MEDICAL RESEARCH

A STATISTICAL AND EPIDEMIOLOGICAL APPROACH

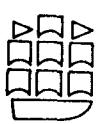
Medical Research

A Statistical and Epidemiological Approach

J. M. England, M.B., B.S., B.Sc., Ph.D.

Member of the Scientific Staff, Medical Research Council's Experimental Haematology Unit, Honorary Lecturer, Department of Haematology, St. Mary's Hospital Medical School, London

Foreword by
W. S. Peart, M.D., F.R.C.P., F.R.S.
Professor of Medicine,
St. Mary's Hospital, London, W. 2.



CHURCHILL LIVINGSTONE
EDINBURGH LONDON AND NEW YORK
1975

CHURCHILL LIVINGSTONE

Medical Division of Longman Group Limited

Distributed in the United States of America by Longman Inc., New York and by associated companies, branches and representatives throughout the world.

First published 1975

ISBN 0 443 01139 7

Library of Congress Catalog Card Number 74-75102

PRINTED BY Unwin Brothers Limited
THE GRESHAM PRESS OLD WOKING SURREY ENGLAND

Produced by 'Uneoprint'

A member of the Staples Printing Group

Foreword

The last twenty years have seen a remarkable growth and respect for numeracy in medicine. The change from clinical impressions and impressionable physicians to the measurement of the probability of accuracy in clinical observations and medical research, have been the areas of greatest advance. Statements of opinion now require statistical support if they are to command clinical backing and action. This all makes for better medicine and a properly sceptical approach, which is all for the good of the patient. But how to teach the student and doctor since most books on the subject lack that close relation to medicine which breathes life into what many have believed are dry bones. This book by a young and active medical research worker fills the need for a clear and explicit guide, linking laboratory, clinical and epidemiological medicine in an outstanding way. While the emphasis throughout is on the solution of practical problems, the author relates this approach to the planning, ethics and humanity of research. A most impressive work.

1974 W. S. PEART

Preface

'The teaching of statistics can contribute to the education of medical undergraduates in two main ways. First, the subject is an integral part of the logic of scientific method and can be conveniently used to introduce ideas about making and interpreting observations and about experimentation. Secondly, statistics comprises a body of techniques for the measurement and assessment of variation, used widely and increasingly in medical research on diagnostic procedures, effectiveness of treatment, development of new drugs, causative factors in disease, laboratory measurement and many other subjects.'

'Some knowledge of the principles of the statistical approach is now necessary so that doctors can make some judgement for themselves of the validity of the claims for medical advances made in journals and in other communications'

Royal Commission on Medical Education 1965-68

The Royal Commission on Medical Education and the Society for Social Medicine have both stressed the importance of medical statistics and epidemiology to the medical student and to the practising doctor. The above quotation well summarises the main reasons and I will not repeat them here.

More recently, the Royal Colleges of Physicians of the United Kingdom have expressed concern at the apparent lack of acquaintance of many post-graduates with the principles of elementary statistics, and have now included questions on statistics in their examinations. The medical research worker has also been criticised. The editors of Clinical Science and Molecular Medicine, for example, have commented in their 'Guidance for Authors' that 'papers are frequently returned for revision (and their publication consequently delayed) because the authors use inappropriate statistical methods'.

I hope that this book will provide a general introduction to the most important statistical and epidemiological aspects of medical research. The first two chapters explain basic statistical terms, and subsequent chapters tell the reader how to design an investigation (Chapter 3), how many subjects to study (Chapter 4) and how to analyse the results (Chapters 5-8). Later chapters are devoted to epidemiology and clinical trials, and the final chapter makes some suggestions about the assessment of an original publication.

Most doctors dislike mathematics and for this reason I have attempted to reduce the mathematics in the text, and place most of the formulae and calculations in a series of appendices at the end of the book. Examples are given of the calculations so that the reader can check that he is performing them correctly. Of course, many of the calculations are tedious, but they can often be avoided by using the simple desk computers which are now available.

More advanced material and some statistical tables are also included in the appendices.

Acknowledgements

I am especially grateful to Professor R. G. Miller, Jr., who read the manuscript of this book. His suggestions have been invaluable. I should also like to thank Professor P. L. Mollison, F.R.S., for his helpful comments, as well as Drs. Helen Dodsworth, M. A. England, P. M. Fraser, R. L. Himsworth, A. Howell, N. C. Hughes-Jones, T. W. Meade and S. N. Wickramasinghe. Dr. K. McPherson allowed me to reproduce Table 4.3, and Dr. P. M. Lambert advised me about standardised mortality ratios.

