Statistical Methods for Research Workers

BY

SIR RONALD A. FISHER, SC.D., F.R.S.

D.Sc. (Ames, Chicago, Harvard, London), LL.D. (Calcutta, Glasgow)

Fellow of Gonville and Caius College, Cambridge Foreign Associate United States National Academy of Science and Foreign Honorary Member American Academy of Arts and Sciences ; Foreign Member of the Royal Swedish Academy of Sciences ; Member of the Royal Danish Academy of Sciences ; Foreign Member American Philosophical Society ; formerly Galton Professor, University of London ; formerly Arthur Balfour Professor of Genetics, University of Cambridge

THIRTEENTH EDITION—REVISED

First Published	1925
Second Edition	1928
Third Edition	1930
Fourth Edition	1932
Fifth Edition	1934
Sixth Edition	1936
Seventh Edition	1938
Eighth Edition	1941
Ninth Edition	1944
Tenth Edition	1946
Tenth Edition Reprinted	1948
Eleventh Edition	1950
Twelfth Edition	1954
Thirteenth Edition	1958

Copyright Translations

French
Presses Universitaires de France
108 Boulevard St Germain
Paris 6e, France

German
Oliver and Boyd Ltd
Tweeddale Court
Edinburgh, Scotland

Italian
Unione Tipografico
Editrice Torinese
Corso Raffaello 28
Torino, Italy

Spanish
Aguilar
S.A. de Ediciodes
Juan Bravo 38
Madrid, Spain

Japanese
Messrs Sobunsha
129 1-Chome, Hinode-cho,
Toshima-ku, Tokyo, Japan

also Door Publications. N.Y. expected 1958

BY THE SAME AUTHOR

THE GENETICAL THEORY OF NATURAL SELECTION, 1930, Oxford Univ. Press

THE DESIGN OF EXPERIMENTS, 1935, 1937, 1942, 1947, 1949, 1951, Oliver and Boyd Ltd.

STATISTICAL TABLES (With Frank Yates), 1938, 1943, 1948, 1953, 1957, Oliver and Boyd Ltd.

THE THEORY OF INBREEDING, 1949, Oliver and Boyd Ltd.

CONTRIBUTIONS TO MATHEMATICAL STATISTICS, 1949, John Wiley and Sons Inc., New York

STATISTICAL METHODS AND SCIENTIFIC INFERENCE, 1956, Oliver and Boyd Ltd.

PRINTED AND PUBLISHED IN GREAT BRITAIN BY
OLIVER AND BOYD LTD., EDINBURGH

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 the practice of older days when animal physiology 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. Often the analysis of large masses of data by statistical methods is necessary, and the biological worker is continually encountering advanced statistical problems the adequate solutions of which are not found in current statistical text-books. To meet these needs the present monograph was prepared, and the early call for the second and later editions indicates the success attained by the author in this project.

F. A. E. C.
D. W. C.

PREFACE TO THIRTEENTH EDITION

For several years prior to the preparation of this book, the author had been working in somewhat intimate co-operation with a number of biological research departments at Rothamsted; the book was very decidedly the product of this circumstance. Daily contact with statistical problems as they presented themselves to laboratory workers stimulated the purely mathematical researches upon which the new methods were based. It was clear that the traditional machinery inculcated by the biometrical school was wholely unsuited to the needs of practical research. The futile elaboration of innumerable measures of correlation, and the evasion of the real difficulties of sampling problems under cover of a contempt for small samples, were obviously beginning to make its pretentions ridiculous. These procedures were not only ill-aimed, but, for all their elaboration, not sufficiently accurate. Only by tackling small sample problems on their merits, in the author's view, did it seem possible to apply accurate tests to practical data. With the encouragement of my colleagues, and the valued help of the late W. S. Gosset ("Student"), his assistant Mr E. Somerfield, and Miss W. A. Mackenzie, the first edition was prepared and weathered the hostile criticisms inevitable to such a venture.

