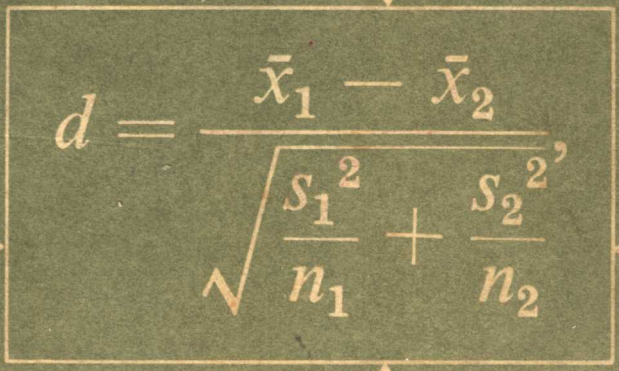
STATISTICAL METHODS IN BIOLOGY



N. T. J. Bailey

STATISTICAL METHODS IN BIOLOGY

by

NORMAN T. J. BAILEY, M.A., D.S.C.

Reader in Biometry, University of Oxford;
Formerly Statistician to the Medical School,
University of Cambridge

THE ENGLISH UNIVERSITIES PRESS LTD
102 NEWGATE STREET
LONDON, E.C.1

First printed 1959

PRINTED AND BOUND IN ENGLAND
FOR THE ENGLISH UNIVERSITIES PRESS LTD
BY HAZELL WATSON AND VINEY LTD, AYLESBURY

PREFACE

The purpose of this book is to provide workers in the biological and medical sciences with an elementary account of the chief statistical methods liable to be needed in practice. Mathematical symbolism has been reduced to the bare minimum required for a concise description of the essential types of analysis. The reader need have no more than an elementary knowledge of algebra. No trigonometry, geometry or calculus is used. In most cases the main text gives a fully worked numerical example of each different type of analysis, in addition to the general discussion and summarising formula.

An important feature of this book is the *Summary of Statistical Formulae*. This is intended to be used for day-to-day reference by the worker who already has some acquaintance with statistical ideas and who merely requires to have his memory refreshed. It is important that the formulae in the *Summary* should not be applied blindly and automatically without proper regard for their suitability. If there is any doubt about the correct way to handle any particular statistical situation, advice should be sought from a qualified statistician. A selection of the most frequently used statistical tables is also provided in an appendix, so as to make the book reasonably self-contained for the purpose of everyday use.

I should like to thank the English Universities Press for asking me to write this book, which has, I believe, a useful function to fulfil. I am also greatly indebted to Professor W. S. Bullough, Dr. H. B. D. Kettlewell and Mr. A. M. Walker, all of whom read the first draft and made many useful suggestions as to content and presentation. In particular, Professor Bullough was responsible for encouraging the development of the *Summary of Statistical Formulae*, intended for purposes of quick reference.

The Appendix Tables have been derived from two well-known sources. I am accordingly indebted to Professor Sir Ronald A. Fisher, Cambridge, to Dr. Frank Yates, Rothamsted, and to Messrs. Oliver & Boyd Limited, Edinburgh, for permission to abridge Tables I, III, IV, V and VI from their book *Statistical Tables for Biological, Agricultural and Medical Research.* Some additional material has

also been incorporated from Tables 12 and 18 of *Biometrika Tables for Statisticians*, Vol. I, by permission of the *Biometrika* Trustees.

Finally, I should like to thank Mrs. Jill Esnouf for drawing the two figures, Mrs. Tamara Hazlewood for carrying out the arithmetical calculations, and Mrs. Daphne Russen and Miss Marian J. Smith for preparing the typescript.

Oxford, NORMAN T. J. BAILEY
January 1959

CONTENTS

CHAPTER 1

INTRODUCTION

The subject-matter of the biological and medical sciences is remarkable for its richness and complexity. Moreover, the wide range of variation observed in both organisms and their environments is frequently analysable into simpler components only with great difficulty. Suppose we want to compare the behaviour of two different animal populations. Not only does each population consist of many individuals differing amongst themselves with regard to factors like sex, age, physical measurements, coloration, susceptibility to disease, aggressiveness etc., but the patterns of behaviour in which we are interested may themselves be fairly complicated. For these reasons, much biological work tends to be comparatively qualitative in nature. In the more exact sciences of physics and chemistry, on the other hand, we find that irreducible variation is usually fairly small, and often consists of little more than experimental errors. The latter can, as a rule, be virtually eliminated by averaging over several repeated determinations.

