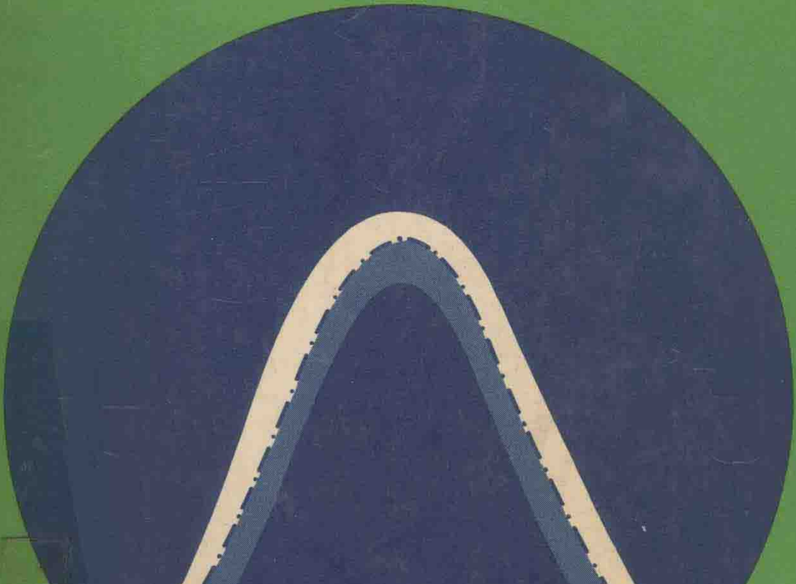
Statistics and Experimental Design

SECOND EDITION

Geoffrey M. Clarke

CONTEMPORARY BIOLOGY



Statistics and Experimental Design

Geoffrey M. Clarke
M.A., Dip. Stat. (Oxon.)

Reader in Statistics
Mathematics Division, University of Sussex

Formerly Statistician at the Long Ashton Research Station
and Lecturer in Statistics, University of Bristol

Edward Arnold

First published 1969
by Edward Arnold (Publishers) Ltd.
41 Bedford Square,
London, WC1B 3DQ

Reprinted 1971, 1973
Reprinted with corrections and additional exercises, 1976
Reprinted 1977
Second edition 1980

Paper edition ISBN: 0 7131 2797 X

Printed and bound in Great Britain at
The Camelot Press Ltd, Southampton

A series of student texts in

CONTEMPORARY BIOLOGY

General Editors:

Professor Arthur J. Willis
Professor Michael A. Sleigh

Preface to the Second Edition

In the ten years that have elapsed since this book was first written, statistics has found its way into a number of biology courses at pre-university and college levels. When students reach university or college courses, however, there still appear to be the same needs and problems as in the past. Most students of biology do little mathematical or numerical work, and find difficulty in thinking in symbols, so that they need to learn statistical methods in an 'applied' way, seeing the use of the methods and the reasons for them. In fact this is a very good way of learning statistics, because many more gross errors in applying statistical methods arise from not knowing when they work than from getting the arithmetic wrong (though there are enough of these, too!).

There has therefore seemed to be no reason to change the original plan, of aiming to present the ideas behind, and limitations of, standard statistical procedures as well as the methods themselves. It must be admitted that people sometimes appear to find this logical thinking about the methods just as hard as understanding what the symbols in the standard formulae mean; but there is no substitute for thinking before using *any* scientific method, statistical or otherwise. In biology it is often possible, with suitable forethought, to plan to a fairly detailed extent the collection of data and the conduct of experiments, and so emphasis is placed in this book on the elementary aspects of experimental design. To study the basic principles of experimental design and analysis helps in understanding some of the commonly-used tests of significance, which were designed for such situations originally.

Some changes have been made in this edition. The question 'Which is the right test for my data?' is often asked, and when it has been sharpened up by thinking what features of the data or the problem are the important ones it can be answered: an attempt to do this is made on pp. 157–8. It is hoped that this will save readers much time in locating the methods as they need them. Additions to Chapters 10 and 17 increase the material on nonparametric methods, though by no means to the level of a catalogue of all the Thirty Nine Steps (or is it Fifty Seven Varieties?) in the nonparametric world; there are of course well-defined problems in which such methods are essential, but not so many of these problems are important biological ones. In experimental design, factorial schemes frequently have factors at more than two levels, and a new section extends the original Chapter 15 on this topic.