To-day exact tests of significance need no apology. The demand, steadily increasing over a long period, for a book designed originally for a much smaller public has justified at least some of the innovations in its plan which at first must have seemed questionable. (The recognition of degrees of freedom ; the use of fixed probability levels in tabulating the functions used in tests of significance ; the analysis of variance ; the need for randomisation in experimental design, etc.) The author was impressed with the practical importance of many recent mathematical advances, which to others seemed to be merely academic refinements. He felt sure, too, that workers with research experience would appreciate a book which, without entering into the mathematical theory of statistical methods, should embody the latest results of that theory, presenting them in the form of practical procedures appropriate to those types of data with which research workers are actually concerned. The practical application of general theorems is a different art from their establishment by mathematical proof. It requires fully as deep an understanding of their meaning, and is, moreover, useful to many to whom the other is unnecessary. To carry out this plan new matter has had to be added with each new edition, to illustrate extensions and improvements, the value of which had in the meantime been established by experience.

In most cases the new methods actually simplify the handling of the data. The conservatism of some university courses in elementary statistics, in stereotyping unnecessary approximations and inappropriate conventions, still hinders many students in the use of exact methods. In reading this book they should try to remember that departures from tradition have not been made capriciously, but only when they have been found to be definitely helpful.

Especially in the order of presentation, the book bears traces of the state of the subject when it first appeared. More recent books have, rightly from the teacher's standpoint, introduced the analysis of variance earlier, and given it more space. They have thus carried further than I the process of abstracting from the field formerly embraced by the correlation coefficient, problems capable of a more direct approach. In excusing myself from the difficult task of a fundamental rearrangement, I may plead that it is of real value to understand the problems discussed by earlier writers, and to be able to translate them into the system of ideas in which they may be more simply or more comprehensively resolved. I have therefore contented myself with indicating the analysis of variance procedure as an alternative approach in some early examples, as in Sections 24 and 24·1.

With a class capable of mastering the whole book, I should now postpone the matter of Sections 30 to 40, dealing with correlation, until further experience has been gained of the applications of the Analysis of Variance, but should later give time to the ideas of correlation and partial correlation, for their importance in understanding the literature of quantitative biology, which has been so largely influenced by them.

In the second edition the importance of providing a striking and detailed illustration of the principles of statistical estimation led to the addition of a ninth chapter. The subject had received only general discussion in the first edition, and, in spite of its practical importance, had not yet won sufficient attention from teachers to drive out of practice the demonstrably defective procedures which were still unfortunately taught to students. The new chapter superseded Section 6 and Example 1 of the first edition; in the third edition it was enlarged by two new sections (57·1 and 57·2) illustrating further the applicability of the method of maximum likelihood, and of the quantitative evaluation of information. Later K. Mather's admirable book *The Measurement of Linkage in Heredity* has illustrated the appropriate procedures for a wider variety of genetical examples.

In Section 27 a general method of constructing the series of orthogonal polynomials was added to the third edition, in response to the need which is felt, with respect to some important classes of data, to use polynomials of higher degree than the fifth. Simple and direct algebraic proofs of the methods of Sections 28 and 28·1 have been published by Miss F. E. Allan.

In the fourth edition the Appendix to Chapter III, on technical notation, was entirely rewritten, since the inconveniences of the moment notation seemed by that time definitely to outweigh the advantages formerly conferred by its familiarity. The principal new matter in that edition was added in response to the increasing use of the analysis of covariance, which is explained in Section 49·1. Since several writers had found difficulty in applying the appropriate tests of significance to deviations from regression formulæ, this section was further enlarged in the fifth edition.

Other new sections in the fifth edition were 21·01, giving a correction for continuity recently introduced by F. Yates, and 21·02 giving the exact test of significance for 2×2 tables. Workers who are accustomed to handle regression equations with a large number of variates will be interested in Section 29·1, which provides the relatively simple adjustments to be made when, at a late stage, it is decided that one or more of the variates used may with advantage be omitted. The possibility of doing this without laborious recalculations should encourage workers to make the list of independent variates included more comprehensive than has, in the past, been thought advisable.