In the biological sciences, therefore, inherent variation must be accepted as basic, and must be handled as such. This certainly makes numerical arguments more difficult. We may talk about the average number of eggs laid by a certain species of bird under particular environmental conditions or the proportion of subjects protected by an immunising vaccine. But these average figures conceal the fact that specific instances may easily show very different results. Some females will produce very large clutches, others will not lay at all. Most vaccinated subjects may be free from infection, but a number of unvaccinated may also escape. This sort of thing makes it difficult to know how much reliance can be placed on the averages; the results in a particular instance might be quite unpredictable. Thus, when comparing the clutch sizes in two species of birds, or the average attack-rates for inoculated and uninoculated subjects, we might be uncertain whether the differences observed were in some sense real or whether they were only due to chance

variations. On the whole, one usually feels that conclusions based on large numbers of observations are more reliable than those based on small numbers. But the question still remains—how large a number is required for adequate reliability? Moreover, how does one measure this reliability?

To some extent, the expert worker in any field learns from experience how to deal with such difficulties. And the continued progress of biological science shows that he is not entirely unsuccessful in his efforts! However, it is frequently advantageous to try to use more precise methods of describing the basic variability, of deciding whether apparent differences are due to chance or not, of estimating unknown constants and so on.

This is where one turns to statistical methods. Some people think that the great variation present in biological material makes statistical methods unreliable. In fact, very nearly the opposite is true. It is precisely because modern statistics is based on a recognition of this variation that it is such a powerful tool for handling numerical data. Great quantities of complicated experimental results can often be reduced to more manageable proportions by the calculation of a few numbers which characterise the whole pattern of events. Again, the application of probability theory, itself based largely on refinements of intuitive common-sense ideas, means that we can assess the odds that some apparent effect is or is not due to chance, or that the unknown true value of some constant lies between certain limits.

Now, it so happens that the application of statistical methods requires comparatively little mathematical knowledge or ability. The majority of procedures required in practice have been reduced to simple arithmetical calculations, many complications being avoided completely by the use of appropriate mathematical tables.

The present book gives a range of elementary methods which should provide a biologist with what he is likely to require for perhaps 95 per cent of the time. Thus a relatively small number of methods will do duty for a large number of situations. It is of paramount importance to understand the general conditions under which any particular method can be used. Statistical tests should never be applied automatically without first giving some thought to their validity. Again, it is more important to be able to recognise the occurrence of some non-standard situation than it is to be able to apply the proper method of analysis. Once the situation is recognised,

expert statistical advice can be sought; if it is not recognised, erroneous and misleading methods may be used.

Another reason why it is useful for the biologist to have some familiarity with statistical ideas is the following. When mathematical difficulties arise, or when there is some doubt as to the best experimental design to be adopted, etc., the advice of a statistician will often be sought. Now, you may usually trust the professional statistician to work out correctly the theoretical consequences of the assumptions of a particular theory. The important point is to know whether the statistician is solving the right problem! If the biologist knows something of the general principles of statistical methods, he can cross-question the statistician who is helping him sufficiently closely to discover whether the latter is on the right track.

Although some of the methods used to deal with complicated experiments may be incomprehensible to the non-mathematician, the final results should be expressible in a form that the biologist can readily appreciate. It is usually safe to distrust any result that cannot, after a reasonable amount of discussion and explanation, be put in terms of the original biological problem. Sometimes it may appear that the wrong problem was tackled or the wrong questions asked, but in such cases adequate explanations ought to be forthcoming.

The general plan of this book is to show first, in Chapter 2, how statistics can describe and handle whole ranges of variation as opposed to mere averages. Then in Chapter 3 we see how more or less uncertain numbers, like averages calculated from relatively limited samples of data, are related to the 'true' values we should get from extremely large samples. Chapter 4 introduces the basic idea of a statistical significance test which is used to decide whether observed differences between factors are likely to be real or due only to chance. Chapters 5 and 6 then describe a number of the most frequently used significance tests. We next consider, in Chapters 7 and 8, ways of testing whether different groups of data are homogeneous, and whether experimental observations agree sufficiently closely with theoretical values for the theory to be regarded as reasonably adequate. After this the problem of associated measurements is introduced, such as occurs when variation in one quantity, like the length of an organism, is closely connected with variation in another quantity, like the organism's weight. The simple case of

a pair of measurements is dealt with in Chapter 9 on correlation and Chapter 10 on regression. More advanced methods of coping with several measurements at once are outlined in Chapter 14.

Chapters 11 and 12 discuss the important questions of how, in specific experiments, it is often possible to choose a pattern of experimentation that is not only highly informative but leads to types of analysis that are simple to perform and interpret. In Chapter 13 we see how to avoid bias in collecting observations or in allocating experimental units, and the final Chapter 15 gives some advice on numerical work and the use of calculating machines.

