	178	175	199·5	166	135	82·5	36·5	20	6·5	4·5	1376

The most obvious feature *of* the table is that cases do not occur in which the father is very tall and the daughter very short, and *vice versa* ; the upper right-hand and lower left-hand corners of the table are blank, so that we may conclude that such occurrences are too rare to occur in a sample of about 1400 cases. The observations recorded lie in a roughly elliptical figure lying diagonally across the table. If we mark out the region in which the frequencies exceed 10 it appears that this region, apart from natural irregularities, is similar, and similarly situated. The frequency of occurrence increases from all sides to the central region of the table, where a few frequencies over 30 may be seen. The lines of equal frequency are roughly similar and similarly situated ellipses. In the outer zone observations occur only occasionally, and therefore irregularly; beyond this we could only explore by taking a much larger sample.

The table has been divided into four quadrants by [p. 143] marking out central values of the two variates; these values, 67.5 inches for the fathers and 63.5 inches for the daughters, are near the means. When the table is so divided it is obvious that the lower right-hand and upper left-hand quadrants are distinctly more populous than the other two; not only are more squares occupied, but the frequencies are higher. It is apparent that tall men have tall daughters more frequently than the short men, and *vice versa*. The method of correlation aims at measuring the degree to which this association exists.

The marginal totals show the frequency distributions of the fathers and the daughters respectively. These are both approximately normal distributions, as is frequently the case with biometrical data collected without selection. This marks a frequent difference between biometrical and experimental data. An experimenter would perhaps have bred from two contrasted groups of fathers of, for example, 63 and 72 inches in height; all his fathers would then belong to these two classes, and the correlation coefficient, if used, would be almost meaningless. Such an experiment would serve to ascertain the regression of daughter's height in father's height, and so to determine the effect on the daughters of selection applied to the fathers, but it would not give us the correlation coefficient which is a descriptive observational feature of the population as it is, and may be wholly vitiated by selection.

Just as normal variation with one variate may be specified by a frequency formula in which the [p. 144] logarithm of the frequency is a quadratic function of the variate, so with two variates the frequency may be expressible in terms of a quadratic function of the values of the two variates. We then have a normal correlation surface, for which the frequency may conveniently be written in the form

$$df = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}} dxdy.$$

In this expression *x* and *y* are the deviations of the two variates from their means, s_1 and s_2 are the two standard deviations, and r is the *correlation* between *x* and *y.* The correlation in the above expression may be positive or negative, but cannot exceed unity in magnitude; it is a pure number without physical dimensions. If r=0, the expression for the frequency degenerates into the product of the two factors

$$\frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_1^2}} dx \cdot \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma_2^2}} dy,$$

showing that the limit of the normal correlation surface, when the correlation vanishes, is merely that of two normally distributed variates varying in complete independence. At the other extreme, when *p* is +1 or -1, the variation of the two variates is in strict proportion, so that the value of either may be calculated accurately from that of the other. In other words, we cease strictly to have two variates, but merely two measures of the same variable quantity.

If we pick out the cases in which one variate has an assigned value, we have what is termed an array; [p. 145] the columns and rows of the table may, except as regards variation within the group limits, be regarded as arrays. With normal correlation the variation within an array may be obtained from the general formula, by giving *x* a constant value, (say) *a*, and dividing by the total frequency with which this value occurs; then we have

$$df = \frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left(y-\rho\frac{a\sigma_2}{\sigma_1}\right)^2},$$

showing (i.) that the variation of *y* within the array is normal ; (ii.) that the mean value of *y* for that array is ras_2/s_1, so that the regression of *y* on *x* is linear, with regression coefficient

$$\rho\frac{\sigma_2}{\sigma_1};$$

and (iii.) that the variance of *y* within the array is $s_2^2(1-r^2)$, and is the same within each array. We may express this by saying that of the total variance of *y* the fraction $(1-r^2)$ is independent of *x*, while the remaining fraction, r^2, is determined by, or calculable from, the value of *x*.

These relations are reciprocal, the regression of *x* on *y* is linear, with regression coefficient rs_1/s_2; the correlation r is thus the geometric mean of the two regressions. The two regression lines representing the mean value of *x* for given *y*, and the mean value of *y* for given *x*, cannot coincide unless r=[plus or minus]1. The variation of *x* within an array in which *y* is fixed, is normal with variance equal to $s_1^2(1-r^2)$, so that we may say that of the variance of *x* the fraction $(1-r^2)$ [p. 146] is independent of *y*, and the remaining fraction, r^2, is determined by, or calculable from, the value of *y*.

Such are the formal mathematical consequences of normal correlation. Much biometric data certainly shows a general agreement with the features to be expected on this assumption; though I am not aware that the question has been subjected to any sufficiently critical enquiry. Approximate agreement is perhaps all that is needed to justify the use of the correlation as a quantity descriptive of the population; its efficacy in this respect is undoubted, and it is not improbable that in some cases it affords a complete description of the simultaneous variation of the variates.

31. The Statistical Estimation of the Correlation

Just as the mean and the standard deviation of a normal population in one variate may be most satisfactorily estimated from the first two moments of the observed distribution, so the only satisfactory estimate of the correlation, when the variates are normally correlated, is found from the "product moment." If *x* and *y* represent the deviations of the two variates from their means, we calculate the three statistics s_1, s_2, r by the three equations

$$ns_1{}^2 = \mathrm{S}(x^2),\ ns_2{}^2 = \mathrm{S}(y^2),\ nrs_1s_2 = \mathrm{S}(xy);$$

then s_1 and s_2 are estimates of the standard deviations s_1, and s_2, and r is an estimate of the correlation r. Such an estimate is called the *correlation coefficient*, or the *product moment correlation*, the latter term [p. 147] referring to the summation of the product terms, *xy*, in the last equation.

The above method of calculation might have been derived from the consideration that the correlation of the population is the geometric mean of the two regression coefficients; for our estimates of these two regressions would be

$$\frac{\mathrm{S}(xy)}{\mathrm{S}(x^2)} \text{ and } \frac{\mathrm{S}(xy)}{\mathrm{S}(y^2)},$$

so that it is in accordance with these estimates to take as our estimate of r

$$r = \frac{S(xy)}{\sqrt{S(x^2) \, . \, S(y^2)}},$$

which is in fact the product moment correlation.

Ex. 25. *Parental correlation in stature.* -- The numerical work required to calculate the correlation coefficient is shown below in Table 32.

The first eight columns require no explanation, since they merely repeat the usual process of finding the mean and standard deviation of the two marginal distributions. It is not necessary actually to find the mean, by dividing the total of the third column, 480.5, by 1376, since we may work all through with the undivided totals. The correction for the fact that our working mean is not the true mean is performed by subtracting $(480.5)^2/1376$ in the 4th column; a similar correction appears at the foot of the 8th column, and at the foot of the last column. The correction for the sum of products is performed by subtracting 4805x2605/1376· This correction of [p. 148] [table] [p. 149] the product term may be positive or negative; if the total deviations of the two variates are of opposite sign, the correction must be added. The sum of squares, with and without Sheppard's correction (1376/12), are shown separately; there is no corresponding correction to be made to the product term.

TABLE 32

Daughters.				Fathers.				Total for Daughters.	Product.
Deviation.	Frequency.			Deviation.	Frequency.				
−11	·5	5·5	60·5	−9	2	18	162	− 8·75	+ 78·75
−10	·5	5	50	−8	4·5	36	288	− 15·25	+122
− 9	−	−	−	−7	7·5	52·5	367·5	− 19	+133
− 8	1	8	64	−6	14·5	87	522	− 23	+138
− 7	4·5	31·5	220·5	−5	45	225	1125	−108·75	+543·75
− 6	14·5	87	522	−4	51·5	206	824	− 81	+324
− 5	15·5	77·5	387·5	−3	92·5	277·5	832·5	− 76·25	+228·75
− 4	48·5	194	776	−2	155	310	620	− 88·50	+177
− 3	99	297	891	−1	178	178	178	−131·25	+131·25
− 2	141·5	283	566	0	175	1390	..	+ 15·5	..
− 1	190·5	190·5	190·5	1	199·5	199·5	199·5	+183·25	+183·25
0	212	−1179	..	2	166	332	664	+197·25	+394·5
1	198·5	198·5	198·5	3	135	405	1215	+245	+735
2	159·5	319	638	4	82·5	330	1320	+174·75	+699
3	142·5	427·5	1282·5	5	36·5	182·5	912·5	+105·25	+526·25
4	77·5	310	1240	6	20	120	720	+ 71·5	+429
5	36	180	900	7	6·5	45·5	318·5	+ 25·25	+176·75
6	19·5	117	702	8	4·5	36	288	+ 14·5	+116
7	9·5	66·5	465·5						
8	4	32	256						
9	1	9	81						
	1376	+1659·5 −1179			1376	+1650·5 −1390		480·5	
	Total .	+480·5	9491·5		Total .	+260·5	10556·5	Total .	+5136·25
Correction for mean			−167·8	Correction for mean . . .			−49·3	Correction for mean .	−90·97
			9323·7				10507·2		
Sheppard's correction . . .			114·7	Sheppard's correction . .			114·7		+5045·28
			9209·0				10392·5		

The 9th column shows the total deviations of the daughter's height for each of the 18 columns in which the table is divided. When the numbers are small, these may usually be written down by inspection of the table. In the present case, where the numbers are large, and the entries are complicated by quartering, more care is required. The total of column 9 checks with that of the 3rd column. In order that it shall do so, the central entry +15.5, Which does not contribute to the products, has to be included. Each entry in the 9th column is multiplied by the paternal deviation to give the 10th column. In the present case all the entries in column 10 are positive; frequently both positive and negative entries occur, and it is then convenient to form a separate column for each. A useful check is afforded by repeating the work of the last two columns, interchanging the variates; we should then find the total deviation of the fathers for each array of daughters, and multiply by the daughters deviation. The uncorrected totals, 5136.25, should then agree. This check is especially useful with small tables, in which the work of the last two columns, carried out rapidly, is liable to error.

The value of the correlation coefficient, using Sheppard's correction, is found by dividing 5045.28 [p. 150] by the geometric mean of 9209.0 and 10,392.5; its value is +.5157· If Sheppard's correction had not been used, we should have obtained +.5097. The difference is in this case not large compared to the errors of random sampling, and the full effects on the distribution in random samples of using Sheppard's correction have never been fully examined, but there can be little doubt that Sheppard's correction should be used, and that its use gives generally an improved estimate of the correlation. On the other hand, the distribution in random samples of the uncorrected value is simpler and better understood, so that the uncorrected value should be used in tests of significance, in which the effect of correction need not, of course, be overlooked. For simplicity coarse grouping should be avoided where such tests are intended. The fact that with small samples the correlation obtained by the use of Sheppard's correction may exceed unity, illustrates the disturbance introduced into the random sampling distribution.

32. Partial Correlations

A great extension of the utility of the idea of correlation lies in its application to groups of more than two variates. In such cases, where the correlation between each pair of three variates is known, it is possible to eliminate any one of them, and so find what the correlation of the other two would be in a population selected so that the third variate was constant.

Ex. 26. *Elimination of age in organic correlations* [p. 151] *with growing children.* -- For example, it was found (Mumford and Young's data) in a group of boys of different ages, that the correlation of *standing height* with *chest girth* was +.836. One might expect that part of this association was due to general growth with age. It would be more desirable for many purposes to know the correlation between the variates for boys of a given age; but in fact only a few of the boys will be exactly of the same age, and even if we make age groups as broad as a year, we shall have in each group much fewer than the total number measured. In order to utilise the whole material, we only need to know the correlations of *standing height* with *age*, and of *chest girth* with *age*. These are given as .714 and .708.·

The fundamental formula in calculating partial correlation coefficients may be written

$$r_{12\cdot 3}=\frac{r_{12}-r_{13}\,r_{23}}{\sqrt{(1-r_{13}^{2})\,(1-r_{23}^{2})}}.$$

Here the three variates are numbered 1, 2, and 3, and we wish to find the correlation between 1 and 2, when 3 is eliminated; this is called the "partial" correlation between 1 and 2, and is designated by $r_{12.3}$, to show that variate 3 has been eliminated. The symbols r_{12}, r_{13}, r_{23}, indicate the correlations found directly between each pair of variates; these correlations being distinguished as "total" correlations.

Inserting the numerical values in the above formula we find $r_{12.3}$ =·.668, showing that when age is eliminated the correlation, though still considerable, [p. 152] has been markedly reduced. The mean value given by the above-mentioned authors for the correlations found by grouping the boys by years, is .653, not a greatly different value. In a similar manner, two or more variates may be eliminated in succession; thus with four variates, we may first eliminate variate 4, by thrice applying the above formula to find $r_{12.4}$, $r_{13.4}$, and $r_{23.4}$. Then applying the same formula again, to these three new values, we have

$$r_{12\cdot 34}=\frac{r_{12\cdot 4}-r_{13\cdot 4}\,r_{23\cdot 4}}{\sqrt{(1-r_{13\cdot 4}^{2})(1-r_{23\cdot 4}^{2})}}.$$

The labour increases rapidly with the number of variates to be eliminated. To eliminate *s* variates, the number of operations involved, each one application of the above formula is 1/6 $s(s+1)(s+2)$; for values of s from 1 to 6 this gives 1, 4, 10, 20, 35, 56 operations. Much of this labour may be saved by using tables of [sqrt]1-r^2 such as that published by J. R. Miner.[1]

The meaning of the correlation coefficient should be borne clearly in mind. The original aim to measure the "strength of heredity" by this method was based clearly on the supposition that the whole class of factors which tend to make relatives alike, in contrast to the unlikeness of unrelated persons, may be grouped together as heredity. That this is so for all practical purposes is, I believe, admitted, but the correlation does not tell us that this is so; it merely [p. 153] tells us the degree of resemblance in the actual population studied, between father and daughter. It tells us to what extent the height of the father is relevant information respecting the height of the daughter, or, otherwise interpreted, it tells us the relative importance of the factors which act alike upon the heights of father and daughter, compared to the totality of factors at work. If we know that B is caused by A, together with other factors independent of A, and that B has no influence on A, then the correlation between A and B does tell us how important, in relation to the other causes at work, is the influence of A. If we have not such knowledge, the correlation does not tell us whether A causes B, or B causes A, or whether both influences are at work, together with the effects of common causes.

This is true equally of partial correlations. If we know that a phenomenon A is not itself influential in determining certain other phenomena B, C, D, ..., but on the contrary is probably directly influenced by them, then the calculation of the partial correlations A with B, C, D, ... in each case eliminating the remaining values, will form a most valuable analysis of the causation of A. If on the contrary we choose a group of social phenomena with no antecedent knowledge of the causation or absence of causation among them, then the calculation of correlation coefficients, total or partial, will not advance us a step towards evaluating the importance of the causes at work.

The correlation between A and B measures, on a [p. 154] conventional scale, the importance of the factors which (on a balance of like and unlike action) act alike in both A and B, as against the remaining factors which affect A and B independently. If we eliminate a third variate C, we are removing from the comparison all those factors which become inoperative when C is fixed. If these are only those which affect A and B independently, then the correlation between A and B, whether positive or negative, will be numerically increased. We shall have eliminated irrelevant disturbing factors, and obtained, as it were, a better controlled experiment. We may also require to eliminate C if these factors act alike, or oppositely on the two variates correlated; in such a case the variability of C actually masks the effect we wish to investigate. Thirdly, C may be one of the chain of events by the mediation of which A affects B, or *vice versa.* The extent to which C is the channel through which the influence passes may be estimated by eliminating C; as one may demonstrate the small effect of latent factors in human heredity by finding the correlation of grandparent and grandchild, eliminating the intermediate parent. In no case, however, can we judge whether or not it is profitable to eliminate a certain variate unless we know, or are willing to assume, a qualitative scheme of causation. For the purely descriptive purpose of specifying a population in respect of a number of variates, either partial or total correlations are effective, and correlations of either type may be of interest.