The manuscript was typed rapidly and faultlessly by Mrs. T. Alasya, and the figures were drawn with great care and attention by Mr. Paul Darton.

I also thank the Journal of the American Statistical Association for permission to reproduce statistical tables, and the Medical Research Council for allowing me to reproduce 'Responsibility in Investigations on Human Subjects'.

Mr. N. D. Palmer kindly checked the references.

Contents

CHAPTER 1

An introduction to the important terminology

POPULATIONS, SAMPLES AND VARIATES

In research it is never possible to study everyone with a certain disease or everyone who is normal. These large groups may be thought to be of infinite size and are technically termed populations. The researcher must be content with studying a small number of people from any population and these people would form what is termed a *sample*.

If, for example, one wishes to investigate whether the incidence or severity of atheroma is increased in diabetes, then one must draw a sample from the population of diabetics, and compare the findings with those in a normal sample. It is essential for the purposes of comparison to study a group of normal people, and they are said to form the *controls*.

Having selected the samples of diabetics and normals one must measure something as an index of atheroma. The parameter which is measured is termed a *variate*. Perhaps one could determine cholesterol in postmortem specimens of aorta, and express the results as milligrams of cholesterol per gram of aorta. If the measurements are rounded up or down to the nearest milligram, the scale of units is discontinuous and the results would be examples of a *discontinuous variate*. However, if the results are expressed extremely accurately, say to the nearest millionth of a milligram, then the scale of units is continuous for practical purposes, and we would speak of a *continuous variate*.

Another approach to estimating the severity of atheroma might be simply to record the presence or absence of a peripheral pulse such as the popliteal. This would be an *all or none variate*.

MEASURES OF LOCATION AND INDICES OF DISPERSION

The nature of the variate determines the way in which the results are analysed. We will now consider a typical continuous variate, such as the level of haemoglobin. Figure 1.1 shows the values of haemoglobin in 32 normal subjects. If we want to summarise the information contained in this graph we must first say whereabouts the results lie within the range of possible haemoglobin values.

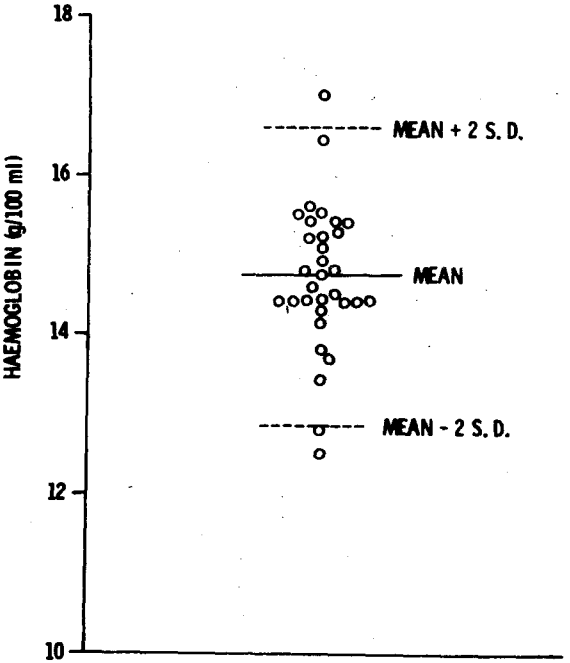


Figure 1.1 Haemoglobin values in 32 normal adult males. The mean is 14.7 g/100 ml, and the standard deviation is 0.9g/100 ml.

Technically speaking, we want a *measure of location*. Secondly we must know how varied the results are, and any estimate of this is termed an *index of dispersion*. A measure of location plus an index of dispersion can therefore provide a good summary of the results.

The most important measure of location is the average or arithmetic *mean*, defined as:

$$\text{Mean} = \frac{\Sigma x}{n} = \bar{x}$$

The Greek letter sigma, Σ, means the sum of, hence Σx means the sum of all the values of x. We divide Σx by the number of observations, n, to get the mean value of x which is written $\bar{x}$.