The succession of subjects introduced is intended to provide a graded course of instruction in elementary statistical methods. Those who are already familiar with some of these may, of course, pick out any chapters in which they are specially interested. Most of the methods recommended are comparatively easy, though they may require a certain amount of practice before a real facility in their application is achieved. Only Chapter 14 is likely to present any real arithmetical difficulty. But the principles involved are important, and should be understood even if the labour of computation is delegated to others.

An important feature of this book is the *Summary of Statistical Formulae*. This is intended for use as a quick-reference guide by the reader who already has some knowledge of statistics. It cannot be emphasised too strongly that standard formulae should not be applied blindly without some understanding of their suitability. Nevertheless, many workers who have already acquired a basic training in statistics, either from this book or elsewhere, will frequently require only to have their memories refreshed.

The five Appendix Tables provided are to enable the reader to carry out the commonest statistical tests without special reference to more extensive compilations. There is, however, some advantage in possessing a good set of tables for more general application. The best collection is probably Fisher & Yates' *Statistical Tables for Biological, Agricultural and Medical Research*. Another useful book is Volume I of *Biometrika Tables for Statisticians*, which contains a rather larger number of tables, many of them not readily available elsewhere. For day-to-day laboratory use the small and cheap *Cambridge Elementary Statistical Tables*, by D. V. Lindley and J. C. P. Miller, can be recommended. *Barlow's Tables* of squares,

cubes, square roots, cube roots and reciprocals is also an extremely useful aid to numerical work.

Finally, a word should be said about further reading. This book attempts to provide the groundwork basic to most statistical methods. However, those workers who are closely concerned with special fields will want to know something about more advanced methods. It is possible in the subject of experimental design, for example, to learn to use relatively sophisticated patterns of experimentation without becoming involved in higher mathematics. To some extent the choice of text-books is a personal one, depending on the reader's own interests and way of looking at things. Specific recommendations are thus liable to be difficult. The section on *Suggestions for More Advanced Reading* therefore includes a variety of statistical books, some of them rather specialised, which the reader may find useful to consult.

CHAPTER 2

VARIABILITY AND FREQUENCY DISTRIBUTIONS

We have seen in Chapter 1 how considerable natural variation is inherent in the subject-matter of practically all biological and medical work. It cannot be effectively disposed of by taking a few averages and then regarding these as more or less precise measurements. We must learn to handle the whole pattern of variation as such. The present chapter introduces some of the more common patterns, and shows how these can be described in fairly simple numerical terms. A clear idea of the basic attitude involved in looking at one's data from this point of view is essential to a proper understanding of the elementary statistical methods recommended in later chapters.

2.1 The normal distribution

We shall begin by considering some simple continuously variable quantity like stature. We know this varies greatly from one individual to another, and may also expect to find certain average differences between people drawn from different social classes or living in different geographical areas, etc. Let us suppose that a socio-medical survey of a particular community has provided us with a representative sample of 117 males whose heights are distributed as shown in the first and third columns of Table 1.

We shall assume that the original measurements were made as accurately as possible, but that they are given here only to the nearest inch. Thus the group labelled '65' contains all those men whose true measurements were between 64·5 and 65·5 inches. One is liable to run into trouble if the exact methods of recording the measurements and grouping them are not specified exactly. In the example just given the mid-point of the interval labelled '65' is, in fact, 65 inches. But suppose that the original readings were made only to the nearest half inch and then 'rounded up', i.e. $64\frac{1}{2}$ then included with 65. The

TABLE 1. *Distribution of stature in 117 males*

Absolute height (in.)	*Measurements from working origin at 67 in. (x)*	*Number of men observed (f)*	*Contributions to the sum (fx)*	*Contributions to the sum of squares (fx^2)*
61	−6	1	−6	36
62	−5	3	−15	75
63	−4	6	−24	96
64	−3	8	−24	72
65	−2	13	−26	52
66	−1	18	−18	18
67	0	19	0	0
68	1	14	14	14
69	2	14	28	56
70	3	9	27	81
71	4	5	20	80
72	5	4	20	100
73	6	2	12	72
74	7	1	7	49
Totals		117	+15	801

interval '65' would then contain all measurements lying between 64·25 and 65·25 inches, for which the mid-point is 64·75 inches. The difference of a quarter of an inch could lead to serious inaccuracy in certain types of investigation.