As an illustration we may consider in what sense [p. 155] the coefficient of correlation does measure the "strength of heredity," assuming that heredity only is concerned in causing the resemblance between relatives; that is, that any environmental effects are distributed at haphazard. In the first place, we may note that if such environmental effects are increased in magnitude, the correlations would be reduced ; thus the same population, genetically speaking, would show higher correlations if reared under relatively uniform nutritional conditions, than they would if the nutritional conditions had been very diverse; although the genetical processes in the two cases were identical. Secondly, if environmental effects were at all influential (as in the population studied seems not to be indeed the case), we should obtain higher correlations from a mixed population of genetically very diverse strains, than we should from a more uniform population. Thirdly, although the influence of father on daughter is in a certain sense direct, in that the father contributes to the germinal composition of his daughter, we must not assume that this fact is necessarily the cause of the whole of the correlation; for it has been shown that husband and wife also show considerable resemblance in stature, and consequently taller fathers tend to have taller daughters partly because they choose, or are chosen by, taller wives. For this reason, for example, we should expect to find a noticeable positive correlation between step-fathers and step-daughters; also that, when the stature of the wife is eliminated, the partial correlation between father and daughter will be found to be lower than the total correlation. [p. 156] These considerations serve to some extent to define the sense in which the somewhat vague phrase, "strength of heredity," must be interpreted, in speaking of the correlation coefficient. It will readily be understood that, in less well understood cases, analogous considerations may be of some importance, and should if possible be critically considered.

33. Accuracy of the Correlation Coefficient

With large samples, and moderate or small correlations, the correlation obtained from a sample of *n* pairs of values is distributed normally about the true value r, with variance,

$$\frac{(1-\rho^2)^2}{n-1};$$

it is therefore usual to attach to an observed value *r*, a standard error $(1-r^2)/[\text{srqt}]n-1$, or $(1-r^2)/[\text{sqrt}]n$. This procedure is only valid under the restrictions stated above; with small samples the value of *r* is often very different from the true value, r, and the factor $1-r^2$, correspondingly in error; in addition the distribution of *r* is far from normal, so that tests of significance based on the above formula are often very deceptive. Since it is with small samples, less than 100, that the practical research worker ordinarily wishes to use the correlation coefficient, we shall give an account of more accurate methods of handling the results.

In all cases the procedure is alike for total and for partial correlations. Exact account may be taken of the differences in the distributions in the two cases, [p. 157] by deducting unity from the sample number for each variate eliminated; thus a partial correlation found by eliminating three variates, and based on data giving 13 values for each variate, is distributed exactly as is a total correlation based on 10 pairs of values.

34. The Significance of an Observed Correlation

In testing the significance of an observed correlation we require to calculate the probability that such a correlation should arise, by random sampling, from an uncorrelated population. If the probability is low we regard the correlation as significant. The table of *t* given at the end of the preceding chapter (p. 137) may be utilised to make an exact test. If *n'* be the number of pairs of observations on which the correlation is based, and *r* the correlation obtained, without using Sheppard's correction, then we take

$$t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n' - 2},$$

$$n = n' - 2,$$

and it may be demonstrated that the distribution of *t* so calculated, will agree with that given in the table.

It should be observed that this test, as is obviously necessary, is identical with that given in the last chapter for testing whether or not the linear regression coefficient differs significantly from zero.

TABLE V.A (p. 174) allows this test to be applied directly from the value of *r*, for samples up to 100 pairs of observations. Taking the four definite levels [p. 158] of significance, represented by P =·.10, .05, .02, and .01, the table shows for each value of *n*, from 1 to 20, and thence by larger intervals to 100, the corresponding values of *r*.

Ex. 27. *Significance of a correlation coefficient between* *autumn rainfall and wheat crop*. -- For the twenty years, 1885-1904, the mean wheat yield of Eastern England was found to be correlated with the autumn rainfall; the correlation found was -.629. Is this value significant? We obtain in succession

$$1-r^2 = \cdot 6044,$$
$$\sqrt{1-r^2} = \cdot 7774,$$
$$r/\sqrt{1-r^2} = -\cdot 8091,$$
$$t = -3\cdot 433.$$

For *n*=18, this shows that P is less than .01, and the correlation is definitely significant. The same conclusion may be read off at once from Table V.A entered with *n*=18.

If we had applied the standard error,

$$\sigma_r = \frac{1-r^2}{\sqrt{n'-1}},$$

we should have

$$t = \frac{r}{\sigma_r} = \frac{r}{1-r^2}\sqrt{n'-1} = 4\cdot 536,$$

a much greater value than the true one, very much exaggerating the significance. In addition, assuming that *r* was normally distributed (*n* = [infinity]), the significance of the result would between further exaggerated. This illustration will suffice to show how deceptive, in small samples, is the use of the standard error of the [p. 159] correlation coefficient, on the assumption that it will be normally distributed. Without this assumption the standard error is without utility. The misleading character of the formula is increased if *n'* is substituted for *n'*-1, as is often done. Judging from the normal deviate 4.536, we should suppose that the correlation obtained would be exceeded in random samples from uncorrelated material only 6 times in a million trials. Actually it would be exceeded about 3000 times in a million trials, or with 500 times the frequency supposed.

It is necessary to warn the student emphatically against the misleading character of the standard error of the correlation coefficient deduced from a small sample, because the principal utility of the correlation coefficient lies in its application to subjects of which little is known, and upon which the data are relatively scanty. With extensive material appropriate for biometrical investigations there is little danger of false conclusions being drawn, whereas with the comparatively few cases to which the experimenter must often look for guidance, the uncritical application of methods standardised in biometry, must be so frequently misleading as to endanger the credit of this most valuable weapon of research. It is not true, as the above example shows, that valid conclusions cannot be drawn from small samples; if accurate methods are used in calculating the probability, we thereby make full allowance for the size of the sample, and should be influenced in our judgment only by the value of the-probability indicated. The great increase of certainty which accrues from increasing data is [p. 160] reflected in the value of P, if accurate methods are used.

Ex. 28. *Significance of a partial correlation coefficient*. -- In a group of 32 poor law relief unions, Yule found that the percentage change from 1881 to 1891 in the percentage of the population in receipt of relief was correlated with the corresponding change in the ratio of the numbers given outdoor relief to the numbers relieved in the workhouse, when two other variates had been eliminated, namely, the corresponding changes in the percentage of the population over 65, and in the population itself.

The correlation found by Yule after eliminating the two variates was +.457; such a correlation is termed a partial correlation of the second order. Test its significance.

It has been demonstrated that the distribution in random samples of partial correlation coefficients may be derived from that of total correlation coefficients merely by deducting from the number of the sample, the number of variates eliminated. Deducting 2 from the 32 unions used, we have 30 as the effective number of the sample; hence

$$n=28$$

Calculating t from r as before, we find

$$t=2.719,$$

whence it appears from the table that P lies between .02 and .01. The correlation is therefore significant. This, of course, as in other cases, is on the assumption [p. 161] that the variates correlated (but not necessarily those eliminated) are normally distributed; economic variates seldom themselves give normal distributions, but the fact that we are here dealing with rates of change makes the assumption of normal distribution much more plausible. The values given in Table V.(A) for $n=25$, and $n=30$, give a sufficient indication of the level of *significance* attained by this observation.

35. Transformed Correlations

In addition to testing the significance of a correlation, to ascertain if there is any substantial evidence of association at all, it is also frequently required to perform one or more of the following operations, for each of which the standard error would be used in the case of a normally distributed quantity. With correlations derived from large samples the standard error may, therefore, be so used, except when the correlation approaches [plus or minus]1; but with small samples such as frequently occur in practice, special methods must be applied to obtain reliable results.

(i.) To test if an observed correlation differs significantly from a given theoretical value.

(ii.) To test if two observed correlations are significantly different.

(iii.) If a number of independent estimates of a correlation are available, to combine them into an improved estimate.

(iv.) To perform tests (i.) and (ii.) with such average values. [p. 162]

Problems of these kinds may be solved by a method analogous to that by which we have solved the problem of testing the significance of an observed correlation. In that case we were able from the given value r to calculate a quantity t which is distributed in a known manner, for which tables were available. The transformation led exactly to a distribution which had already been studied. The transformation which we shall now employ leads approximately to the normal distribution in which all the above tests may be carried out without difficulty. Let

$$z = \tfrac{1}{2}\{\log_e(1+r) - \log_e(1-r)\}$$

then as r changes from 0 to 1, z will pass from 0 to [infinity]. For small values of r, z is nearly equal to r, but as r approaches unity, z increases without limit. For negative values of r, z is negative. The advantage of this transformation lies in the distribution of the two quantities in random samples. The standard deviation of *r* depends on the true value of the correlation, r; as is seen from the formula

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{n' - 1}}.$$

Since r is unknown, we have to substitute for it the observed value *r*, and this value will not, in small samples, be a very accurate estimate of r. The standard error of *z* is simpler in form,

$$\sigma_z = \frac{1}{\sqrt{n' - 3}},$$

and is practically independent of the value of the [p. 163] correlation in the population from which the sample is drawn.

In the second place the distribution of *r* is skew in small samples, and even for large samples it remains very skew for high correlations. The distribution of *z* is not strictly normal, but it tends to normality rapidly as the sample is increased, whatever may be the value of the correlation. We shall give examples to test the effect of the departure of the *z* distribution from normality.

Finally the distribution of *r* changes its form rapidly as r is changed ; consequently no attempt can be made, with reasonable hope of success, to allow for the skewness of the distribution. On the contrary, the distribution of *z* is nearly constant in form, and the accuracy of tests may be improved by small corrections for skewness; such corrections are, however, in any case somewhat laborious, and we shall not deal with them. The simple assumption that *z* is normally distributed will in all ordinary cases be sufficiently accurate.

These three advantages of the transformation from *r* to *z* may be seen by comparing Figs. 7 and 8. In Fig. 7 are shown the actual distributions of *r*, for 8 pairs of observations, from populations having correlations 0 and 0.8; Fig. 8 shows the corresponding distribution curves for *z*. The two curves in Fig. 7 are widely different in their modal heights; both are distinctly non-normal curves; in form also they are strongly contrasted, the one being symmetrical, the other highly unsymmetrical. On the contrary, in [p. 164] Fig. 8 the two curves do not differ greatly in height; although not exactly normal in form, they come so close to it, even for a small sample of 8 pairs of observations, [p. 165] that the eye cannot detect the difference; and this approximate normality holds up to the extreme limits r=[plus or minus]1. One additional feature is brought out by Fig. 8 ; in the distribution for r=0.8, although the curve itself is as symmetrical as the eye can judge of, yet the ordinate of zero error is not centrally placed. The figure, in fact, reveals the small bias which is introduced into the estimate of the correlation coefficient as ordinarily calculated; we shall treat further of this bias in the next section, and in the following chapter shall deal with a similar bias introduced in the calculation of intraclass correlations.

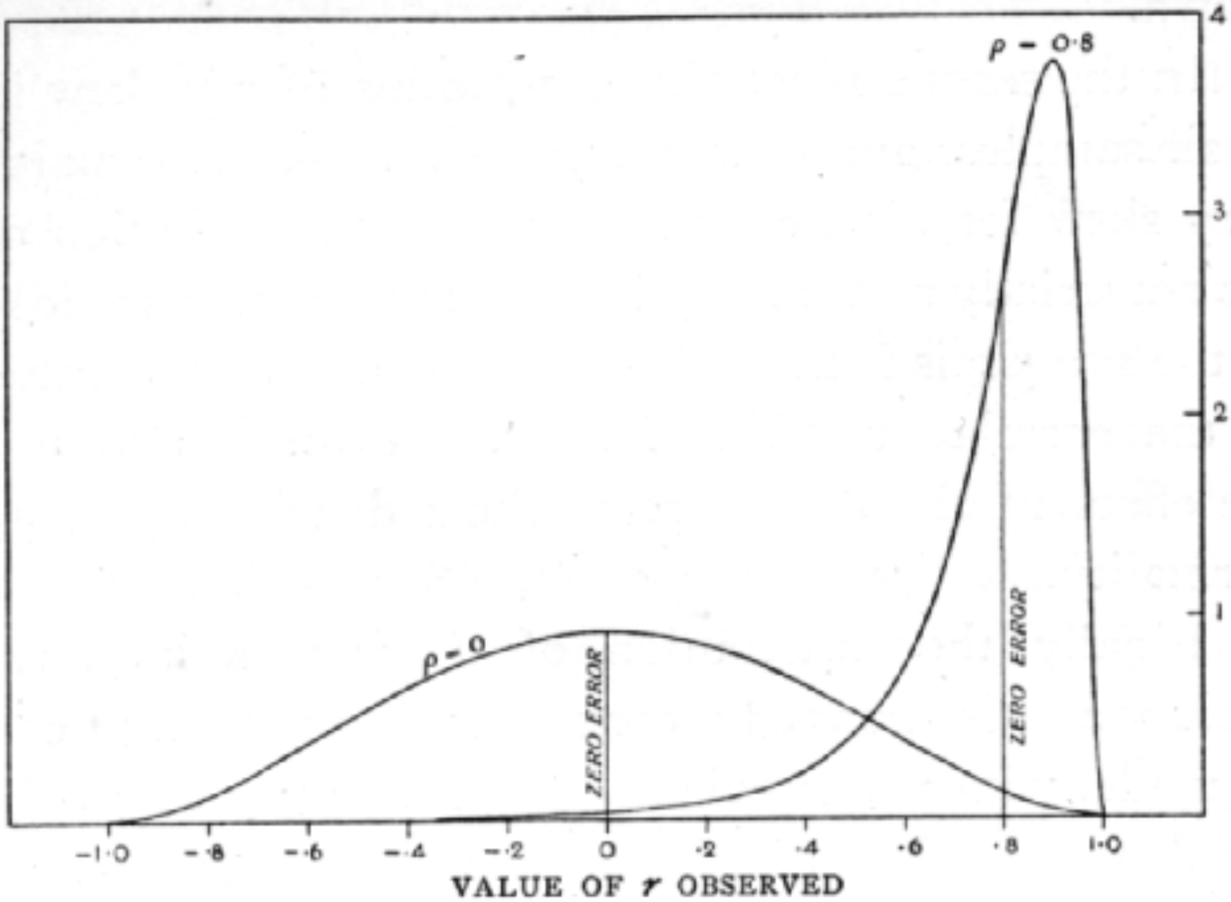

FIG. 7.