Section 5, formerly occupied by an account of the tables available for testing significance, was given to a historical note on the principal contributors to the development of statistical reasoning.

In the sixth edition Example 15·1, Section 22, gave a new test of homogeneity for data with hierarchical subdivisions. Attention was also called to Working and Hotelling's formula for the sampling error of values estimated by regression, and in Section 29·2 to an extended use of successive summation in fitting polynomials.

I am indebted to Dr W. E. Deming for the extension of the Table of z to the 0·1 per cent. level of significance. Such high levels of significance are especially useful when the test we make is the most favourable out of a number which *a priori* might equally well have been chosen.

Two changes in the seventh edition may be mentioned. Section 27 was expanded so as to give a fuller introduction to the theory of orthogonal polynomials, by way of orthogonal comparisons between observations, which most practical workers find easier to grasp. The arithmetical construction is simpler by this path, and the full generality of the original treatment can be retained without very complicated algebraic expressions. A useful range of tables giving the serial values to the fifth degree is now available in *Statistical Tables.*

Section 49·2 was added to give an outline of the important new subject of the use of multiple measurements to form the best discriminant functions of which they are capable. The tests of significance appropriate to this process are approximate and deserve further study. The diversity of problems which yield to this method is very striking.

A section new in the ninth edition is given to the test of homogeneity of evidence used in estimation, since this subject is the natural and logical complement to the methods of combining independent evidence illustrated in the previous examples. In the tenth edition is an extension of the t-test to find fiducial limits for the ratio of means or regression coefficients (Section 26·2).

The sections of Chapter VIII, the Principles of Experimentation, which have always been too short to do justice to aspects of the subject other than the purely statistical, have since developed into an independent book, *The Design of Experiments* (Oliver and Boyd, 1935, 1937, 1942, 1947, 1949, 1951, 1953). The tables of this book, together with a number of others calculated for a variety of statistical purposes, with illustrations of their use, are now available under the title of *Statistical Tables* (Oliver and Boyd, 1938, 1943, 1948, 1953, 1957). Both of these publications relieve the present work of claims for expansion in directions which threatened to obstruct its usefulness as a single course of study. The serious student should make sure that these volumes also are accessible to him.

Since the middle of this century a flood of literature has appeared bearing on statistical methods. The authors are largely in mathematical teaching departments, and better trained as mathematicians than some of their predecessors. Too often, however, their experience has not included the training and mental discipline of the natural sciences, and much space is given to the trivial and the irrelevant. Competition also has led to methods of publicity with propagandist zeal. Though the methods of this book, and of the *Design of Experiments* have been widely used, the underlying logic has been often misapprehended, and erroneous numerical tables have been published. In 1957, therefore, the previous books were supplemented by one devoted to the logic of induction, under the title of *Statistical Methods and Scientific Inference*. Detailed and explicit demonstration is then given of the logical concepts such as fiducial probability, taken for granted in this book, as distinct from " Decision Functions," " Inverse Probability " and like approaches.

It should be noted that numbers of sections, tables and examples have been unaltered by the insertion of fresh material, so that references to them, though not to pages, will be valid irrespective of the edition used.

DEPARTMENT OF GENETICS, CAMBRIDGE
1958

CONTENTS

TABLES

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 had usually been overlooked, the more naturally because the development of the underlying mathematical theory had 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" 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 always 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. In Wave Mechanics this is now clearly recognised. 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 unfortunate misapprehension that statistics is to be regarded as a branch of economics, whereas in truth methods adequate to the treatment of economic data, in so far as these exist, have mostly been developed in the study of biology and the other sciences.

The idea of a population is to be applied not only to living, or even to material, individuals. If an observation, such as a simple measurement, be repeated indefinitely, 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 possible 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 standard errors shows a deliberate attempt to learn 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 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* was the object of researches up to 1915. 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; the number of classes in which the population is distributed may be finite or infinite; again, in the case of quantitative variates, the intervals by which the classes differ may be 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. 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