Statistics in Medical Papers

D. G. ALTMAN, B.SC.

Department of Community Medicine,
St Thomas's Hospital Medical School, London

Clinical decisions are necessarily an amalgam of objective evidence and subjective opinion. But medical research, to be of value, must strive to be objective, and a growing awareness of this in medicine is shown by the increasing use of statistics. Deviation from fundamental statistical principles, however, is likely to produce erroneous results which may not be recognised by those without some statistical training. Unfortunately, despite the availability of several good books on the subject (Armitage 1971; Hill, 1971), there is still much published research displaying inadequate or incorrect understanding of statistics.

This article covers some of the most important aspects that ought to be considered by researcher and reader alike. The progress of a research project will be followed from the design to the drawing of conclusions. In general, experiments (including trials) are carried out to measure the effect of some action by the experimenter, whereas surveys observe existing relationships. Most of the following applies equally to experiments and surveys so that no particular importance should be attached to the choice of examples.

PLANNING AN INVESTIGATION
Experiments and surveys are usually carried out to measure effects or to test hypotheses. Since these intentions will initiate the research, it should be possible at the planning stage to decide on an appropriate design for the investigation, and to determine what analyses will need to be carried out.

If a statistician is to be involved, perhaps advising on or performing the analysis, it is best to seek his advice at the planning stage to ensure that the investigation will yield the desired information. If brought in after data collection he may not be able to overcome deficiencies in the design.

DESIGN
The design of an experiment includes —

1. the choice of treatments for comparison;
2. the choice of units (the sample);
3. the rules by which the treatments are allocated; and
4. the specification of the measurements to be recorded (Finney, 1955).
For surveys usually only 2 and 4 apply. Because of the importance of the design in determining what analyses will be possible and the validity of conclusions (including consideration of whether they answer the original questions), the benefit of statistical advice is greatest at this stage.

Much emphasis has been placed on the importance of having a representative sample, but the importance of this depends on the design of the investigation and the subsequent analyses. If it is intended to extrapolate the findings to a larger population for example in the form of an estimate of prevalence, the sample will be very important. But if it is the intention to look only at relationships within the sample, as for example when comparing treatment and control groups, it is only necessary that the various groups are comparable. This is often achieved by matching or random allocation. The principle is to compare groups that are comparable except for the one factor being studied, be it an experimental intervention or an observed characteristic. However, such results could be generalised only if the sample was representative or, possibly if other studies produced similar results.

A common error is to omit a control group when one is needed. The results of a drug trial may be meaningless in the absence of a control or placebo group without which it is impossible to conclude that any observed improvement was due to the drug. Jellinek (1946) gave an excellent example of the importance of the placebo in the comparison of the effectiveness of three headache remedies. He found that people with headaches could be separated into those who responded to the placebo and those who did not. The first group found all three drugs and the placebo equally effective, but the second group differentiated between the effectiveness of the three drugs. Conclusions about the drugs were made from the results of the second group only.

Any errors at the design stage may well be irremediable later. However, if it is realised that the design of a study was deficient, the flaw should be admitted and its effect on the results estimated. There is a good example of this approach in the paper by 'Student' (1931) in which he attempted to retrieve something from the results of a pre-war milk feeding trial in which children were randomly allocated to feeding and control groups. The problem here was that some of the people in charge of giving the milk to the children gave milk to those in the control group who looked as if they needed it. However, because the deficiencies in the design were stated and discussed, it was possible to draw some conclusions from this study. Conversely, the absence of vital information, whether deliberate or not, will cast doubt on the respectability of a study. Obscure phrases such as 'nearly random' should be avoided.

The common neglect of correct experimental design was illustrated by Gifford and Feinstein (1969) who compared the methodology in 32 papers reporting the use of anticoagulant therapy for acute myocardial infarction. Less than half had carried out a trial and over a quarter did not have concurrent control groups. They
concluded that the controversy over the benefit of such therapy was largely due to the inadequacy of the design of most of the research.

The choice of design is made easier by a clear statement of the aims of the investigation and of any hypotheses to be tested. In general, proven and simple designs are to be preferred as they make for less complicated analyses. Full details of the design should always be presented so that readers can assess for themselves the validity of the subsequent analyses and the credibility of the conclusions.

ANALYSIS
The aim of experiments in medical research is sometimes to measure an ‘effect’ such as a dose-response relationship, and this is usually a descriptive procedure. But both experiments and surveys are frequently set up to test some hypothesis about the data, and it is here that many problems arise.