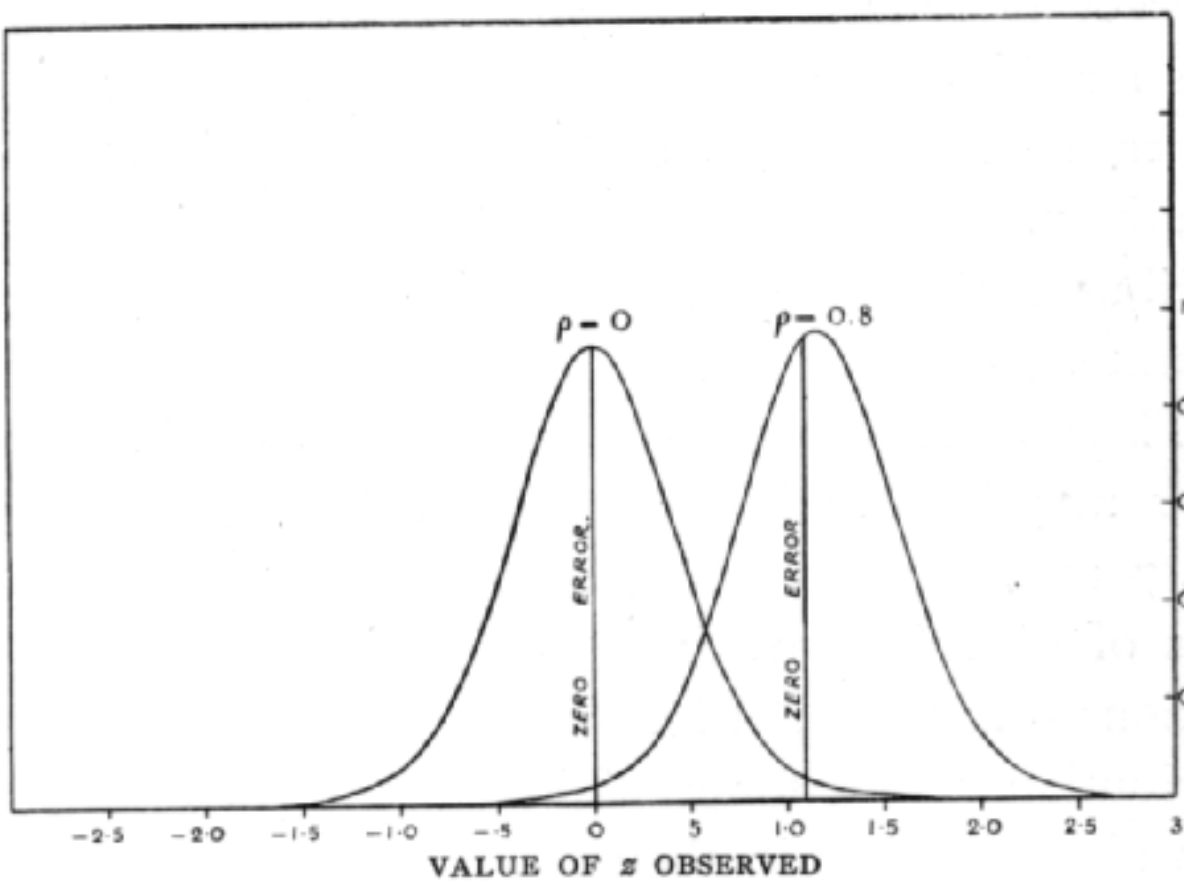

FIG. 8.

To facilitate the transformation we give in Table V.(B) (p. 175) the values of *r* corresponding to values of *z*, proceeding by intervals of ,01, from 0 to 3. In the earlier part of this table it will be seen that the values of *r* and z do not differ greatly; but with higher correlations small changes in *r* correspond to relatively large changes in *z*. In fact, measured on the *z*-scale, a correlation of .99 differs from a correlation .95 by more than a correlation .6 exceeds zero. The values of *z* give a truer picture of the relative importance of correlations of different sizes, than do the values of *r*.

To find the value of *z* corresponding to a given value of *r*, say .6, the entries in the table lying on either side of .6, are first found, whence we see at once that *z* lies between .69 and .70; the interval between these entries is then divided proportionately to find the fraction to be added to .69. In this case we have 20/64, or .31, so that *z*=.6931. Similarly, in finding [p. 166] the value of *r* corresponding to any value of *z*, say .9218, we see at once that it lies between .7259 and .7306; the difference is 47, and 18 per cent of this gives 8 to be added to the former value, giving us finally *r*=.7267. The same table may thus be used to transform *r* into *z*, and to reverse the process.

Ex. 29. *Test of the approximate normality of the distribution of z*. -- In order to illustrate the kind of accuracy obtainable by the use of *z*, let us take the case that has already been treated by an exact method in Ex. 26. A correlation of -.629 has been obtained from 20 pairs of observations ; test its significance.

For r=-.629 we have, using either a table of natural logarithms, or the special table for z, z=.7398. To divide this by its standard error is equivalent to multiplying it by [sqrt]17.· This gives -3.050, which we interpret as a normal deviate. From the table of normal deviates it appears that this value will be exceeded about 23 times in 10,000 trials. The true frequency, as we have seen, is about 30 times in 10,000 trials. The error tends slightly to exaggerate the significance of the result.

Ex. 30.· *Further test of the normality of the distribution of z*. -- A partial correlation +.457 was obtained from a sample of 32, after eliminating two variates. Does this differ significantly from zero? Here z =.4935; deducting the two eliminated variates the effective size of the sample is 30, and the standard error of z is 1/[sqrt]27; multiplying z by [sqrt]27, we have as a. normal variate 2.564. Table IV. shows, as before, that P is just over ·.01. There is a slight exaggeration [p. 167] of significance, but it is even slighter than in the previous example.

The above examples show that the z transformation will give a variate which, for most practical purposes, may be taken to be normally distributed. In the case of simple tests of significance the use of the table of t is to be preferred ; in the following examples this method is not available, and the only method available which is both tolerably accurate and sufficiently rapid for practical use lies in the use of z.

Ex. 31· *Significance of deviation from expectation of an observed correlation coefficient*. -- In a sample of 25 pairs of parent and child the correlation was found to be .60. Is this value consistent with the view that the true correlation in that character was .46?

The first step is to find the difference of the corresponding values of *z*. This is shown below:

TABLE 33

	r.	*z*.
Sample value . . .	·60	·6931
Population value . .	·46	·4973
Difference . . .		·1958

To obtain the normal deviate we multiply by [sqrt]22, and obtain .918. The deviation is less than the standard deviation, and the value obtained is therefore quite in accordance with the hypothesis. [p. 168]

Ex. 32. *Significance of difference between two observed correlations*. -- Of two samples the first, of 20 pairs, gives a correlation .6, the second, of 25 pairs, gives a correlation .8: are these values significantly different?

In this case we require not only the difference of the values of *z*, but the standard error of the difference. The variance of the difference is the sum of the reciprocals of 17 and 22; the work is shown below:

TABLE 34

	r.	z.	$n'-3$.	Reciprocal.
1st sample .	·60	·6931	17	·05882
2nd sample .	·80	1·0986	22	·04545
Difference .		·4055 ± ·3230	Sum .	·10427

The standard error which is appended to the difference of the values of z is the square root of the variance found on the same line. The difference does not exceed twice the standard error, and cannot therefore be judged significant. There is thus no sufficient evidence to conclude that the two samples are not drawn from equally correlated populations.

Ex. 33· *Combination of values from small samples.* -- Assuming that the two samples in the last example were drawn from equally correlated populations, estimate the value of the correlation.

The two values of z must be given weight inversely proportional to their variance. We therefore [p. 169] multiply the first by 17, the second by 22 and add, dividing the total by 39. This gives an estimated value of z for the population, and the corresponding value of r may be found from the table.

TABLE 35

	r.	z.	$n-3$.	$(n-3)z$.
1st sample . . .	·60	·6930	17	11·7810
2nd sample . . .	·80	1·0986	22	24·1692
	·7267	·9218	39	35·9502

The weighted average value of *z* is .9218, to which corresponds the value *r*=.7267; the value of *z* so obtained may be regarded as subject to normally distributed errors of random sampling with variance equal to 1/39· The accuracy is therefore equivalent to that of a single value obtained from 42 pairs of observations. Tests of significance may thus be applied to such averaged values of *z*, as to individual values.

36. Systematic Errors

In connexion with the averaging of correlations obtained from small samples it is worth while to consider the effects of two classes of systematic errors, which, although of little or no importance when single values only are available, become of increasing importance as larger numbers of samples are averaged.

The value of *z* obtained from any sample is an estimate of a true value, *r*, belonging to the sampled [p. 170] population, just as the value of *r* obtained from a sample is an estimate of a population value, r. If the method of obtaining the correlation were free from bias, the values of *z* would be normally distributed about a mean *z*[bar], which would agree in value with z. Actually there is a small bias which makes the mean value of *z* somewhat greater numerically than z; thus the correlation, whether positive or negative, is slightly exaggerated. This bias may effectively be corrected by subtracting from the value of *z* the correction

$$\frac{\rho}{2(n'-1)}.$$

For single samples this correction is unimportant, being small compared to the standard error of *z*. For example, if *n'*=10, the standard error of *z* is .378, while the correction is r/l8 and cannot exceed .056. If, however, *z*[bar] were the mean of 1000 such values of *z*, derived from samples of 10, the standard error of *z*[bar] is only .012, and the correction, which is unaltered by taking the mean, may become of great importance.

The second type of systematic error is that introduced by neglecting Sheppard's correction. In calculating the value of *z*, we must always take the value of *r* found without using Sheppard's correction, since the latter complicates the distribution.

But the omission of Sheppard's correction introduces a systematic error, in the opposite direction to that mentioned above; and which, though normally very small, appears in large as well as in small samples. In case of averaging the correlations from a number of [p. 171] coarsely grouped small samples, the average z should be obtained from values of r found without Sheppard's correction, and to the result a correction, representing the average effect of Sheppard's correction, may be applied.

37. Correlation between Series

The extremely useful case in which it is required to find the correlation between two series of quantities, such as annual figures, arranged in order at equal intervals of time, is in reality a case of partial correlation, although it may be treated more directly by the method of fitting curved regression lines given in the last chapter (p. 128).

If, for example, we had a record of the number of deaths from a certain disease for successive years, and wished to study if this mortality were associated with meteorological conditions, or the incidence of some other disease, or the mortality of some other age group, the outstanding difficulty in the direct application of the correlation coefficient is that the number of deaths considered probably exhibits a progressive change during the period available. Such changes may be due to changes in the population among which the deaths occur, whether it be the total population of a district, or that of a particular age group, or to changes in the sanitary conditions in which the population lives, or in the skill and availability of medical assistance, or to changes in the racial or genetic composition of the population. In any case it is usually found that the changes are still apparent [p. 172] when the number of deaths is converted into a death-rate on the existing population in each year, by which means one of the direct effects of changing population is eliminated.

If the progressive change could be represented effectively by a straight line it would be sufficient to consider the time as a third variate, and to eliminate it by calculating the corresponding partial correlation coefficient. Usually, however, the change is not so simple, and would need an expression involving the square and higher powers of the time adequately to represent it. The partial correlation required is one found by eliminating not only t, but t^2, t^3, t^4, ..., regarding these as separate variates; for if we have eliminated all of these up to (say) the fourth degree, we have incidentally eliminated from the correlation any function of the time *of* the fourth degree, including that by which the progressive change is best represented.

This partial correlation may be calculated directly from the coefficients of the regression function obtained as in the last chapter (p. 128). If y and y' are the two quantities to be correlated, we obtain for y the coefficients A, B, C,..., and for y' the corresponding coefficients A', B', C', .. .; the sum of the squares of the deviations of the variates from the curved regression lines are obtained as before, from the equations

$$S(y-Y)^2 = S(y^2) - n'A^2 - \frac{n'(n'^2-1)}{12}B^2 - \ldots,$$

$$S(y'-Y')^2 = S(y'^2) - n'A'^2 - \frac{n'(n'^2-1)}{12}B'^2 - \ldots;$$

[p. 173]

while the sum of the products may be obtained from the similar equation

$$S\{(y-Y)(y'-Y')\} = S(yy') - n'AA' - \frac{n'(n'^2-1)}{12}BB' - \ldots,$$

the required partial correlation being, then,

$$r = \frac{S\{(y-Y)(y'-Y')\}}{\sqrt{S(y-Y)^2 \,.\, S(y'-Y')^2}}$$

In this process the number of variates eliminated is equal to the degree of t to which the fitting has been carried; it will be understood that both variates must be fitted to the same degree, even if one of them is capable of adequate representation by a curve of lower degree than is the other. [p. 174]

TABLE V. (A).—VALUES OF THE CORRELATION COEFFICIENT FOR DIFFERENT LEVELS OF SIGNIFICANCE

n	P = ·1.	·05.	·02.	·01.
1	·98769	·996917	·9995066	·9998766
2	·90000	·95000	·98000	·990000
3	·8054	·8783	·93433	·95873
4	·7293	·8114	·8822	·91720
5	·6694	·7545	·8329	·8745
6	·6215	·7067	·7887	·8343
7	·5822	·6664	·7498	·7977
8	·5494	·6319	·7155	·7646
9	·5214	·6021	·6851	·7348
10	·4973	·5760	·6581	·7079
11	·4762	·5529	·6339	·6835
12	·4575	·5324	·6120	·6614
13	·4409	·5139	·5923	·6411
14	·4259	·4973	·5742	·6226
15	·4124	·4821	·5577	·6055
16	·4000	·4683	·5425	·5897
17	·3887	·4555	·5285	·5751
18	·3783	·4438	·5155	·5614
19	·3687	·4329	·5034	·5487
20	·3598	·4227	·4921	·5368
25	·3233	·3809	·4451	·4869
30	·2960	·3494	·4093	·4487
35	·2746	·3246	·3810	·4182
40	·2573	·3044	·3578	·3932
45	·2428	·2875	·3384	·3721
50	·2306	·2732	·3218	·3541
60	·2108	·2500	·2948	·3248
70	·1954	·2319	·2737	·3017
80	·1829	·2172	·2565	·2830
90	·1726	·2050	·2422	·2673
100	·1638	·1946	·2301	·2540

For a total correlation n is 2 less than the number of pairs in the sample; for a partial correlation the number of eliminated variates also should be subtracted.

TABLE V. (B).—TABLE OF r, FOR VALUES OF z FROM 0 TO 3

z.	·01.	·02.	·03.	·04.	·05.	·06.	·07.	·08.	·09.	·10.
·0	·0100	·0200	·0300	·0400	·0500	·0599	·0699	·0798	·0898	·0997
·1	·1096	·1194	·1293	·1391	·1489	·1586	·1684	·1781	·1877	·1974
·2	·2070	·2165	·2260	·2355	·2449	·2543	·2636	·2729	·2821	·2913
·3	·3004	·3095	·3185	·3275	·3364	·3452	·3540	·3627	·3714	·3800
·4	·3885	·3969	·4053	·4136	·4219	·4301	·4382	·4462	·4542	·4621
·5	·4699	·4777	·4854	·4930	·5005	·5080	·5154	·5227	·5299	·5370
·6	·5441	·5511	·5580	·5649	·5717	·5784	·5850	·5915	·5980	·6044
·7	·6107	·6169	·6231	·6291	·6351	·6411	·6469	·6527	·6584	·6640
·8	·6696	·6751	·6805	·6858	·6911	·6963	·7014	·7064	·7114	·7163
·9	·7211	·7259	·7306	·7352	·7398	·7443	·7487	·7531	·7574	·7616
1·0	·7658	·7699	·7739	·7779	·7818	·7857	·7895	·7932	·7969	·8005
1·1	·8041	·8076	·8110	·8144	·8178	·8210	·8243	·8275	·8306	·8337
1·2	·8367	·8397	·8426	·8455	·8483	·8511	·8538	·8565	·8591	·8617
1·3	·8643	·8668	·8692	·8717	·8741	·8764	·8787	·8810	·8832	·8854
1·4	·8875	·8896	·8917	·8937	·8957	·8977	·8996	·9015	·9033	·9051
1·5	·9069	·9087	·9104	·9121	·9138	·9154	·9170	·9186	·9201	·9217
1·6	·9232	·9246	·9261	·9275	·9289	·9302	·9316	·9329	·9341	·9354
1·7	·9366	·9379	·9391	·9402	·9414	·9425	·9436	·9447	·9458	·94681
1·8	·94783	·94884	·94983	·95080	·95175	·95268	·95359	·95449	·95537	·95624
1·9	·95709	·95792	·95873	·95953	·96032	·96109	·96185	·96259	·96331	·96403
2·0	·96473	·96541	·96609	·96675	·96739	·96803	·96865	·96926	·96986	·97045
2·1	·97103	·97159	·97215	·97269	·97323	·97375	·97426	·97477	·97526	·97574
2·2	·97622	·97668	·97714	·97759	·97803	·97846	·97888	·97929	·97970	·98010
2·3	·98049	·98087	·98124	·98161	·98197	·98233	·98267	·98301	·98335	·98367
2·4	·98399	·98431	·98462	·98492	·98522	·98551	·98579	·98607	·98635	·98661
2·5	·98688	·98714	·98739	·98764	·98788	·98812	·98835	·98858	·98881	·98903
2·6	·98924	·98945	·98966	·98987	·99007	·99026	·99045	·99064	·99083	·99101
2·7	·99118	·99136	·99153	·99170	·99186	·99202	·99218	·99233	·99248	·99263
2·8	·99278	·99292	·99306	·99320	·99333	·99346	·99359	·99372	·99384	·99396
2·9	·99408	·99420	·99431	·99443	·99454	·99464	·99475	·99485	·99495	·99505

For greater accuracy, and for values beyond the table,

$$r = (e^{2z} - 1)/(e^{2z} + 1);$$

$$z = \tfrac{1}{2}\{\log(1 + r) - \log(1 - r)\}.$$

Footnotes

[1] *Tables of* [sqrt]*1-r^2 for Use in Partial Correlation, and in Trigonometry*. Johns Hopkins Press, 1922.