Worked Examples are included in the text; and there are answers and comments, collected at the end of the book, for the Exercises that follow the chapters. A student working on his own would be well advised to attempt all the Exercises before moving to a new chapter.

Many of the principles of elementary statistical methods apply quite generally over a wide range of sciences, and it is hoped that the book may be of use to scientists in disciplines other than biology, provided they do not mind the examples being from biology or biochemistry—no great depth of knowledge in these subjects is required to appreciate them!

I am indebted to the Biometrika Trustees for permission to reprint parts of Tables 8, 12, 13 and 18 from *Biometrika Tables for Statisticians*, third edition; also to the Literary Executor of the late Sir Ronald A. Fisher, F.R.S., to Dr Frank Yates, F.R.S. and to Oliver & Boyd Ltd., Edinburgh, for permission to reprint parts of Tables III, V and VI from their book *Statistical Tables for Biological, Agricultural and Medical Research*, sixth edition.

I would like to acknowledge the help and encouragement received from Professor Arthur Willis, General Editor of the series, and from the Publishers, when I first wrote this book, and also their continuing interest and the various comments which I have received from users. All these have, I hope, helped to reduce the number of errors or obscurities that might otherwise be present. Comments will continue to be welcome.

The University of Sussex G. M. C.
Falmer, Brighton, 1980

Symbols

r_i	= a typical value of a discrete variate r
x_i	= a typical value of a continuous variate x
f_i	= the frequency with which x_i occurs in a sample
$\bar{x}$ or $\bar{r}$	= the mean of a sample of observations
N	= the number of observations in a sample
M	= the median of a sample of observations
s^2	= the variance of a sample of observations
μ, σ^2	= mean and variance of a probability distribution
$\mathcal{N}(\mu, \sigma^2)$	= the normal distribution whose parameters are μ, σ^2
n, p	= the parameters in a binomial distribution
q	$= 1 - p$
$\hat{p}$	= estimate of p calculated from a sample
λ	= the parameter (mean) of a Poisson distribution
$\hat{\lambda}$	= estimate of λ calculated from a sample
N.H.	= Null Hypothesis upon which the calculations of a significance test are based
A.H.	= Alternative Hypothesis, accepted when N.H. is rejected
d.f.	= degrees of freedom
$\boldsymbol{t}_{(f)}$	= Student's $\boldsymbol{t}$ distribution with f degrees of freedom
$\chi^2_{(f)}$	= the χ^2-distribution with f degrees of freedom
$\mathrm{F}_{(m, n)}$	= the F (variance-ratio) distribution with m and n d.f.
ρ	= correlation coefficient in a population
$\boldsymbol{r}$	= estimate of ρ calculated from a sample
$\boldsymbol{r}_s$	= Spearman's rank correlation coefficient
b	= slope of regression line
$\hat{b}$	= estimate of b calculated from a sample
d	= the unit (standard) normal deviate (often also called z)
S.S.	= sums of squares of deviations about the mean
M.S.	= mean square (S.S./d.f.)
$\hat{\sigma}^2$	= estimate of variance calculated in analysis of variance
e_{ij}	= error term in a linear (analysis of variance) model

Introduction

Experiments in the physical sciences often aim to estimate a numerical *constant*, and a student whose estimate is far away from the known true value has made an error at some stage of the experiment. Perhaps the error arises in using the equipment, or in not allowing for some change from normal environmental conditions when carrying out the experiment and the calculations involved.