The most commonly used index of dispersion is the *standard deviation*, written s. The standard deviation is sometimes called the *root mean square deviation*, which tells us how it may be derived:

For every value of x calculate the deviation from the true population mean, μ. Note that the population mean is not the same as the sample mean $\bar{x}$, and that small Greek letters are generally used to denote population values whilst ordinary letters are reserved for sample values. The deviation is:

$$(x - \mu)$$

The squared deviation is:

$$(x - \mu)^2$$

The mean squared deviation is

$$\frac{\Sigma (x - \mu)^2}{n}$$

The root mean squared deviation or standard deviation is:

$$\sqrt{\frac{\Sigma (x - \mu)^2}{n}}$$

Using this formula, though, we are unable to calculate the standard deviation because we do not know the value of μ. The best estimate we have is the sample mean, $\bar{x}$, and a little thought will verify the fact that using $\bar{x}$ instead of μ would often tend to underestimate the standard deviation. This is because $\bar{x}$ will often differ from μ, and in this case the deviation of the results would be smaller in relation to $\bar{x}$ than they would be in relation to μ. An attempt to allow for this deficiency is to divide the sum of the squared deviations by $(n - 1)$ instead of by n.

$$\text{Standard deviation, } s = \sqrt{\frac{\Sigma (x - \bar{x})^2}{(n - 1)}}$$

This is the reason that a peculiar number such as $(n - 1)$ appears in the formula, and $(n - 1)$ is also termed the numbers of degrees of freedom ('degrees of freedom' is a difficult concept which is explained in Appendix 8 for those who wish to refer to it).

Having derived $\bar{x}$ and s, there are three other things which may be calculated.

$$\text{Variance} = s^2$$

$$\text{Coefficient of variation} = \text{C.V.} = \frac{s}{\bar{x}} \times 100\%$$

Standard error of the mean = S.E.M. $= \frac{s}{\sqrt{n}}$

Example The haemoglobin levels in g/100 ml of the 32 normal men shown in Figure 1.1 are:-

14.8, 15.4, 15.5, 13.7, 14.4, 14.1, 14.4, 14.4, 15.1, 15.3, 14.2, 14.8, 14.9, 14.3, 12.8, 13.4, 15.6, 14.6, 14.7, 14.4, 13.8, 15.4, 14.5, 16.4, 15.2, 12.5, 15.9, 15.5, 14.4, 17.0, 15.2, 14.4.

Calculate

$$\Sigma x = 14.8 + 15.4 + 15.5 \ldots + 17.0 + 15.2 + 14.4 = 471$$
$$\Sigma x^2 = (14.8)^2 + (15.4)^2 + (15.5)^2 \ldots + (17.0)^2 + (15.2)^2 + (14.4)^2 = 6959.6$$
$$n = 32$$

Then

$$\text{mean} = \bar{x} = \frac{\Sigma x}{n} = \frac{471}{32} = 14.7188$$

$$\text{Standard deviation} = \sqrt{\frac{\Sigma (x - \bar{x})^2}{n-1}} = \sqrt{\frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}}$$

$$= \sqrt{\frac{6959.6 - \frac{471^2}{32}}{32-1}} = 0.9344$$

Variance $= s^2 = 0.8731$

Coefficient of variation = C.V. $= \frac{s}{\bar{x}} \times 100\%$

$$= \frac{0.9344}{14.7188} \times 100\% = 6.3483\%$$

Standard error of the mean = S.E.M. $= \frac{s}{\sqrt{n}} = \frac{0.9344}{\sqrt{32}} = 0.1652$

The *variance* is a useful thing to calculate for many reasons. For example, if we wish to add two things together and see how variable the sum is, then we merely add the variances too. The variance of differences, sums, products etc. is summarised in Appendix 1.

The *coefficient of variation*, which is expressed usually as a percentage, can and should always be calculated. If the C.V.

exceeds 12 per cent it may well be that problems will arise in analysing the data (see Chapter 2, the lognormal distribution). The C.V. is also important because it is unaffected by the units which are used to express the results. Haemoglobin could be expressed as g/100 ml. of blood or g/litre of blood, the coefficient of variation would be the same.

The *standard error of the mean* is smaller than the standard deviation. It is derived from s by dividing s by the square root of the number of observations which are used to calculate the mean. If one imagines taking repeated samples of a certain size, and calculating the mean of each sample, then the S.E.M. is the standard deviation of the sample means. A smaller S.E.M. means that the sample mean is a more accurate estimator of the mean of the population from which the sample is derived. It is important to realise that the S.E.M. only decreases as the square root of the number of observations. Table 1.1 illustrates this point, which should be borne in mind by anyone collecting data for analysis.