Classics in the History of Psychology

An internet resource developed by
Christopher D. Green
York University, Toronto, Ontario
ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted April 2000

VI

THE CORRELATION COEFFICIENT

30. No quantity is more characteristic of modern statistical work than the correlation coefficient, and no method has been applied successfully to such various data as the method of correlation. Observational data in particular, in cases where we can observe the occurrence of various possible contributory causes of a phenomenon, but cannot control them, has been given by its means an altogether new importance. In experimental work proper its position is much less central; it will be found useful in the exploratory stages of an enquiry, as when two factors which had been thought independent appear to be associated in their occurrence; but it is seldom, with controlled experimental conditions, that it is desired to express our conclusion in the form of a correlation coefficient.

One of the earliest and most striking successes of the method of correlation was in the biometrical study of inheritance. At a time when nothing was known of the mechanism of inheritance, or of the structure of the germinal material, it was possible by this method to demonstrate the existence of inheritance, and to [p. 139] "measure its intensity"; and this in an organism in which experimental breeding could not be practised, namely, Man. By comparison of the results obtained from the physical measurements in man with those obtained from other organisms, it was established that man's nature is not less governed by heredity than that of the rest of the animate world. The scope of the analogy was further widened by demonstrating that correlation coefficients of the same magnitude were obtained for the mental and moral qualities in man as for the physical measurements.

These results are still of fundamental importance, for not only is inheritance in man still incapable of experimental study, and existing methods of mental testing are still unable to analyse the mental disposition, but even with organisms suitable for experiment and measurement, it is only in the most favourable cases that the several factors causing fluctuating variability can be resolved, and their effects studied, by Mendelian methods. Such fluctuating variability, with an approximately normal distribution, is characteristic of the majority of the useful qualities of domestic plants and animals; and although there is strong reason to think that inheritance in such cases is ultimately Mendelian, the biometrical method of study is at present alone capable of holding out hopes of immediate progress.

We give in Table 31 an example of a correlation table. It consists of a record in compact form of the stature of 1376 fathers and daughters. (Pearson and Lee's data.) The measurements are grouped in [p. 140-141] [table] [p. 142] inches, and those whose measurement was recorded as an integral number of inches have been split; thus a father recorded as of 67 inches would appear as 1/2 under 66.5 and 1/2 under 67.5. Similarly with the daughters; in consequence, when both measurements are whole numbers the case appears in four quarters. This gives the table a confusing appearance, since the majority of entries are fractional, although they represent frequencies. It is preferable, if bias in measurement can be avoided, to group the observations in such a way that each possible observation lies wholly within one group.

TABLE 31

Height of Daughters in Inches.	Height of Fathers in Inches.																		
	58·5	59·5	60·5	61·5	62·5	63·5	64·5	65·5	66·5	67·5	68·5	69·5	70·5	71·5	72·5	73·5	74·5	75·5	Total.
52·5	..	..	..	..	·25	·25	..	..	..	..	..	..	..	..	..	..	..	..	·5
53·5	..	..	..	..	·25	·25	..	..	..	..	..	..	..	..	..	..	..	..	·5
54·5	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..	..
55·5	..	..	..	..	..	..	..	1	..	..	..	..	..	..	..	..	..	..	1
56·5	·25	·25	..	·25	1·25	·5	..	1	·5	·5	..	..	..	..	..	..	..	..	4·5
57·5	·25	·25	·5	1·5	4·5	1	1·5	1·5	2·5	..	·5	·5	..	..	..	..	..	..	14·5
58·5	·25	·75	·5	·75	·75	1	1·75	1·25	5	2·75	·5	·25	..	..	..	..	..	..	15·5
59·5	·5	1	2	..	6	4·75	5	6·25	11·75	3·5	3·5	2	1·75	·5	..	..	..	..	48·5
60·5	·75	·75	..	2·5	8	6·25	12·5	18·25	20·25	11	9	4·75	2·5	1·25	1·25	..	..	..	99
61·5	..	·5	1·75	2	9·75	11·5	13	23·75	23·75	20·25	16·5	10·25	4·25	3	1·25	..	..	..	141·5
62·5	..	1	2·25	2	4·5	12	22·75	26	33	28·25	24·75	14·25	13·75	4·75	·75	·5	..	..	190·5
63·5	..	..	·25	2	6	8·25	11	27·25	35·75	37·25	31·5	26·25	16·25	7·75	1·5	·75	·25	..	212
64·5	..	..	·25	2·5	1·75	3·25	9·25	23	18·75	28·5	33	34·25	24·5	11·75	5·5	1	·25	1	198·5
65·5	..	..	..	·5	1	·5	11	12·25	9·25	19·75	30	26·5	22·25	15	4·75	3·75	2	1	159·5
66·5	..	..	..	·5	·5	1·5	3·25	7·25	8·75	16	26·25	26·75	20·5	18·5	7·75	4·25	·25	·5	142·5
67·5	..	..	..	..	..	..	1	5·75	7	4	14·25	13·25	12	11·25	4·5	3·75	·75	..	77·5
68·5	..	..	..	..	·25	·25	·25	·25	1·5	3	5·5	4·25	5·75	5·25	3·75	2·5	1·5	2	36
69·5	..	..	..	..	·25	·25	·25	·25	·25	·25	1	2·5	6·5	2·25	2·75	2	1	..	19·5
70·5	..	..	..	..	..	..	..	..	..	..	1·75	·25	4·5	·75	1·25	·75	·25	..	9·5
71·5	..	..	..	..	..	..	..	..	..	..	·5	..	·5	·5	1·5	·75	·25	..	4
72·5	..	..	..	..	..	..	..	..	..	..	1	..	..	..	..	..	..	..	1
Total	2	4·5	7·5	14·5	45	51·5	92·5	155	178	175	199·5	166	135	82·5	36·5	20	6·5	4·5	1376

The most obvious feature *of* the table is that cases do not occur in which the father is very tall and the daughter very short, and *vice versa* ; the upper right-hand and lower left-hand corners of the table are blank, so that we may conclude that such occurrences are too rare to occur in a sample of about 1400 cases. The observations recorded lie in a roughly elliptical figure lying diagonally across the table. If we mark out the region in which the frequencies exceed 10 it appears that this region, apart from natural irregularities, is similar, and similarly situated. The frequency of occurrence increases from all sides to the central region of the table, where a few frequencies over 30 may be seen. The lines of equal frequency are roughly similar and similarly situated ellipses. In the outer zone observations occur only occasionally, and therefore irregularly; beyond this we could only explore by taking a much larger sample.

The table has been divided into four quadrants by [p. 143] marking out central values of the two variates; these values, 67.5 inches for the fathers and 63.5 inches for the daughters, are near the means. When the table is so divided it is obvious that the lower right-hand and upper left-hand quadrants are distinctly more populous than the other two; not only are more squares occupied, but the frequencies are higher. It is apparent that tall men have tall daughters more frequently than the short men, and *vice versa*. The method of correlation aims at measuring the degree to which this association exists.

The marginal totals show the frequency distributions of the fathers and the daughters respectively. These are both approximately normal distributions, as is frequently the case with biometrical data collected without selection. This marks a frequent difference between biometrical and experimental data. An experimenter would perhaps have bred from two contrasted groups of fathers of, for example, 63 and 72 inches in height; all his fathers would then belong to these two classes, and the correlation coefficient, if used, would be almost meaningless. Such an experiment would serve to ascertain the regression of daughter's height in father's height, and so to determine the effect on the daughters of selection applied to the fathers, but it would not give us the correlation coefficient which is a descriptive observational feature of the population as it is, and may be wholly vitiated by selection.

Just as normal variation with one variate may be specified by a frequency formula in which the [p. 144] logarithm of the frequency is a quadratic function of the variate, so with two variates the frequency may be expressible in terms of a quadratic function of the values of the two variates. We then have a normal correlation surface, for which the frequency may conveniently be written in the form

$$df = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{x^2}{\sigma_1^2} - \frac{2\rho xy}{\sigma_1\sigma_2} + \frac{y^2}{\sigma_2^2}\right\}} dxdy.$$

In this expression *x* and *y* are the deviations of the two variates from their means, s_1 and s_2 are the two standard deviations, and r is the *correlation* between *x* and *y.* The correlation in the above expression may be positive or negative, but cannot exceed unity in magnitude; it is a pure number without physical dimensions. If r=0, the expression for the frequency degenerates into the product of the two factors

$$\frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma_1^2}} dx \cdot \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{y^2}{2\sigma_2^2}} dy,$$

showing that the limit of the normal correlation surface, when the correlation vanishes, is merely that of two normally distributed variates varying in complete independence. At the other extreme, when *p* is +1 or -1, the variation of the two variates is in strict proportion, so that the value of either may be calculated accurately from that of the other. In other words, we cease strictly to have two variates, but merely two measures of the same variable quantity.

If we pick out the cases in which one variate has an assigned value, we have what is termed an array; [p. 145] the columns and rows of the table may, except as regards variation within the group limits, be regarded as arrays. With normal correlation the variation within an array may be obtained from the general formula, by giving *x* a constant value, (say) *a*, and dividing by the total frequency with which this value occurs; then we have

$$df = \frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-\rho^2}} \cdot e^{-\frac{1}{2(1-\rho^2)\sigma_2^2}\left(y-\rho\frac{a\sigma_2}{\sigma_1}\right)^2},$$

showing (i.) that the variation of *y* within the array is normal ; (ii.) that the mean value of *y* for that array is ras_2/s_1, so that the regression of *y* on *x* is linear, with regression coefficient

$$\rho\frac{\sigma_2}{\sigma_1};$$

and (iii.) that the variance of *y* within the array is $s_2^2(1-r^2)$, and is the same within each array. We may express this by saying that of the total variance of *y* the fraction $(1-r^2)$ is independent of *x*, while the remaining fraction, r^2, is determined by, or calculable from, the value of *x*.

These relations are reciprocal, the regression of *x* on *y* is linear, with regression coefficient rs_1/s_2; the correlation r is thus the geometric mean of the two regressions. The two regression lines representing the mean value of *x* for given *y*, and the mean value of *y* for given *x*, cannot coincide unless r=[plus or minus]1. The variation of *x* within an array in which *y* is fixed, is normal with variance equal to $s_1^2(1-r^2)$, so that we may say that of the variance of *x* the fraction $(1-r^2)$ [p. 146] is independent of *y*, and the remaining fraction, r^2, is determined by, or calculable from, the value of *y*.

Such are the formal mathematical consequences of normal correlation. Much biometric data certainly shows a general agreement with the features to be expected on this assumption; though I am not aware that the question has been subjected to any sufficiently critical enquiry. Approximate agreement is perhaps all that is needed to justify the use of the correlation as a quantity descriptive of the population; its efficacy in this respect is undoubted, and it is not improbable that in some cases it affords a complete description of the simultaneous variation of the variates.

31. The Statistical Estimation of the Correlation

Just as the mean and the standard deviation of a normal population in one variate may be most satisfactorily estimated from the first two moments of the observed distribution, so the only satisfactory estimate of the correlation, when the variates are normally correlated, is found from the "product moment." If x and y represent the deviations of the two variates from their means, we calculate the three statistics s_1, s_2, r by the three equations

$$ns_1^2 = \mathrm{S}(x^2),\ ns_2^2 = \mathrm{S}(y^2),\ nrs_1s_2 = \mathrm{S}(xy);$$

then s_1 and s_2 are estimates of the standard deviations s_1, and s_2, and r is an estimate of the correlation r. Such an estimate is called the *correlation coefficient*, or the *product moment correlation*, the latter term [p. 147] referring to the summation of the product terms, xy, in the last equation.

The above method of calculation might have been derived from the consideration that the correlation of the population is the geometric mean of the two regression coefficients; for our estimates of these two regressions would be

$$\frac{\mathrm{S}(xy)}{\mathrm{S}(x^2)} \text{ and } \frac{\mathrm{S}(xy)}{\mathrm{S}(y^2)},$$

so that it is in accordance with these estimates to take as our estimate of r

$$r = \frac{S(xy)}{\sqrt{S(x^2) \cdot S(y^2)}},$$

which is in fact the product moment correlation.

Ex. 25. *Parental correlation in stature.* -- The numerical work required to calculate the correlation coefficient is shown below in Table 32.

The first eight columns require no explanation, since they merely repeat the usual process of finding the mean and standard deviation of the two marginal distributions. It is not necessary actually to find the mean, by dividing the total of the third column, 480.5, by 1376, since we may work all through with the undivided totals. The correction for the fact that our working mean is not the true mean is performed by subtracting $(480.5)^2/1376$ in the 4th column; a similar correction appears at the foot of the 8^{th} column, and at the foot of the last column. The correction for the sum of products is performed by subtracting 4805x2605/1376· This correction of [p. 148] [table] [p. 149] the product term may be positive or negative; if the total deviations of the two variates are of opposite sign, the correction must be added. The sum of squares, with and without Sheppard's correction (1376/12), are shown separately; there is no corresponding correction to be made to the product term.