A convenient visual way of presenting such data is shown in Fig. 1, in which the area of each rectangle is, on the scale used, equal to the observed proportion or percentage of individuals whose height falls in the corresponding group. The total area covered by all the rectangles therefore adds up to unity or 100 per cent. This diagram is called a *histogram*. It is easily constructed when, as here, all the groups are of the same width. It is also easily adapted to the case when the intervals are unequal, provided we remember that the *areas* of the rectangles must be proportional to the numbers of units concerned. If, for example, we wished to group together the entries for 72, 73 and 74 inches, totalling 7 individuals or 6 per cent of the total, then we should need a rectangle whose base covered 3 inches on the horizontal scale but whose height was only 2 units on the vertical scale shown in the diagram. In this way we can make allowance for unequal grouping intervals, but it is usually less troublesome

if we can manage to keep them all the same width. In some books histograms are drawn so that the area of each rectangle is equal to the actual number (instead of the proportion) of individuals in the corresponding group. It is better, however, to use proportions, as different histograms can then be compared directly.

The general appearance of the rectangles in Fig. 1 is quite striking, especially the tall hump in the centre and the rapidly falling tails on each side. There are certain minor irregularities in the pattern, and

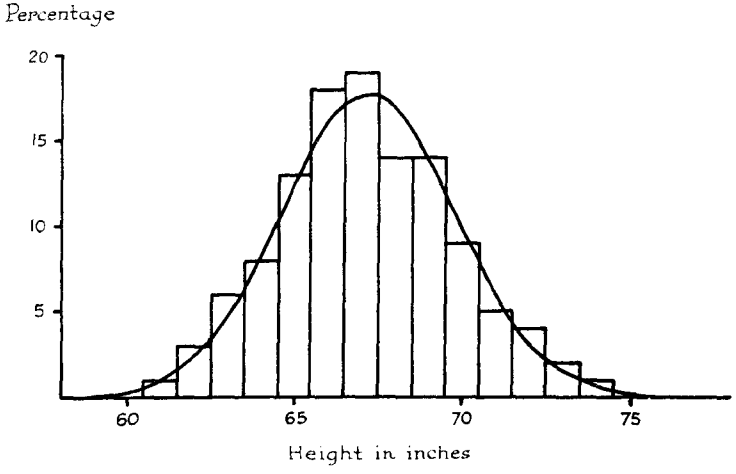


FIG. 1.—The diagram shows the observed distribution of the heights of 117 males, exhibited in the form of a histogram (rectangles), together with a fitted 'normal' curve (smooth curve).

these would, in general, be more pronounced if the size of the sample were smaller. Conversely, with larger samples we usually find that the set of rectangles presents a more regular appearance. This suggests that if we had a very large number of measurements, the ultimate shape of the picture for a suitably small width of rectangle would be something very like a smooth curve. Such a curve could be regarded as representing the true, theoretical or ideal distribution of heights in a very (or, better, infinitely) large population of individuals.

What sort of ideal curve can we expect? There are several theoretical reasons for expecting the so-called *Gaussian* or *'normal' curve*

to turn up in practice; and it is an empirical fact that such a curve often describes with sufficient accuracy the shape of histograms based on large numbers of observations. Moreover, the normal curve is one of the easiest to handle theoretically, and it leads to types of statistical analysis that can be carried out with a minimum amount of computation. Hence the central importance of this distribution in statistical work.

The actual mathematical equation of the normal curve is

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

where μ is the *mean* or *average* value and σ is the *standard deviation*, which is a measure of the concentration of frequency about the mean. More will be said about μ and σ later. The ideal variable x may take any value from $-\infty$ to $+\infty$. However, some real measurements, like stature, may be essentially positive. But if small values are very rare, the ideal normal curve may be a sufficiently close approximation. Those readers who are anxious to avoid as much algebraic manipulation as possible can be reassured by the promise that no *direct* use will be made in this book of the equation shown. Most of the practical numerical calculations to which it leads are fairly simple.

Fig. 1 shows a normal curve, with its typical symmetrical bell shape, fitted by suitable methods to the data embodied in the rectangles. This is not to say that the fitted curve is actually the true, ideal one to which the histogram approximates; it is merely the best approximation we can find.

The normal curve used above is the curve we have chosen to represent the *frequency distribution* of stature for the ideal or infinitely large *population*. This ideal population should be contrasted with the limited *sample* of observed values that turns up on any occasion when we make actual measurements in the real world. In the survey mentioned above we had a sample of 117 men. If the community were sufficiently large for us to collect several samples of this size, we should find that few if any of the corresponding histograms were exactly the same, although they might all be taken as illustrating the underlying frequency distribution. The differences between such histograms constitute what we call *sampling variation*, and this becomes more prominent as the size of sample decreases.