In biology, a further important factor must be recognized. This is also true of other subjects as widely different as industrial experiments in engineering, metallurgy, textile production or pharmaceutical chemistry; so statistical methods developed for one subject very often apply to another. The all-important common factor is that individual items of material, individual plants or animals or engineering specimens or units of textile, *vary among themselves naturally even when they are treated alike.* Two plants, grown originally from the same batch of seed in identical pots side by side on a greenhouse bench, and given the same amount of fertilizer and water over the same period, will not grow to exactly the same height; nor will twin animals of the same sex, living in the same cage and receiving the same diet, put on exactly the same amount of weight. ***Experimental Error*** in this sense (a universal but perhaps unfortunate phrase) means the natural variation which is present among the individuals or units concerned, even when they are treated alike and are in identical conditions.

Now suppose that we examine a considerable number of individuals. We may measure the heights of many plants, or we may weigh many animals of the same type and so collect a whole set of height or weight records, from a whole ***population*** of individuals. When looking at the natural variation present in these records of height or weight, we often see that it has a pattern: for example, some plants are much taller than the general average, some much shorter, but the majority are fairly near this average. Biological data can often be explained by one of a few standard statistical patterns of natural variation—***distributions***, as these patterns are called: thus we shall speak of the distribution of height through a population of plants, meaning the way in which height varies when we measure a number of similar plants all growing in the same environment and conditions.

In order to make sense of a whole population of records, we often try to see the pattern in them by drawing a suitable diagram. Also we shall want to summarize the information about the whole population by calculating suitable numbers: a good method of summarizing a large collection of measurements of heights of plants is simply to calculate their *average* height. Frequently, in experiments, two populations of plants will be exactly comparable except for one factor: for example, one may receive a standard fertilizer treatment while the other has this plus additional nitrogen, growing conditions and environment being otherwise very similar. We shall want to compare these two populations, and it is very often useful to do this by comparing the two average heights. Finally, we may not have very many plants available for an experiment, or very much time or labour available to make large numbers of measurements, and so it is important for biologists to study the statistical methods of making comparisons based on small numbers of observations.

Table of Contents

I

Populations, Samples and Variates

POPULATIONS

A population may consist of things rather than of people. We have mentioned, in the Introduction, a population of plants; if the height of each plant is measured, the whole set of heights may also be called a population—this time it is a population of measurements.

Before a definition of the word 'population' is attempted, let us consider a few examples.

1 Humans, of specified age-group, racial group and sex may each have their height measured.
2 A field of wheat may be divided into square units ('plots') of fixed size, the crop yield on each plot being weighed.
3 Mice may be bred under controlled conditions in an animal house, and the coat colour of each mouse noted.
4 A properly balanced six-sided die, whose faces are numbered 1, 2, 3, 4, 5, 6, may be thrown a large number of times and the score for each throw recorded.

In (1), the people form a population, and from them we collect a population of height measurements; in (2) the 'plots' form the population (not the individual plants, because we did not weigh each separately) and give us a population of weights; in (3) the animals form the population, from which a population of observations of colour (not, this time, of numbers such as measurements of height or weight) is collected; and in (4) there is only one die, but it is thrown many times to give a population of throws, and hence a population of scores.

We shall use the word 'population' to mean a set of individuals or objects, upon each member of which a numerical measurement is taken or an observation of a particular characteristic made; and we shall also call the set of measurements or observations a population.

Membership of a population must be precisely specified: in (1), we know that, on average, men are taller than women, sons taller than fathers, and some races taller than others, so unless we do have a fairly rigid specification we shall be measuring different ages, races and sexes and collecting a very heterogeneous population of measurements in the process. In (2), the crop records would only be useful if we knew that seed had been sown uniformly and the same variety (or cultivar) of wheat used throughout; in (3) animals should live in the same environment and should be similar genetically (descendants of the same original animals); in (4) the same die should be thrown each time, with adequate shaking to prevent each new result depending on previous ones. Unless we specify a population carefully, we often find some 'unusual' results which, it may be claimed, are not really part of this population; in scientific work we cannot have arguments of this sort casting doubt on the generality of results. When we can say that all measurements in an experiment were taken on members of a carefully specified population, then the experimental results do apply to that population.

Sometimes there are two distinct parts to a population, for example male and female animals, or some plants in a field given additional fertilizer and the remainder not. Then it is useful to take measurements separately on the two parts or *sub-populations*. It is of very little use to make measurements on a very heterogeneous population.