TABLE 32

Daughters.				Fathers.				Total for Daughters.	Product.
Deviation.	Frequency.			Deviation.	Frequency.				
−11	·5	5·5	60·5	−9	2	18	162	− 8·75	+ 78·75
−10	·5	5	50	−8	4·5	36	288	− 15·25	+122
− 9	–	–	–	−7	7·5	52·5	367·5	− 19	+133
− 8	1	8	64	−6	14·5	87	522	− 23	+138
− 7	4·5	31·5	220·5	−5	45	225	1125	−108·75	+543·75
− 6	14·5	87	522	−4	51·5	206	824	− 81	+324
− 5	15·5	77·5	387·5	−3	92·5	277·5	832·5	− 76·25	+228·75
− 4	48·5	194	776	−2	155	310	620	− 88·50	+177
− 3	99	297	891	−1	178	178	178	−131·25	+131·25
− 2	141·5	283	566	0	175	1390	..	+ 15·5	..
− 1	190·5	190·5	190·5	1	199·5	199·5	199·5	+183·25	+183·25
0	212	−1179	..	2	166	332	664	+197·25	+394·5
1	198·5	198·5	198·5	3	135	405	1215	+245	+735
2	159·5	319	638	4	82·5	330	1320	+174·75	+699
3	142·5	427·5	1282·5	5	36·5	182·5	912·5	+105·25	+526·25
4	77·5	310	1240	6	20	120	720	+ 71·5	+429
5	36	180	900	7	6·5	45·5	318·5	+ 25·25	+176·75
6	19·5	117	702	8	4·5	36	288	+ 14·5	+116
7	9·5	66·5	465·5						
8	4	32	256						
9	1	9	81						
	1376	+1659·5 −1179			1376	+1650·5 −1390		480·5	
	Total .	+480·5	9491·5		Total .	+260·5	10556·5	Total .	+5136·25
Correction for mean			−167·8	Correction for mean . . .			−49·3	Correction for mean .	−90·97
			9323·7				10507·2		
Sheppard's correction . . .			114·7	Sheppard's correction . .			114·7		+5045·28
			9209·0				10392·5		

The 9th column shows the total deviations of the daughter's height for each of the 18 columns in which the table is divided. When the numbers are small, these may usually be written down by inspection of the table. In the present case, where the numbers are large, and the entries are complicated by quartering, more care is required. The total of column 9 checks with that of the 3rd column. In order that it shall do so, the central entry +15.5, Which does not contribute to the products, has to be included. Each entry in the 9th column is multiplied by the paternal deviation to give the 10th column. In the present case all the entries in column 10 are positive; frequently both positive and negative entries occur, and it is then convenient to form a separate column for each. A useful check is afforded by repeating the work of the last two columns, interchanging the variates; we should then find the total deviation of the fathers for each array of daughters, and multiply by the daughters deviation. The uncorrected totals, 5136.25, should then agree. This check is especially useful with small tables, in which the work of the last two columns, carried out rapidly, is liable to error.

The value of the correlation coefficient, using Sheppard's correction, is found by dividing 5045.28 [p. 150] by the geometric mean of 9209.0 and 10,392.5; its value is +.5157· If Sheppard's correction had not been used, we should have obtained +.5097. The difference is in this case not large compared to the errors of random sampling, and the full effects on the distribution in random samples of using Sheppard's correction have never been fully examined, but there can be little doubt that Sheppard's correction should be used, and that its use gives generally an improved estimate of the correlation. On the other hand, the distribution in random samples of the uncorrected value is simpler and better understood, so that the uncorrected value should be used in tests of significance, in which the effect of correction need not, of course, be overlooked. For simplicity coarse grouping should be avoided where such tests are intended. The fact that with small samples the correlation obtained by the use of Sheppard's correction may exceed unity, illustrates the disturbance introduced into the random sampling distribution.

32. Partial Correlations

A great extension of the utility of the idea of correlation lies in its application to groups of more than two variates. In such cases, where the correlation between each pair of three variates is known, it is possible to eliminate any one of them, and so find what the correlation of the other two would be in a population selected so that the third variate was constant.

Ex. 26. *Elimination of age in organic correlations* [p. 151] *with growing children.* -- For example, it was found (Mumford and Young's data) in a group of boys of different ages, that the correlation of *standing height* with *chest girth* was +.836. One might expect that part of this association was due to general growth with age. It would be more desirable for many purposes to know the correlation between the variates for boys of a given age; but in fact only a few of the boys will be exactly of the same age, and even if we make age groups as broad as a year, we shall have in each group much fewer than the total number measured. In order to utilise the whole material, we only need to know the correlations of *standing height* with *age*, and of *chest girth* with *age*. These are given as .714 and .708.·

The fundamental formula in calculating partial correlation coefficients may be written

$$r_{12\cdot 3}=\frac{r_{12}-r_{13}\,r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}.$$

Here the three variates are numbered 1, 2, and 3, and we wish to find the correlation between 1 and 2, when 3 is eliminated; this is called the "partial" correlation between 1 and 2, and is designated by $r_{12.\,3}$, to show that variate 3 has been eliminated. The symbols r_{12}, r_{13}, r_{23}, indicate the correlations found directly between each pair of variates; these correlations being distinguished as "total" correlations.

Inserting the numerical values in the above formula we find $r_{12.\,3}$ =·.668, showing that when age is eliminated the correlation, though still considerable, [p. 152] has been markedly reduced. The mean value given by the above-mentioned authors for the correlations found by grouping the boys by years, is .653, not a greatly different value. In a similar manner, two or more variates may be eliminated in succession; thus with four variates, we may first eliminate variate 4, by thrice applying the above formula to find $r_{12.\,4}$, $r_{13.\,4}$, and $r_{23.\,4}$. Then applying the same formula again, to these three new values, we have

$$r_{12\cdot 34}=\frac{r_{12\cdot 4}-r_{13\cdot 4}\,r_{23\cdot 4}}{\sqrt{(1-r_{13\cdot 4}^2)(1-r_{23\cdot 4}^2)}}.$$

The labour increases rapidly with the number of variates to be eliminated. To eliminate *s* variates, the number of operations involved, each one application of the above formula is 1/6 $s(s+1)(s+2)$; for values of s from 1 to 6 this gives 1, 4, 10, 20, 35, 56 operations. Much of this labour may be saved by using tables of [sqrt]$1-r^2$ such as that published by J. R. Miner.[1]

The meaning of the correlation coefficient should be borne clearly in mind. The original aim to measure the "strength of heredity" by this method was based clearly on the supposition that the whole class of factors which tend to make relatives alike, in contrast to the unlikeness of unrelated persons, may be grouped together as heredity. That this is so for all practical purposes is, I believe, admitted, but the correlation does not tell us that this is so; it merely [p. 153] tells us the degree of resemblance in the actual population studied, between father and daughter. It tells us to what extent the height of the father is relevant information respecting the height of the daughter, or, otherwise interpreted, it tells us the relative importance of the factors which act alike upon the heights of father and daughter, compared to the totality of factors at work. If we know that B is caused by A, together with other factors independent of A, and that B has no influence on A, then the correlation between A and B does tell us how important, in relation to the other causes at work, is the influence of A. If we have not such knowledge, the correlation does not tell us whether A causes B, or B causes A, or whether both influences are at work, together with the effects of common causes.

This is true equally of partial correlations. If we know that a phenomenon A is not itself influential in determining certain other phenomena B, C, D, ..., but on the contrary is probably directly influenced by them, then the calculation of the partial correlations A with B, C, D, ... in each case eliminating the remaining values, will form a most valuable analysis of the causation of A. If on the contrary we choose a group of social phenomena with no antecedent knowledge of the causation or absence of causation among them, then the calculation of correlation coefficients, total or partial, will not advance us a step towards evaluating the importance of the causes at work.

The correlation between A and B measures, on a [p. 154] conventional scale, the importance of the factors which (on a balance of like and unlike action) act alike in both A and B, as against the remaining factors which affect A and B independently. If we eliminate a third variate C, we are removing from the comparison all those factors which become inoperative when C is fixed. If these are only those which affect A and B independently, then the correlation between A and B, whether positive or negative, will be numerically increased. We shall have eliminated irrelevant disturbing factors, and obtained, as it were, a better controlled experiment. We may also require to eliminate C if these factors act alike, or oppositely on the two variates correlated; in such a case the variability of C actually masks the effect we wish to investigate. Thirdly, C may be one of the chain of events by the mediation of which A affects B, or *vice versa.* The extent to which C is the channel through which the influence passes may be estimated by eliminating C; as one may demonstrate the small effect of latent factors in human heredity by finding the correlation of grandparent and grandchild, eliminating the intermediate parent. In no case, however, can we judge whether or not it is profitable to eliminate a certain variate unless we know, or are willing to assume, a qualitative scheme of causation. For the purely descriptive purpose of specifying a population in respect of a number of variates, either partial or total correlations are effective, and correlations of either type may be of interest.

As an illustration we may consider in what sense [p. 155] the coefficient of correlation does measure the "strength of heredity," assuming that heredity only is concerned in causing the resemblance between relatives; that is, that any environmental effects are distributed at haphazard. In the first place, we may note that if such environmental effects are increased in magnitude, the correlations would be reduced ; thus the same population, genetically speaking, would show higher correlations if reared under relatively uniform nutritional conditions, than they would if the nutritional conditions had been very diverse; although the genetical processes in the two cases were identical. Secondly, if environmental effects were at all influential (as in the population studied seems not to be indeed the case), we should obtain higher correlations from a mixed population of genetically very diverse strains, than we should from a more uniform population. Thirdly, although the influence of father on daughter is in a certain sense direct, in that the father contributes to the germinal composition of his daughter, we must not assume that this fact is necessarily the cause of the whole of the correlation; for it has been shown that husband and wife also show considerable resemblance in stature, and consequently taller fathers tend to have taller daughters partly because they choose, or are chosen by, taller wives. For this reason, for example, we should expect to find a noticeable positive correlation between step-fathers and step-daughters; also that, when the stature of the wife is eliminated, the partial correlation between father and daughter will be found to be lower than the total correlation. [p. 156] These considerations serve to some extent to define the sense in which the somewhat vague phrase, "strength of heredity," must be interpreted, in speaking of the correlation coefficient. It will readily be understood that, in less well understood cases, analogous considerations may be of some importance, and should if possible be critically considered.

33. Accuracy of the Correlation Coefficient

With large samples, and moderate or small correlations, the correlation obtained from a sample of *n* pairs of values is distributed normally about the true value r, with variance,

$$\frac{(1-\rho^2)^2}{n-1};$$

it is therefore usual to attach to an observed value *r*, a standard error $(1-r^2)$/[srqt]n-1, or $(1-r^2)$/[sqrt]n. This procedure is only valid under the restrictions stated above; with small samples the value of *r* is often very different from the true value, r, and the factor $1-r^2$, correspondingly in error; in addition the distribution of *r* is far from normal, so that tests of significance based on the above formula are often very deceptive. Since it is with small samples, less than 100, that the practical research worker ordinarily wishes to use the correlation coefficient, we shall give an account of more accurate methods of handling the results.

In all cases the procedure is alike for total and for partial correlations. Exact account may be taken of the differences in the distributions in the two cases, [p. 157] by deducting unity from the sample number for each variate eliminated; thus a partial correlation found by eliminating three variates, and based on data giving 13 values for each variate, is distributed exactly as is a total correlation based on 10 pairs of values.

34. The Significance of an Observed Correlation

In testing the significance of an observed correlation we require to calculate the probability that such a correlation should arise, by random sampling, from an uncorrelated population. If the probability is low we regard the correlation as significant. The table of *t* given at the end of the preceding chapter (p. 137) may be utilised to make an exact test. If *n'* be the number of pairs of observations on which the correlation is based, and *r* the correlation obtained, without using Sheppard's correction, then we take

$$t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n' - 2},$$

$$n = n' - 2,$$

and it may be demonstrated that the distribution of *t* so calculated, will agree with that given in the table.

It should be observed that this test, as is obviously necessary, is identical with that given in the last chapter for testing whether or not the linear regression coefficient differs significantly from zero.

TABLE V.A (p. 174) allows this test to be applied directly from the value of *r*, for samples up to 100 pairs of observations. Taking the four definite levels [p. 158] of significance, represented by P =·.10, .05, .02, and .01, the table shows for each value of *n*, from 1 to 20, and thence by larger intervals to 100, the corresponding values of *r*.

Ex. 27. *Significance of a correlation coefficient between* *autumn rainfall and wheat crop*. -- For the twenty years, 1885-1904, the mean wheat yield of Eastern England was found to be correlated with the autumn rainfall; the correlation found was -.629. Is this value significant? We obtain in succession

$$1-r^2 = \cdot 6044,$$
$$\sqrt{1-r^2} = \cdot 7774,$$
$$r/\sqrt{1-r^2} = -\cdot 8091,$$
$$t = -3\cdot 433.$$

For *n*=18, this shows that P is less than .01, and the correlation is definitely significant. The same conclusion may be read off at once from Table V.A entered with *n*=18.

If we had applied the standard error,

$$\sigma_r = \frac{1-r^2}{\sqrt{n'-1}},$$

we should have

$$t = \frac{r}{\sigma_r} = \frac{r}{1-r^2}\sqrt{n'-1} = 4\cdot 536,$$

a much greater value than the true one, very much exaggerating the significance. In addition, assuming that *r* was normally distributed (*n* = [infinity]), the significance of the result would between further exaggerated. This illustration will suffice to show how deceptive, in small samples, is the use of the standard error of the [p. 159] correlation coefficient, on the assumption that it will be normally distributed. Without this assumption the standard error is without utility. The misleading character of the formula is increased if *n'* is substituted for *n'*-1, as is often done. Judging from the normal deviate 4.536, we should suppose that the correlation obtained would be exceeded in random samples from uncorrelated material only 6 times in a million trials. Actually it would be exceeded about 3000 times in a million trials, or with 500 times the frequency supposed.

It is necessary to warn the student emphatically against the misleading character of the standard error of the correlation coefficient deduced from a small sample, because the principal utility of the correlation coefficient lies in its application to subjects of which little is known, and upon which the data are relatively scanty. With extensive material appropriate for biometrical investigations there is little danger of false conclusions being drawn, whereas with the comparatively few cases to which the experimenter must often look for guidance, the uncritical application of methods standardised in biometry, must be so frequently misleading as to endanger the credit of this most valuable weapon of research. It is not true, as the above example shows, that valid conclusions cannot be drawn from small samples; if accurate methods are used in calculating the probability, we thereby make full allowance for the size of the sample, and should be influenced in our judgment only by the value of the-probability indicated. The great increase of certainty which accrues from increasing data is [p. 160] reflected in the value of P, if accurate methods are used.

Ex. 28. *Significance of a partial correlation coefficient*. -- In a group of 32 poor law relief unions, Yule found that the percentage change from 1881 to 1891 in the percentage of the population in receipt of relief was correlated with the corresponding change in the ratio of the numbers given outdoor relief to the numbers relieved in the workhouse, when two other variates had been eliminated, namely, the corresponding changes in the percentage of the population over 65, and in the population itself.

The correlation found by Yule after eliminating the two variates was +.457; such a correlation is termed a partial correlation of the second order. Test its significance.

It has been demonstrated that the distribution in random samples of partial correlation coefficients may be derived from that of total correlation coefficients merely by deducting from the number of the sample, the number of variates eliminated. Deducting 2 from the 32 unions used, we have 30 as the effective number of the sample; hence

$$n=28$$

Calculating t from r as before, we find

$$t=2.719,$$

whence it appears from the table that P lies between .02 and .01. The correlation is therefore significant. This, of course, as in other cases, is on the assumption [p. 161] that the variates correlated (but not necessarily those eliminated) are normally distributed; economic variates seldom themselves give normal distributions, but the fact that we are here dealing with rates of change makes the assumption of normal distribution much more plausible. The values given in Table V.(A) for $n=25$, and $n=30$, give a sufficient indication of the level of *significance* attained by this observation.

35. Transformed Correlations

In addition to testing the significance of a correlation, to ascertain if there is any substantial evidence of association at all, it is also frequently required to perform one or more of the following operations, for each of which the standard error would be used in the case of a normally distributed quantity. With correlations derived from large samples the standard error may, therefore, be so used, except when the correlation approaches [plus or minus]1; but with small samples such as frequently occur in practice, special methods must be applied to obtain reliable results.

(i.) To test if an observed correlation differs significantly from a given theoretical value.

(ii.) To test if two observed correlations are significantly different.

(iii.) If a number of independent estimates of a correlation are available, to combine them into an improved estimate.