Table 1.1 Effect of increasing sample size upon the standard error of the mean. Samples drawn from population with standard deviation of 100

Size of sample	*Standard error of mean*
1	100.0
2	70.7
3	57.7
4	50.0
16	25.0
64	12.5
256	6.25

Small initial increases in sample size repay well in terms of increased accuracy. Increasing the sample size from 1 to 4 halves the S.E.M. but it is necessary to increase the sample to 16 to halve it again. To reduce the standard error still further becomes progressively more difficult. We need a sample of 256 items to reduce the S.E.M. to 6 per cent of the standard deviation. Such large samples are rarely needed, and when needed may be both expensive and difficult to obtain.

All the most important indices of dispersion have therefore been described. Some important measures of location have not been described but are listed in Appendix 2. These are the median, mode, harmonic mean and geometric mean, and their role is also described.

THE NULL HYPOTHESIS AND SIGNIFICANCE TESTING

Having obtained our mean haemoglobin level from the information in Figure 1.1, we may then wish to consider whether the value which we have obtained is the same as a normal value of 14 g/100 ml which we have found in a textbook. To do this we set up the hypothesis that our observations are derived from a population of mean 14 g/100 ml and then calculate the number of times in which our result would occur by chance. The hypothesis that the sample is derived from a population of mean = 14 is termed the *null hypothesis*, and the calculation of the frequency with which our results would occur by chance is termed a *significance test*. The precise mathematical nature of the individual significance test need not concern us here, it will be dealt with in detail much later. All that we need to know is that the significance test tells us the number of times our result would occur by chance if the null hypothesis were true. Suppose we found that our results occurred by chance quite frequently, then it would be reasonable to say that the null hypothesis was true (i.e. there is no significant difference between the mean of our sample and 14). If our results only occurred by chance very infrequently, then the null hypothesis would be unlikely to be true, and it would be likely that the mean of our sample differed from 14.

Obviously when we calculate the likelihood of our result occurring by chance we must express our conclusion somehow. We could say that our result or a more extreme result would occur, for example, once in every 100 times that we took a sample from the population of mean 14. This mode of expression would be quite adequate. Alternatively we could say that the probability of our result or a more extreme result occurring is 1/100 or 0.01. The symbol for probability is p, hence we write $p = 0.01$. The scale for probability expressed in these terms can extend from 0 (never happens) to 1.0 (always happens).

There is nothing magical about a value of p, it is merely an expression of how frequently the experimental results would occur if the specified null hypothesis is true.

Most experimenters consider that if p is 0.05 or less then the null hypothesis is unlikely to be true. That is to say, there is a significant difference between the sample mean and 14 in our example. The choice of 0.05 is arbitrary, there being no reason why one could not insist on results reaching the 0.01 level of significance before accepting them.

PRACTICAL SIGNIFICANCE AND STATISTICAL SIGNIFICANCE

Gertrude Stein said, 'A difference to be a difference, must make

a difference'. A statistician would say that just because an experimental result reaches the desired significance level, this does not mean that the result has any practical significance. This may seem self evident, but one need only to look at the medical literature to find many examples of confused thinking about this point. A new drug for treating carcinoma may on average prolong survival for one day. If one studied enough patients a significant difference in survival could be shown between patients treated with the drug and those not treated, but would this be of any practical value?

Conversely, a treatment may prolong life by one or two useful years, but if the normal survival is very variable, and the person studying the drug does not investigate enough people, then the difference between the treated and untreated groups may not be statistically significant. In this way an important new drug could be missed. The way to avoid this difficulty will be discussed later (Chapter 6), but calculation of the confidence limits are often useful.

CONFIDENCE LIMITS

Calculation of confidence limits is the converse of calculating a significance test. With the significance test we ask how often a certain thing would happen by chance if a given null hypothesis is true. When we calculate confidence limits we stipulate the level of probability which interests us, and then determine the range of null hypotheses which will agree at the required probability level. It is common to use the 0.05 significance level. If for example we were working out the confidence limits for the mean of a sample we would calculate the number above the mean such that the null hypothesis was just rejected at the 0.05 level. This would be called the upper confidence limit, and likewise we could calculate the number below the mean such that the null hypothesis is just rejected, and this would be the lower confidence limit. Therefore the probability that the population mean is between the two limits is 0.95 (the sum of all possible probabilities is 1.0), and these limits are usually termed 95 per cent confidence limits.

We could therefore say that we were 95 per cent confident that our sample was drawn from a population whose mean was within a certain range.

The precise way to calculate these confidence limits varies under different circumstances and will be explained later. Confidence limits provide a very useful alternative to significance tests. In the previous section we mentioned that variable results could