The members actually observed will almost always be a part only, and that often a very small one, of a large population of similar individuals or objects that could have been observed: the observed members form a small ***sample*** from a larger population. We probably could, with much labour, measure all English males between the ages of 20 and 25 to find their average height; but it would be much more satisfactory if we could measure only a sample, calculate the average height in this sample, and then apply the information from the sample in some way to refer to the whole population. We shall return to this; but quite clearly if the original population has not been precisely defined it will be impossible to produce a sample with any confidence that it represents the original. In some of the examples above, the *complete* original population is infinite and can be imagined only: we could go on breeding mice or throwing dice for ever, so that whenever we stop to look at the results these can be only a sample.

In the social sciences, much use is made of *surveys*, where people are asked questions about such things as their political sympathies or what sort of washing powder they use. There are some notorious pitfalls in obtaining information in this manner, and fortunately the biologist does not often need to work this way. But in a new field of inquiry, the basic problems may have to be discovered by this method: veterinary scientists have had to rely on the observations of farmers and practitioners to find out what may be the urgent problems needing attention, or the conditions which seem to be conducive to particular diseases or disorders.

VARIATES

The measurement (e.g. of height, crop yield, score with a die) or observation (e.g. coat colour of mouse) made on each member of a population or sample is called a ***variate*** (or sometimes, confusingly, a *variable*). Variates are of two types, ***continuous*** and ***discrete*** (or discontinuous). *Continuous* variates are those in which any value whatever (sometimes within certain upper and lower limits) is possible. For example, human height could be 174 cm, 172·72 cm, and even a figure such as 175·0498573 cm if we could measure so accurately: within the known range of human heights, no figure specified, to whatever length in decimals, is impossible. We shall denote the values of continuous variates by x. An example of a *discrete* variate is the score made on throwing a die: it can be only 1, 2, 3, 4, 5 or 6, and any other number is quite impossible—there is no such thing as a face having 1·58 dots on it. Thus we have a limited list of possible numerical results; we denote these by r for a discrete variate. Both of these types may be referred to as ***quantitative*** (i.e. measurable) variates, whereas the colour of a mouse's coat is an example of a ***qualitative*** (i.e. observable) variate; these latter variates can frequently be divided into a small number of definite categories, e.g. in mice, white coat, grey coat, brown coat. With sufficient care, quantitative variates can be measured accurately, but qualitative variates lend themselves to subjective errors in their compilation according to which observer actually records them, and so it is extremely important to use a qualitative variate only when there seems to be no equally good quantitative one.

RANDOM SAMPLING

A random sample may be defined for our purposes as one in which every member of the original population has an equal chance of appearing (readers will be able to refine the wording of this definition after the remarks on *probability*). Thus if we have five plants a, b, c, d, e, growing in pots and we decide to choose *at random* two of these for examination in the laboratory, we might equally well find ourselves with a, b; a, c; a, d; a, e; b, c; b, d; b, e; c, d; c, e; or d, e. This would, however, by no means be true if we were to look at the plants and choose two to *represent* the five. Suppose we have labelled them in order of height, *a* being slightly larger than *b*, *b* than *c*, and so on; if we pick *a*, we are most likely to feel that we should take a smallish one, say *d* or *e*, to match it, and we are unlikely to choose *b* deliberately. But if we are going to take a measurement on each of the two sample members, and use the average of these two measurements to make a general statement about what the average for the whole population of five plants might be, the statistical methods of *confidence limits* (Chapter 9) are needed, and these can be used only if the sample consists wholly of members chosen at random. The literature contains plenty of examples of unexpected, subjective errors which have crept in when observers have tried to do their own sampling visually.