(iv.) To perform tests (i.) and (ii.) with such average values. [p. 162]

Problems of these kinds may be solved by a method analogous to that by which we have solved the problem of testing the significance of an observed correlation. In that case we were able from the given value r to calculate a quantity t which is distributed in a known manner, for which tables were available. The transformation led exactly to a distribution which had already been studied. The transformation which we shall now employ leads approximately to the normal distribution in which all the above tests may be carried out without difficulty. Let

$$z = \tfrac{1}{2}\{\log_e(1+r) - \log_e(1-r)\}$$

then as r changes from 0 to 1, z will pass from 0 to [infinity]. For small values of r, z is nearly equal to r, but as r approaches unity, z increases without limit. For negative values of r, z is negative. The advantage of this transformation lies in the distribution of the two quantities in random samples. The standard deviation of r depends on the true value of the correlation, r; as is seen from the formula

$$\sigma_r = \frac{1 - \rho^2}{\sqrt{n' - 1}}.$$

Since r is unknown, we have to substitute for it the observed value r, and this value will not, in small samples, be a very accurate estimate of r. The standard error of z is simpler in form,

$$\sigma_z = \frac{1}{\sqrt{n' - 3}},$$

and is practically independent of the value of the [p. 163] correlation in the population from which the sample is drawn.

In the second place the distribution of r is skew in small samples, and even for large samples it remains very skew for high correlations. The distribution of z is not strictly normal, but it tends to normality rapidly as the sample is increased, whatever may be the value of the correlation. We shall give examples to test the effect of the departure of the z distribution from normality.

Finally the distribution of r changes its form rapidly as r is changed ; consequently no attempt can be made, with reasonable hope of success, to allow for the skewness of the distribution. On the contrary, the distribution of z is nearly constant in form, and the accuracy of tests may be improved by small corrections for skewness; such corrections are, however, in any case somewhat laborious, and we shall not deal with them. The simple assumption that z is normally distributed will in all ordinary cases be sufficiently accurate.

These three advantages of the transformation from r to z may be seen by comparing Figs. 7 and 8. In Fig. 7 are shown the actual distributions of r, for 8 pairs of observations, from populations having correlations 0 and 0.8; Fig. 8 shows the corresponding distribution curves for z. The two curves in Fig. 7 are widely different in their modal heights; both are distinctly non-normal curves; in form also they are strongly contrasted, the one being symmetrical, the other highly unsymmetrical. On the contrary, in [p. 164] Fig. 8 the two curves do not differ greatly in height; although not exactly normal in form, they come so close to it, even for a small sample of 8 pairs of observations, [p. 165] that the eye cannot detect the difference; and this approximate normality holds up to the extreme limits r=[plus or minus]1. One additional feature is brought out by Fig. 8 ; in the distribution for r=0.8, although the curve itself is as symmetrical as the eye can judge of, yet the ordinate of zero error is not centrally placed. The figure, in fact, reveals the small bias which is introduced into the estimate of the correlation coefficient as ordinarily calculated; we shall treat further of this bias in the next section, and in the following chapter shall deal with a similar bias introduced in the calculation of intraclass correlations.

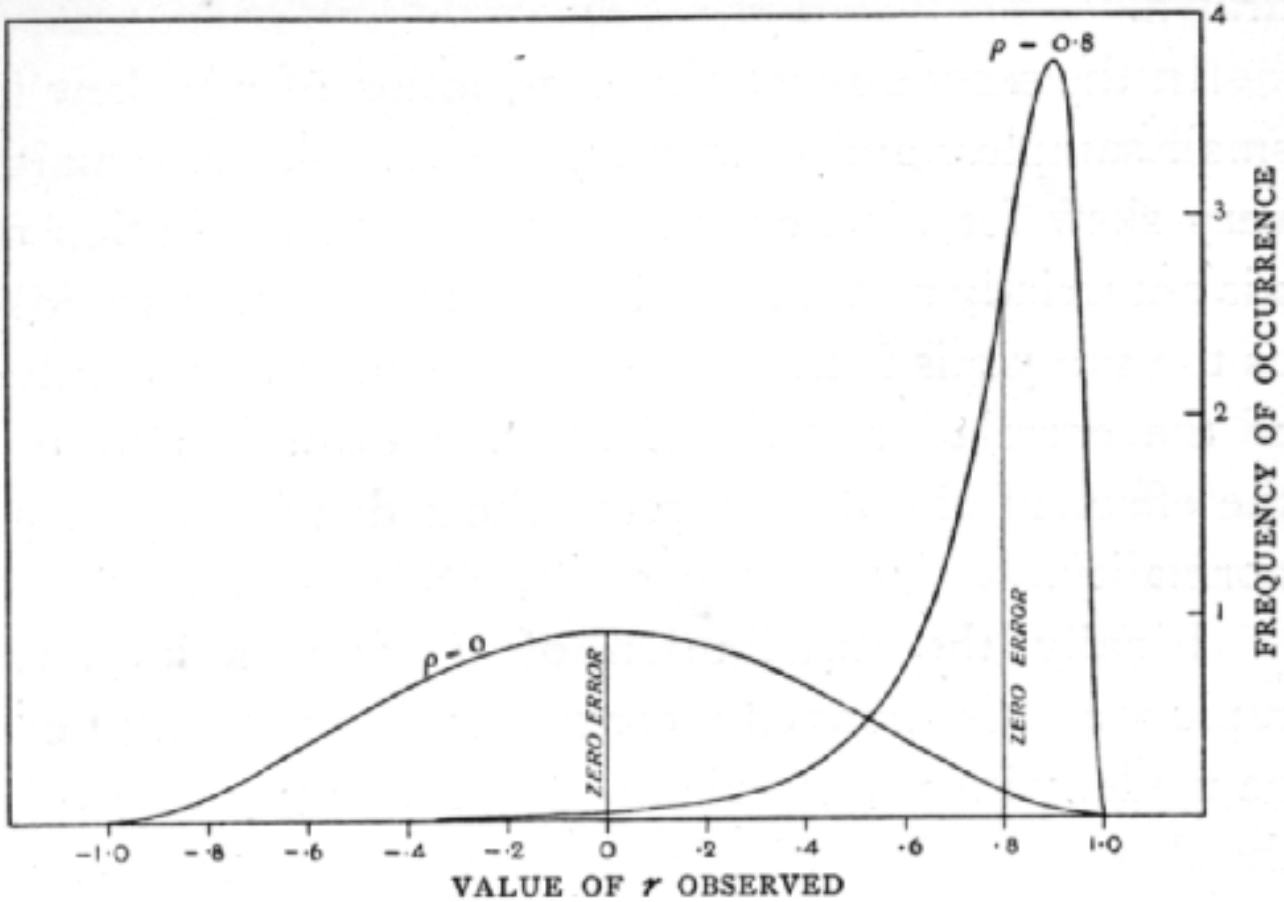

FIG. 7.

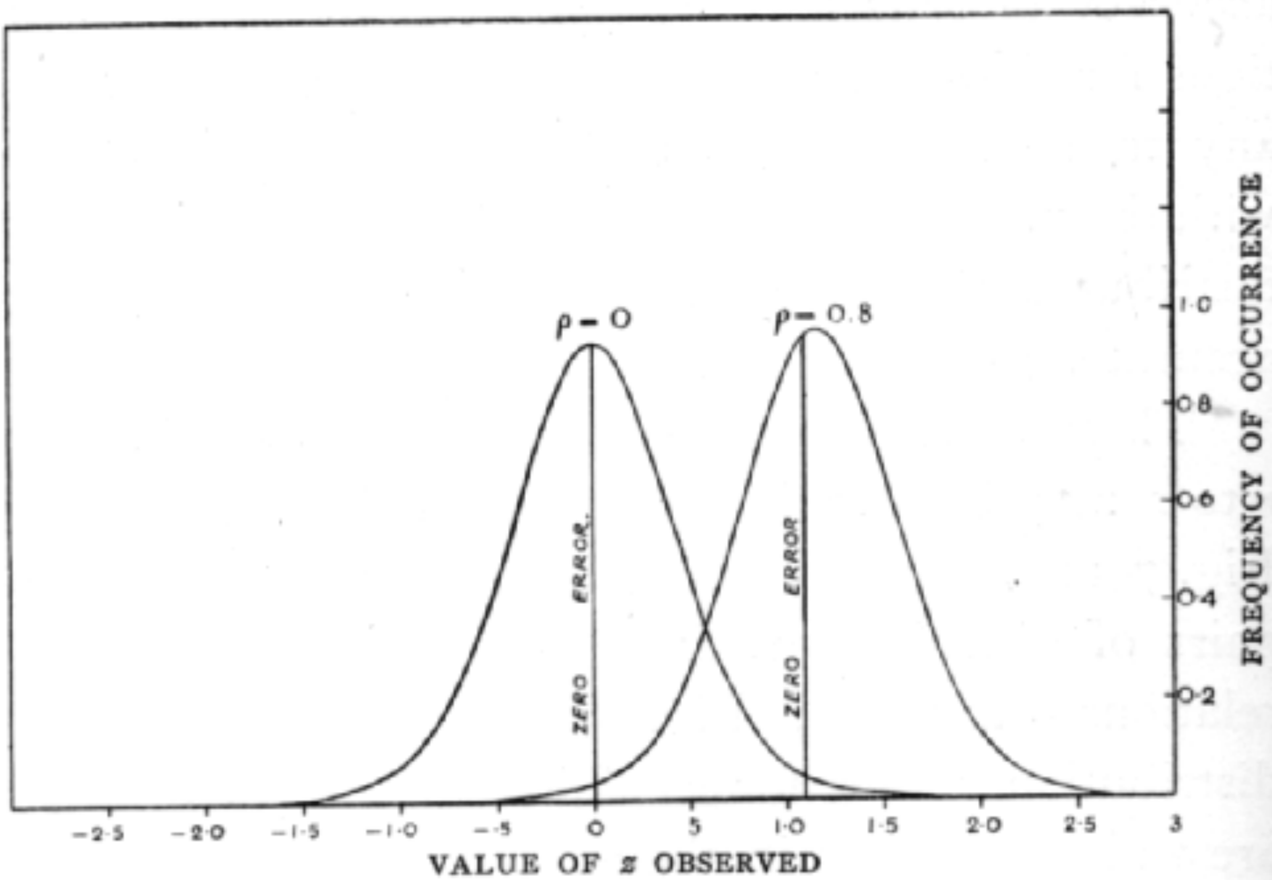

FIG. 8.

To facilitate the transformation we give in Table V.(B) (p. 175) the values of r corresponding to values of z, proceeding by intervals of ,01, from 0 to 3. In the earlier part of this table it will be seen that the values of *r* and z do not differ greatly; but with higher correlations small changes in *r* correspond to relatively large changes in *z*. In fact, measured on the *z*-scale, a correlation of .99 differs from a correlation .95 by more than a correlation .6 exceeds zero. The values of *z* give a truer picture of the relative importance of correlations of different sizes, than do the values of *r*.

To find the value of *z* corresponding to a given value of *r*, say .6, the entries in the table lying on either side of .6, are first found, whence we see at once that *z* lies between .69 and .70; the interval between these entries is then divided proportionately to find the fraction to be added to .69. In this case we have 20/64, or .31, so that *z*=.6931. Similarly, in finding [p. 166] the value of *r* corresponding to any value of *z*, say .9218, we see at once that it lies between .7259 and .7306; the difference is 47, and 18 per cent of this gives 8 to be added to the former value, giving us finally *r*=.7267. The same table may thus be used to transform *r* into *z*, and to reverse the process.

Ex. 29. *Test of the approximate normality of the distribution of z*. -- In order to illustrate the kind of accuracy obtainable by the use of *z*, let us take the case that has already been treated by an exact method in Ex. 26. A correlation of -.629 has been obtained from 20 pairs of observations ; test its significance.

For *r*=-.629 we have, using either a table of natural logarithms, or the special table for *z*, *z*=.7398. To divide this by its standard error is equivalent to multiplying it by [sqrt]17.· This gives -3.050, which we interpret as a normal deviate. From the table of normal deviates it appears that this value will be exceeded about 23 times in 10,000 trials. The true frequency, as we have seen, is about 30 times in 10,000 trials. The error tends slightly to exaggerate the significance of the result.

Ex. 30.· *Further test of the normality of the distribution of z.* -- A partial correlation +.457 was obtained from a sample of 32, after eliminating two variates. Does this differ significantly from zero? Here *z* =.4935; deducting the two eliminated variates the effective size of the sample is 30, and the standard error of *z* is 1/[sqrt]27; multiplying *z* by [sqrt]27, we have as a. normal variate 2.564. Table IV. shows, as before, that P is just over ·.01. There is a slight exaggeration [p. 167] of significance, but it is even slighter than in the previous example.

The above examples show that the *z* transformation will give a variate which, for most practical purposes, may be taken to be normally distributed. In the case of simple tests of significance the use of the table of *t* is to be preferred ; in the following examples this method is not available, and the only method available which is both tolerably accurate and sufficiently rapid for practical use lies in the use of *z*.

Ex. 31· *Significance of deviation from expectation of an observed correlation coefficient*. -- In a sample of 25 pairs of parent and child the correlation was found to be .60. Is this value consistent with the view that the true correlation in that character was .46?

The first step is to find the difference of the corresponding values of *z*. This is shown below:

TABLE 33

	r.	z.
Sample value . . .	·60	·6931
Population value . .	·46	·4973
Difference . . .		·1958

To obtain the normal deviate we multiply by [sqrt]22, and obtain .918. The deviation is less than the standard deviation, and the value obtained is therefore quite in accordance with the hypothesis. [p. 168]

Ex. 32. *Significance of difference between two observed correlations*. -- Of two samples the first, of 20 pairs, gives a correlation .6, the second, of 25 pairs, gives a correlation .8: are these values significantly different?

In this case we require not only the difference of the values of *z*, but the standard error of the difference. The variance of the difference is the sum of the reciprocals of 17 and 22; the work is shown below:

TABLE 34

	r.	z.	$n'-3$.	Reciprocal.
1st sample .	·60	·6931	17	·05882
2nd sample .	·80	1·0986	22	·04545
Difference .		·4055 ± ·3230	Sum .	·10427

The standard error which is appended to the difference of the values of z is the square root of the variance found on the same line. The difference does not exceed twice the standard error, and cannot therefore be judged significant. There is thus no sufficient evidence to conclude that the two samples are not drawn from equally correlated populations.

Ex. 33· *Combination of values from small samples.* -- Assuming that the two samples in the last example were drawn from equally correlated populations, estimate the value of the correlation.

The two values of z must be given weight inversely proportional to their variance. We therefore [p. 169] multiply the first by 17, the second by 22 and add, dividing the total by 39. This gives an estimated value of z for the population, and the corresponding value of r may be found from the table.

TABLE 35

	r.	z.	$n-3$.	$(n-3)z$.
1st sample . . .	·60	·6930	17	11·7810
2nd sample . . .	·80	1·0986	22	24·1692
	·7267	·9218	39	35·9502

The weighted average value of z is .9218, to which corresponds the value r=.7267; the value of z so obtained may be regarded as subject to normally distributed errors of random sampling with variance equal to 1/39· The accuracy is therefore equivalent to that of a single value obtained from 42 pairs of observations. Tests of significance may thus be applied to such averaged values of z, as to individual values.

36. Systematic Errors

In connexion with the averaging of correlations obtained from small samples it is worth while to consider the effects of two classes of systematic errors, which, although of little or no importance when single values only are available, become of increasing importance as larger numbers of samples are averaged.

The value of z obtained from any sample is an estimate of a true value, r, belonging to the sampled [p. 170] population, just as the value of r obtained from a sample is an estimate of a population value, r. If the method of obtaining the correlation were free from bias, the values of *z* would be normally distributed about a mean *z*[bar], which would agree in value with z. Actually there is a small bias which makes the mean value of *z* somewhat greater numerically than z; thus the correlation, whether positive or negative, is slightly exaggerated. This bias may effectively be corrected by subtracting from the value of *z* the correction

$$\frac{\rho}{2(n' - 1)}.$$

For single samples this correction is unimportant, being small compared to the standard error of *z*. For example, if *n*'=10, the standard error of *z* is .378, while the correction is r/l8 and cannot exceed .056. If, however, *z*[bar] were the mean of 1000 such values of *z*, derived from samples of 10, the standard error of *z*[bar] is only .012, and the correction, which is unaltered by taking the mean, may become of great importance.

The second type of systematic error is that introduced by neglecting Sheppard's correction. In calculating the value of *z*, we must always take the value of *r* found without using Sheppard's correction, since the latter complicates the distribution.

But the omission of Sheppard's correction introduces a systematic error, in the opposite direction to that mentioned above; and which, though normally very small, appears in large as well as in small samples. In case of averaging the correlations from a number of [p. 171] coarsely grouped small samples, the average z should be obtained from values of r found without Sheppard's correction, and to the result a correction, representing the average effect of Sheppard's correction, may be applied.

37. Correlation between Series

The extremely useful case in which it is required to find the correlation between two series of quantities, such as annual figures, arranged in order at equal intervals of time, is in reality a case of partial correlation, although it may be treated more directly by the method of fitting curved regression lines given in the last chapter (p. 128).

If, for example, we had a record of the number of deaths from a certain disease for successive years, and wished to study if this mortality were associated with meteorological conditions, or the incidence of some other disease, or the mortality of some other age group, the outstanding difficulty in the direct application of the correlation coefficient is that the number of deaths considered probably exhibits a progressive change during the period available. Such changes may be due to changes in the population among which the deaths occur, whether it be the total population of a district, or that of a particular age group, or to changes in the sanitary conditions in which the population lives, or in the skill and availability of medical assistance, or to changes in the racial or genetic composition of the population. In any case it is usually found that the changes are still apparent [p. 172] when the number of deaths is converted into a death-rate on the existing population in each year, by which means one of the direct effects of changing population is eliminated.

If the progressive change could be represented effectively by a straight line it would be sufficient to consider the time as a third variate, and to eliminate it by calculating the corresponding partial correlation coefficient. Usually, however, the change is not so simple, and would need an expression involving the square and higher powers of the time adequately to represent it. The partial correlation required is one found by eliminating not only t, but t^2, t^3, t^4, ..., regarding these as separate variates; for if we have eliminated all of these up to (say) the fourth degree, we have incidentally eliminated from the correlation any function of the time *of* the fourth degree, including that by which the progressive change is best represented.

This partial correlation may be calculated directly from the coefficients of the regression function obtained as in the last chapter (p. 128). If y and y' are the two quantities to be correlated, we obtain for y the coefficients A, B, C,..., and for y' the corresponding coefficients A', B', C', .. .; the sum of the squares of the deviations of the variates from the curved regression lines are obtained as before, from the equations

$$S(y-Y)^2 = S(y^2) - n'A^2 - \frac{n'(n'^2-1)}{12}B^2 - \ldots,$$

$$S(y'-Y')^2 = S(y'^2) - n'A'^2 - \frac{n'(n'^2-1)}{12}B'^2 - \ldots;$$

[p. 173]

while the sum of the products may be obtained from the similar equation

$$S\{(y-Y)(y'-Y')\} = S(yy') - n'AA' - \frac{n'(n'^2-1)}{12}BB' - \ldots,$$

the required partial correlation being, then,

$$r = \frac{S\{(y-Y)(y'-Y')\}}{\sqrt{S(y-Y)^2 . S(y'-Y')^2}}.$$

In this process the number of variates eliminated is equal to the degree of *t* to which the fitting has been carried; it will be understood that both variates must be fitted to the same degree, even if one of them is capable of adequate representation by a curve of lower degree than is the other. [p. 174]

TABLE V. (A).—VALUES OF THE CORRELATION COEFFICIENT FOR DIFFERENT LEVELS OF SIGNIFICANCE

n	P = ·1.	·05.	·02.	·01.
1	·98769	·996917	·9995066	·9998766
2	·90000	·95000	·98000	·990000
3	·8054	·8783	·93433	·95873
4	·7293	·8114	·8822	·91720
5	·6694	·7545	·8329	·8745
6	·6215	·7067	·7887	·8343
7	·5822	·6664	·7498	·7977
8	·5494	·6319	·7155	·7646
9	·5214	·6021	·6851	·7348
10	·4973	·5760	·6581	·7079
11	·4762	·5529	·6339	·6835
12	·4575	·5324	·6120	·6614
13	·4409	·5139	·5923	·6411
14	·4259	·4973	·5742	·6226
15	·4124	·4821	·5577	·6055
16	·4000	·4683	·5425	·5897
17	·3887	·4555	·5285	·5751
18	·3783	·4438	·5155	·5614
19	·3687	·4329	·5034	·5487
20	·3598	·4227	·4921	·5368
25	·3233	·3809	·4451	·4869
30	·2960	·3494	·4093	·4487
35	·2746	·3246	·3810	·4182
40	·2573	·3044	·3578	·3932
45	·2428	·2875	·3384	·3721
50	·2306	·2732	·3218	·3541
60	·2108	·2500	·2948	·3248
70	·1954	·2319	·2737	·3017
80	·1829	·2172	·2565	·2830
90	·1726	·2050	·2422	·2673
100	·1638	·1946	·2301	·2540

For a total correlation n is 2 less than the number of pairs in the sample; for a partial correlation the number of eliminated variates also should be subtracted.

TABLE V. (B).—TABLE OF r, FOR VALUES OF z FROM 0 TO 3

z.	·01.	·02.	·03.	·04.	·05.	·06.	·07.	·08.	·09.	·10.
·0	·0100	·0200	·0300	·0400	·0500	·0599	·0699	·0798	·0898	·0997
·1	·1096	·1194	·1293	·1391	·1489	·1586	·1684	·1781	·1877	·1974
·2	·2070	·2165	·2260	·2355	·2449	·2543	·2636	·2729	·2821	·2913
·3	·3004	·3095	·3185	·3275	·3364	·3452	·3540	·3627	·3714	·3800
·4	·3885	·3969	·4053	·4136	·4219	·4301	·4382	·4462	·4542	·4621
·5	·4699	·4777	·4854	·4930	·5005	·5080	·5154	·5227	·5299	·5370
·6	·5441	·5511	·5580	·5649	·5717	·5784	·5850	·5915	·5980	·6044
·7	·6107	·6169	·6231	·6291	·6351	·6411	·6469	·6527	·6584	·6640
·8	·6696	·6751	·6805	·6858	·6911	·6963	·7014	·7064	·7114	·7163
·9	·7211	·7259	·7306	·7352	·7398	·7443	·7487	·7531	·7574	·7616
1·0	·7658	·7699	·7739	·7779	·7818	·7857	·7895	·7932	·7969	·8005
1·1	·8041	·8076	·8110	·8144	·8178	·8210	·8243	·8275	·8306	·8337
1·2	·8367	·8397	·8426	·8455	·8483	·8511	·8538	·8565	·8591	·8617
1·3	·8643	·8668	·8692	·8717	·8741	·8764	·8787	·8810	·8832	·8854
1·4	·8875	·8896	·8917	·8937	·8957	·8977	·8996	·9015	·9033	·9051
1·5	·9069	·9087	·9104	·9121	·9138	·9154	·9170	·9186	·9201	·9217
1·6	·9232	·9246	·9261	·9275	·9289	·9302	·9316	·9329	·9341	·9354
1·7	·9366	·9379	·9391	·9402	·9414	·9425	·9436	·9447	·9458	·94681
1·8	·94783	·94884	·94983	·95080	·95175	·95268	·95359	·95449	·95537	·95624
1·9	·95709	·95792	·95873	·95953	·96032	·96109	·96185	·96259	·96331	·96403
2·0	·96473	·96541	·96609	·96675	·96739	·96803	·96865	·96926	·96986	·97045
2·1	·97103	·97159	·97215	·97269	·97323	·97375	·97426	·97477	·97526	·97574
2·2	·97622	·97668	·97714	·97759	·97803	·97846	·97888	·97929	·97970	·98010
2·3	·98049	·98087	·98124	·98161	·98197	·98233	·98267	·98301	·98335	·98367
2·4	·98399	·98431	·98462	·98492	·98522	·98551	·98579	·98607	·98635	·98661
2·5	·98688	·98714	·98739	·98764	·98788	·98812	·98835	·98858	·98881	·98903
2·6	·98924	·98945	·98966	·98987	·99007	·99026	·99045	·99064	·99083	·99101
2·7	·99118	·99136	·99153	·99170	·99186	·99202	·99218	·99233	·99248	·99263
2·8	·99278	·99292	·99306	·99320	·99333	·99346	·99359	·99372	·99384	·99396
2·9	·99408	·99420	·99431	·99443	·99454	·99464	·99475	·99485	·99495	·99505

For greater accuracy, and for values beyond the table,

$$r = (e^{2z} - 1)/(e^{2z} + 1);$$

$$z = \tfrac{1}{2}\{\log(1+r) - \log(1-r)\}.$$

Footnotes

[1] *Tables of* [sqrt]*1-r*^{2} *for Use in Partial Correlation, and in Trigonometry*. Johns Hopkins Press, 1922.

Classics in the History of Psychology

An internet resource developed by
Christopher D. Green
York University, Toronto, Ontario
ISSN 1492-3173

(Return to index)

STATISTICAL METHODS FOR RESEARCH WORKERS

By Ronald A. Fisher (1925)

Posted April 2000

SOURCES USED FOR DATA AND METHODS

J. W. BISPHAM (1923). An experimental determination of the distribution of the partial correlation coefficient in samples of thirty. Metron, ii. 684-696.

J. BLAKEMAN (1905). On tests for linearity of regression in frequency distributions. Biometrika, iv. 332.·

J. T. ROTTOMLEY. Four-figure mathematical tables. Macmillan and Co.

M. BRISTOL-ROACH (1925). On the relation of certain soil algæ to some soluble organic compounds. Ann. Bot., xxxix.

J. BURGESS (1895). On the definite integral, etc. Trans. Roy. Soc. Edin., xxxix. 257-321.

W. P. ELDERTON (1902). Tables for testing the goodness of fit of theory to observation. Biometrika, i. 155.·

R. A. FISHER (1921). On the mathematical foundations of theoretical statistics. Phil. Trans., A. ccxxii. 309-368.

R. A. FISHER (1921). An examination of the yield of dressed grain from Broadbalk. Journal of Agricultural Science, xi. 107-135.

R, A. FISHER (1921). On the " probable error " of a coefficient of correlation deduced from a small sample. Metron, i. pt. 4, 1-32.

R. A. FISHER (1922). On the interpretation of c^2 from contingency tables, and on the calculation of P. Journal of the Royal Statistical Society, lxxxv. 81-94.

R. A. FISHER (1922). The goodness of fit of regression formula, and the distribution of regression coefficients. Journal of the Royal Statistical Society, lxxxv. 597-612.

R. A. FISHER, H. G. THORNTON, and W. A. MACKENZIE (1922). The accuracy of the plating method of estimating the density of bacterial populations. Annals of Applied

Biology, ix. 325-359. [p. 234] ·

R. A. FISHER and W. A. MACKENZIE (1923). The manurial response of different potato varieties. Journal of Agricultural Science, xiii. 311-320.

R. A. FISHER (1924). The distribution of the partial correlation coefficient. Metron, iii. 329-332.·

R. A. FISHER (1924). The conditions under which c^2 measures the discrepancy between observation and hypothesis. Journal of the Royal Statistical Society, lxxxvii. 442-449.

R. A. FISHER (1924). The influence of rainfall upon the yield of wheat at Rothamsted. Phil. Trans., B. ccxiii. 89-142.

J. G. H. FREW (1924). On *Chlorops Tœniopus* Meig. (The gout fly of barley.) Annals of Applied Biology, xi. 175-219.

GEISSLER (1889). Beiträge zur Frage des Geschlechts verhältnisses der Geborenen. Zeitschift des K. Sachsischen Statistischen Bureaus.

J. W. L. GLAISHER (1871). On a class of definite integrals. Phil. Mag., Series IV. xlii. 421-436.

M. GREENWOOD and G. U. YULE (1915) The statistics of antityphoid and anticholera inoculations, and the interpretation of such statistics in general. Pro. Roy. Sec. Medicine; section of epidemiology and state medicine, viii.

113.·

R. P. GREGORY, D. DE WINTON, and W. BATESON (1923). Genetics of Primula Sinensis. Journal of Genetics, xiii. 219-253.

W. HALL. Four-figure tables and constants.

J. A. HARRIS (1916). A contribution to the problem of homotyposis. Biometrika, xi. 201-214.

A. H. HERSH (1924). The effects of temperature upon the heterozygotes in the bar series of Drosophila. Journal of Experimental Zoology, xxxix. 55-71.·

J. S. HUXLEY (1923). Further data on linkage in Gammarus Chevreuxi and its relation to cytology. British Journal of Exp. Biology, i. 79-96.

T.L.KELLEY(1923). Statistical method. Macmillan and Co.

"MATHETES" (1924). Statistical study on the effect of manuring on infestation of barley by gout fly. Annals of Applied Biology, xi. 220-235.

W. B. MERCER and A. D. HALL (1911). The experimental error of field trials. Journal of Agricultural Science, iv. 107-132.

J. R. MINER (1922). Tables of [sqrt]1-r^2 and 1-r^2. Johns Hopkins Press, Baltimore. [p. 235]

A. A. MUMFORD and M. YOUNG (1923). The interrelationships of the physical measurements and the vital capacity. Biometrika, xv. 109-133.·

K. PEARSON (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Phil. Mag., Series V. 1. 157-175.

K. PEARSON and A. LEE (1903). Inheritance of physical characters. Biometrika, ii. 357-462.

K. PEARSON (1922). Tables of the incomplete G-function. Stationery Office.

N. SHAW (1922). The air and its ways. Cambridge University Press.

W. F. SHEPPARD (1907) Table of deviates of the normal curve. Biometrika, v. 404-406.

"Student" (1907). On the error of counting with a hæemocytometer. Biometrika, v. 351-360.

"Student" (1908). The probable error of a mean. Biometrika, vi. 1-25.·

"Student" (1917) Tables for estimating the probability that the mean of a unique sample of observations lies between -[infinity] and any given distance of the mean of the population from which the sample is drawn. Biometrika, xi.

414-417.·

J. F. TOCHER (1908). Pigmentation survey of school children in Scotland. Biometrika, vi. 129-235.

G. U. YULE (1917). An introduction to the theory of statistics. C. Griffin and Co., London.

G. U. YULE (I923). On the application of the c^2 method to association and contingency tables, with experimental illustrations. Journal of the Royal Statistical Society, lxxxv. 95-104.

TABLE I

TABLE OF x

The deviation in the normal distribution in terms of the standard deviation.