In order to choose a random sample from a given (finite) population, it is first necessary to number the members of the population systematically, starting at 1. We next want to obtain a random sample of numbers. Finally we use as our sample the members bearing these numbers. A table of random digits is required, such as that given by Fisher and Yates[7] (Table XXXIII). The aim when producing a table of random digits is that every entry in it is equally likely to be any one of the digits 0, 1, 2, . . ., up to 9, irrespective of entries in any previous position. Random digit tables have been produced by various methods; some simple computer systems have standard programs for generating them when required. Whatever method is used, a set of tests (based on the χ^2 *distribution*, see Exercise 8.11) is made to check whether the digits do appear to be equally likely to come up. If one tries to write down a run of numbers haphazardly out of the head, it is very unlikely that they will conform to this (cf. Exercise 1.3).

Let us suppose that we have a table of random digits, a part of which (the starting-point chosen at random) is as follows:

0743552718349456231435822724961128859777434258800571209 . . . .

This could be used to select 20 members from 1000, which had already been numbered, by grouping the random digits in threes, 074, 355, 271, 834, 945, 623, etc., and choosing as the sample the members of the population carrying these numbers (if 000 appeared, this would be number 1000).

If, for example, number 74 appeared a second time before the full quota of 20 was obtained, there are two choices of procedure:

Sampling with replacement where the 74th member would actually be used a second time; or *Sampling without replacement*, which is used where the first procedure is impossible for physical reasons, e.g. if an amount of material corresponding to 20 plants was needed for a destructive laboratory test.

Statistical methods are simpler when sampling is with replacement, and we shall assume that this can be done. (However, in the most common situation where the sample forms a very small part, say no more than 5%, of the population, the corrections in the calculations needed *without replacement* are quite negligibly small.)

In everything which follows, we shall assume that populations have been properly specified and samples randomly chosen from them, so that we may concentrate attention on the variates measured or observed.

EXERCISES

(For Answers and Comments, see p. 160)

1.1 Consider how to define precisely these populations:

(a) of students, when an inquiry about expenditure on books is being carried out;

(b) of tomato plants grown in greenhouses, when the number of leaves produced during a given time is measured;

(c) of bean foliage in a smallholding, when numbers of black-fly are counted;

(d) of leaves as in (b) removed for chemical analysis to estimate nitrogen content.

1.2 What type of variate is shown in each part of Exercise 1.1, and in Example (2) at the beginning of the chapter?

1.3 Write down haphazardly a run of 200 digits. Count the number of 0's, 1's, . . . , 9's. Count also the number of times 0 is followed by 0, or by 1, or by 2, etc., and make similar counts for the digits 1 to 9. (The results can be used later as examples of the χ^2-test.)

1.4 How would you use a table of random numbers to select:

(a) 15 members from a population of 750?

(b) 10 members from a population of 250?

(c) 10 members from a population of 300?

(d) 20 sample units, each 1 ft^2, in a rectangular field 25 ft × 40 ft in size?

1.5 Suppose that, in a survey, the following variates have been recorded for several individual people: (a) age, (b) year of birth, (c) sex, (d) height, (e) weight, (f) colour of hair, (g) colour of eyes, (h) surname, (j) age rank in household (i.e. whether they are oldest, second oldest, . . ., youngest person in the household where they live), (k) whether they hold a driving licence. For each of the variates (a)–(k), state whether it is qualitative or quantitative; and, if quantitative, state whether it is discrete or continuous.

1.6 The following practical methods have been suggested for selecting random samples. Say whether each is likely to be satisfactory, and if you consider that it is not, then attempt to propose a better method.

(a) A list of all adults living in a village is available (in the form of the voters' list for elections). In order to survey a random sample of households in the village, to see what sort of fuel they use for domestic heating, names are picked at random from the voters' list.

(b) The doctor who serves this village has a card-index, arranged by families (one card per family). In order to survey the state of the teeth of the children in the village, a random sample of cards is taken, and if there are children in the family shown on a selected card then a random choice is made from among these.

(c) A horticulturalist wants to take a 1-in-10 sample of the strawberry plants growing in rows in a field. He does not have time to select several random numbers and so he chooses a starting-point at random among the first 10 plants in the first row. After this he takes every tenth plant down the first row, back along the second row, down the third row and so on through the field.

1.7 Using the run of random digits near the foot of p. 4 make a random selection of eight letters (without replacement) from the 26 letters of the English alphabet.