	·01.	·02.	·03.	·04.	·05.	·06.	·07.	·08.	·09.	·10.
·00	2·575829	2·326348	2·170090	2·053749	1·959964	1·880794	1·811911	1·750686	1·695398	1·644854
·10	1·598193	1·554774	1·514102	1·475791	1·439521	1·405072	1·372204	1·340755	1·310579	1·281552
·20	1·253565	1·226528	1·200359	1·174987	1·150349	1·126391	1·103063	1·080319	1·058122	1·036433
·30	1·015222	·994458	·974114	·954165	·934589	·915365	·896473	·877896	·859617	·841621
·40	·823894	·806421	·789192	·772193	·755415	·738847	·722479	·706303	·690309	·674490
·50	·658838	·643345	·628006	·612813	·597760	·582841	·568051	·553385	·538836	·524401
·60	·510073	·495850	·481727	·467699	·453762	·439913	·426148	·412463	·398855	·385320
·70	·371856	·358459	·345125	·331853	·318639	·305481	·292375	·279319	·266311	·253347
·80	·240426	·227545	·214702	·201893	·189118	·176374	·163658	·150969	·138304	·125661
·90	·113039	·100434	·087845	·075270	·062707	·050154	·037608	·025069	·012533	0

The value of P for each entry is found by adding the column heading to the value in the left-hand margin. The corresponding value of x is the deviation such that the probability of an observation falling outside the range from $-x$ to $+x$ is P. For example, $P = \cdot 03$ for $x = 2\cdot 170090$; so that 3 per cent of normally distributed values will have positive or negative deviations exceeding the standard deviation in the ratio 2·170090.

TABLE II

VALUES OF x FOR SMALL VALUES OF P

P . .	·001	·000,1	·000,01	·000,001	·000,000,1	·000,000,01	·000,000,001
x . .	3·29053	3·89059	4·41717	4·89164	5·32672	5·73073	6·10941

TABLE III TABLE OF χ^2

n.	P = ·99.	·98.	·95.	·90.	·80.	·70.	·50.	·30	·20.	·10.	·05.	·02.	·01.
1	·000157	·000628	·00393	·0158	·0642	·148	·455	1·074	1·642	2·706	3·841	5·412	6·635
2	·0201	·0404	·103	·211	·446	·713	1·386	2·408	3·219	4·605	5·991	7·824	9·210
3	·115	·185	·352	·584	1·005	1·424	2·366	3·665	4·642	6·251	7·815	9·837	11·341
4	·297	·429	·711	1·064	1·649	2·195	3·357	4·878	5·989	7·779	9·488	11·668	13·277
5	·554	·752	1·145	1·610	2·343	3·000	4·351	6·064	7·289	9·236	11·070	13·388	15·086
6	·872	1·134	1·635	2·204	3·070	3·828	5·348	7·231	8·558	10·645	12·592	15·033	16·812
7	1·239	1·564	2·167	2·833	3·822	4·671	6·346	8·383	9·803	12·017	14·067	16·622	18·475
8	1·646	2·032	2·733	3·490	4·594	5·527	7·344	9·524	11·030	13·362	15·507	18·168	20·090
9	2·088	2·532	3·325	4·168	5·380	6·393	8·343	10·656	12·242	14·684	16·919	19·679	21·666
10	2·558	3·059	3·940	4·865	6·179	7·267	9·342	11·781	13·442	15·987	18·307	21·161	23·209
11	3·053	3·609	4·575	5·578	6·989	8·148	10·341	12·899	14·631	17·275	19·675	22·618	24·725
12	3·571	4·178	5·226	6·304	7·807	9·034	11·340	14·011	15·812	18·549	21·026	24·054	26·217
13	4·107	4·765	5·892	7·042	8·634	9·926	12·340	15·119	16·985	19·812	22·362	25·472	27·688
14	4·660	5·368	6·571	7·790	9·467	10·821	13·339	16·222	18·151	21·064	23·685	26·873	29·141
15	5·229	5·985	7·261	8·547	10·307	11·721	14·339	17·322	19·311	22·307	24·996	28·259	30·578
16	5·812	6·614	7·962	9·312	11·152	12·624	15·338	18·418	20·465	23·542	26·296	29·633	32·000
17	6·408	7·255	8·672	10·085	12·002	13·531	16·338	19·511	21·615	24·769	27·587	30·995	33·409
18	7·015	7·906	9·390	10·865	12·857	14·440	17·338	20·601	22·760	25·989	28·869	32·346	34·805
19	7·633	8·567	10·117	11·651	13·716	15·352	18·338	21·689	23·900	27·204	30·144	33·687	36·191
20	8·260	9·237	10·851	12·443	14·578	16·266	19·337	22·775	25·038	28·412	31·410	35·020	37·566
21	8·897	9·915	11·591	13·240	15·445	17·182	20·337	23·858	26·171	29·615	32·671	36·343	38·932
22	9·542	10·600	12·338	14·041	16·314	18·101	21·337	24·939	27·301	30·813	33·924	37·659	40·289
23	10·196	11·293	13·091	14·848	17·187	19·021	22·337	26·018	28·429	32·007	35·172	38·968	41·638
24	10·856	11·992	13·848	15·659	18·062	19·943	23·337	27·096	29·553	33·196	36·415	40·270	42·980
25	11·524	12·697	14·611	16·473	18·940	20·867	24·337	28·172	30·675	34·382	37·652	41·566	44·314
26	12·198	13·409	15·379	17·292	19·820	21·792	25·336	29·246	31·795	35·563	38·885	42·856	45·642
27	12·879	14·125	16·151	18·114	20·703	22·719	26·336	30·319	32·912	36·741	40·113	44·140	46·963
28	13·565	14·847	16·928	18·939	21·588	23·647	27·336	31·391	34·027	37·916	41·337	45·419	48·278
29	14·256	15·574	17·708	19·768	22·475	24·577	28·336	32·461	35·139	39·087	42·557	46·693	49·588
30	14·953	16·306	18·493	20·599	23·364	25·508	29·336	33·530	36·250	40·256	43·773	47·962	50·892

For larger values of n, the expression $\sqrt{2\chi^2} - \sqrt{2n-1}$ may be used as a normal deviate with unit standard error.

TABLE IV.—TABLE OF *t*

n.	P = ·9.	·8.	·7.	·6.	·5.	·4.	·3.	·2.	·1.	·05.	·02.	·01.
1	·158	·325	·510	·727	1·000	1·376	1·963	3·078	6·314	12·706	31·821	63·657
2	·142	·289	·445	·617	·816	1·061	1·386	1·886	2·920	4·303	6·965	9·925
3	·137	·277	·424	·584	·765	·978	1·250	1·638	2·353	3·182	4·541	5·841
4	·134	·271	·414	·569	·741	·941	1·190	1·533	2·132	2·776	3·747	4·604
5	·132	·267	·408	·559	·727	·920	1·156	1·476	2·015	2·571	3·365	4·032
6	·131	·265	·404	·553	·718	·906	1·134	1·440	1·943	2·447	3·143	3·707
7	·130	·263	·402	·549	·711	·896	1·119	1·415	1·895	2·365	2·998	3·499
8	·130	·262	·399	·546	·706	·889	1·108	1·397	1·860	2·306	2·896	3·355
9	·129	·261	·398	·543	·703	·883	1·100	1·383	1·833	2·262	2·821	3·250
10	·129	·260	·397	·542	·700	·879	1·093	1·372	1·812	2·228	2·764	3·169
11	·129	·260	·396	·540	·697	·876	1·088	1·363	1·796	2·201	2·718	3·106
12	·128	·259	·395	·539	·695	·873	1·083	1·356	1·782	2·179	2·681	3·055
13	·128	·259	·394	·538	·694	·870	1·079	1·350	1·771	2·160	2·650	3·012
14	·128	·258	·393	·537	·692	·868	1·076	1·345	1·761	2·145	2·624	2·977
15	·128	·258	·393	·536	·691	·866	1·074	1·341	1·753	2·131	2·602	2·947
16	·128	·258	·392	·535	·690	·865	1·071	1·337	1·746	2·120	2·583	2·921
17	·128	·257	·392	·534	·689	·863	1·069	1·333	1·740	2·110	2·567	2·898
18	·127	·257	·392	·534	·688	·862	1·067	1·330	1·734	2·101	2·552	2·878
19	·127	·257	·391	·533	·688	·861	1·066	1·328	1·729	2·093	2·539	2·861
20	·127	·257	·391	·533	·687	·860	1·064	1·325	1·725	2·086	2·528	2·845
21	·127	·257	·391	·532	·686	·859	1·063	1·323	1·721	2·080	2·518	2·831
22	·127	·256	·390	·532	·686	·858	1·061	1·321	1·717	2·074	2·508	2·819
23	·127	·256	·390	·532	·685	·858	1·060	1·319	1·714	2·069	2·500	2·807
24	·127	·256	·390	·531	·685	·857	1·059	1·318	1·711	2·064	2·492	2·797
25	·127	·256	·390	·531	·684	·856	1·058	1·316	1·708	2·060	2·485	2·787
26	·127	·256	·390	·531	·684	·856	1·058	1·315	1·706	2·056	2·479	2·779
27	·127	·256	·389	·531	·684	·855	1·057	1·314	1·703	2·052	2·473	2·771
28	·127	·256	·389	·530	·683	·855	1·056	1·313	1·701	2·048	2·467	2·763
29	·127	·256	·389	·530	·683	·854	1·055	1·311	1·699	2·045	2·462	2·756
30	·127	·256	·389	·530	·683	·854	1·055	1·310	1·697	2·042	2·457	2·750
∞	·12566	·25335	·38532	·52440	·67449	·84162	1·03643	1·28155	1·64485	1·95996	2·32634	2·57582

TABLE V. (A).—VALUES OF THE CORRELATION COEFFICIENT FOR DIFFERENT LEVELS OF SIGNIFICANCE

n	P = ·1.	·05.	·02.	·01.
1	·98769	·996917	·9995066	·9998766
2	·90000	·95000	·98000	·990000
3	·8054	·8783	·93433	·95873
4	·7293	·8114	·8822	·91720
5	·6694	·7545	·8329	·8745
6	·6215	·7067	·7887	·8343
7	·5822	·6664	·7498	·7977
8	·5494	·6319	·7155	·7646
9	·5214	·6021	·6851	·7348
10	·4973	·5760	·6581	·7079
11	·4762	·5529	·6339	·6835
12	·4575	·5324	·6120	·6614
13	·4409	·5139	·5923	·6411
14	·4259	·4973	·5742	·6226
15	·4124	·4821	·5577	·6055
16	·4000	·4683	·5425	·5897
17	·3887	·4555	·5285	·5751
18	·3783	·4438	·5155	·5614
19	·3687	·4329	·5034	·5487
20	·3598	·4227	·4921	·5368
25	·3233	·3809	·4451	·4869
30	·2960	·3494	·4093	·4487
35	·2746	·3246	·3810	·4182
40	·2573	·3044	·3578	·3932
45	·2428	·2875	·3384	·3721
50	·2306	·2732	·3218	·3541
60	·2108	·2500	·2948	·3248
70	·1954	·2319	·2737	·3017
80	·1829	·2172	·2565	·2830
90	·1726	·2050	·2422	·2673
100	·1638	·1946	·2301	·2540

For a total correlation n is 2 less than the number of pairs in the sample; for a partial correlation the number of eliminated variates also should be subtracted.

TABLE V. (B).—TABLE OF r, FOR VALUES OF z FROM 0 TO 3

z.	·01.	·02.	·03.	·04.	·05.	·06.	·07.	·08.	·09.	·10.
·0	·0100	·0200	·0300	·0400	·0500	·0599	·0699	·0798	·0898	·0997
·1	·1096	·1194	·1293	·1391	·1489	·1586	·1684	·1781	·1877	·1974
·2	·2070	·2165	·2260	·2355	·2449	·2543	·2636	·2729	·2821	·2913
·3	·3004	·3095	·3185	·3275	·3364	·3452	·3540	·3627	·3714	·3800
·4	·3885	·3969	·4053	·4136	·4219	·4301	·4382	·4462	·4542	·4621
·5	·4699	·4777	·4854	·4930	·5005	·5080	·5154	·5227	·5299	·5370
·6	·5441	·5511	·5580	·5649	·5717	·5784	·5850	·5915	·5980	·6044
·7	·6107	·6169	·6231	·6291	·6351	·6411	·6469	·6527	·6584	·6640
·8	·6696	·6751	·6805	·6858	·6911	·6963	·7014	·7064	·7114	·7163
·9	·7211	·7259	·7306	·7352	·7398	·7443	·7487	·7531	·7574	·7616
1·0	·7658	·7699	·7739	·7779	·7818	·7857	·7895	·7932	·7969	·8005
1·1	·8041	·8076	·8110	·8144	·8178	·8210	·8243	·8275	·8306	·8337
1·2	·8367	·8397	·8426	·8455	·8483	·8511	·8538	·8565	·8591	·8617
1·3	·8643	·8668	·8692	·8717	·8741	·8764	·8787	·8810	·8832	·8854
1·4	·8875	·8896	·8917	·8937	·8957	·8977	·8996	·9015	·9033	·9051
1·5	·9069	·9087	·9104	·9121	·9138	·9154	·9170	·9186	·9201	·9217
1·6	·9232	·9246	·9261	·9275	·9289	·9302	·9316	·9329	·9341	·9354
1·7	·9366	·9379	·9391	·9402	·9414	·9425	·9436	·9447	·9458	·94681
1·8	·94783	·94884	·94983	·95080	·95175	·95268	·95359	·95449	·95537	·95624
1·9	·95709	·95792	·95873	·95953	·96032	·96109	·96185	·96259	·96331	·96403
2·0	·96473	·96541	·96609	·96675	·96739	·96803	·96865	·96926	·96986	·97045
2·1	·97103	·97159	·97215	·97269	·97323	·97375	·97426	·97477	·97526	·97574
2·2	·97622	·97668	·97714	·97759	·97803	·97846	·97888	·97929	·97970	·98010
2·3	·98049	·98087	·98124	·98161	·98197	·98233	·98267	·98301	·98335	·98367
2·4	·98399	·98431	·98462	·98492	·98522	·98551	·98579	·98607	·98635	·98661
2·5	·98688	·98714	·98739	·98764	·98788	·98812	·98835	·98858	·98881	·98903
2·6	·98924	·98945	·98966	·98987	·99007	·99026	·99045	·99064	·99083	·99101
2·7	·99118	·99136	·99153	·99170	·99186	·99202	·99218	·99233	·99248	·99263
2·8	·99278	·99292	·99306	·99320	·99333	·99346	·99359	·99372	·99384	·99396
2·9	·99408	·99420	·99431	·99443	·99454	·99464	·99475	·99485	·99495	·99505

For greater accuracy, and for values beyond the table,

$$r = (e^{2z} - 1)/(e^{2z} + 1);$$

$$z = \tfrac{1}{2}\{\log(1 + r) - \log(1 - r)\}.$$

TABLE VI

TABLE OF z. $P = \cdot 05$

Values of n_2.	Values of n_1. 1.	2.	3.	4.	5.	6.	8.	12.	24.	∞.
1	2·5421	2·6479	2·6870	2·7071	2·1974	2·7276	2·7380	2·7484	2·7588	2·7693
2	1·4592	1·4722	1·4765	1·4787	1·4800	1·4808	1·4819	1·4830	1·4840	1·4851
3	1·1577	1·1284	1·1137	1·1051	1·0994	1·0953	1·0899	1·0842	1·0781	1·0716
4	1·0212	·9690	·9429	·9272	·9168	·9093	·8993	·8885	·8767	·8639
5	·9441	·8777	·8441	·8236	·8097	·7997	·7862	·7714	·7550	·7368
6	·8948	·8188	·7798	·7558	·7394	·7274	·7112	·6931	·6729	·6499
8	·8355	·7475	·7014	·6725	·6525	·6378	·6175	·5945	·5682	·5371
12	·7788	·6786	·6250	·5907	·5666	·5487	·5234	·4941	·4592	·4156
24	·7246	·6123	·5508	·5106	·4817	·4598	·4283	·3904	·3425	·2749
∞	·6729	·5486	·4787	·4319	·3974	·3706	·3309	·2804	·2085	0