The hypothesis, which should have been formulated at the design stage, will determine which analysis is appropriate. Suppose we are investigating the association between age and bronchitis and have a frequency table giving the number of people with or without bronchitis in several age groups. If our hypothesis was that the prevalence of bronchitis is different for different age groups then the appropriate analysis would be to perform a $\chi^2$ test. However, if our hypothesis was that there was a significant trend in the prevalence of bronchitis by age then we should use the $\chi^2$ test for trend (Armitage, 1971), which is quite different. It is essential to match the test to the hypothesis.

The t test (for comparing means), which is the next most common in the medical literature, also has several forms, and it is not uncommon for the wrong one to be used. Again, the correct version will depend on the hypothesis.

Suppose we have two sets of observed diastolic blood pressure readings for a group of 12 men taken before and after administration of a hypotensive drug —

| Blood Pressure | 1  | 2  | 3  | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 11 | 12 |
|----------------|----|----|----|----|----|----|----|----|----|----|----|----|
| Before         | 108| 95 | 89 | 85 | 99 | 111| 100| 102| 94 | 88 | 97 | 109|
| After          | 102| 90 | 88 | 86 | 97 | 107| 92 | 99 | 96 | 84 | 93 | 102|
| Difference     | 6  | 5  | 1  | -1 | 2  | 4  | 8  | 3  | -2 | 4  | 4  | 7  |

We wish to test the hypothesis that the drug is effective in reducing blood pressure. Because we have pairs of before and after readings we should compare the differences between the before and after readings (shown above) with zero. To do this we should use the paired $t$ test. We could use the ordinary $t$ test, which would compare the mean blood pressures before and after the drug was administered, but this would not test the stated hypothesis as it would not take account of the fact that we are looking at the same men each time.

It is easy for the wrong form of the $t$ test to be used but this cannot always be
detected in a paper because analyses are often described simply as ‘t tests’ or they may not be described at all.

For more complicated analyses the issues are perhaps not so clear, but the principle is the same: to convince the reader of the correctness of your methods it is necessary to supply unambiguous information on what the hypotheses were and what tests or analyses were carried out.

Most statistical tests (including the t test and $\chi^2$ test) and analyses are valid only under certain conditions. If these have not been met, this should be mentioned, with comments on the possible effect on results. Unusual analyses should be avoided, if possible. If they are used, their purpose should be explained and readers referred to an explanatory text.

**SIGNIFICANCE TESTS**

These merit further attention because the vast majority of analyses reported in the medical literature involve significance tests. Despite their widespread use they can cause conceptual problems.

Significance tests are used to assess the probability that a result found for a sample is also true in the population; in other words that it is a ‘real’ finding, not a chance occurrence. These usually compare a ‘null’ hypothesis, such as ‘treatments A and B are equally effective’ with an alternative hypothesis, such as ‘A is better than B’ or ‘A and B are not equally effective’. The form of the test will depend upon the alternative hypothesis. It is important to remember that statistics can never prove an hypothesis is false but only give the probability that this is so.

A significance test compares observed data with what would have been expected if the null hypothesis were true. It is a way of calculating the probability $p$ of achieving such a result under the null hypothesis. This indicates how likely is the null hypothesis. Before carrying out a significance test the experimenter should decide what level of $p$ will be considered significant. This level is denoted by $\alpha$, and is by convention most frequently 0.05 (that is, 5 per cent). If $p$ is less than $\alpha$ the data are said to be significantly different from what would have been expected under the null hypothesis. Accordingly, the null hypothesis is rejected and the alternative accepted. (Reject and accept are convenient ways of expressing extreme doubt and a high degree of certainty respectively. They are not as absolute as they sound.) However, in doing this, there is a small probability ($\alpha$) that the null hypothesis will be rejected when true. If $\alpha$ is 0.05, we are prepared to reject the null hypothesis in 5 per cent of the cases when it is true. Clearly this probability is reduced by reducing $\alpha$; many feel that $\alpha = 0.01$ is a safer criterion. The level is arbitrary and the choice of the researcher. If both $\alpha$ and $p$ are quoted for each test the reader of a paper can decide for himself how to interpret the result. This may be particularly helpful; for example, when a single significance test just fails to reach statistical significance. If $p$ is greater than $\alpha$, the result is not significant and the alternative hypothesis not proven (Armitage, 1971).
The second error is to fail to accept an alternative hypothesis when it is true. The probability of this is denoted by $\beta$. It is not usual to quote $\beta$ in a paper; indeed it is very often not considered when carrying out the test. However, it is worth knowing that $\beta$ can be quite large for small samples but decreases as the sample size $n$ increases. Also decreasing $\alpha$ (as above) will increase $\beta$.

As an example of some of these ideas consider the blood pressure data presented in the previous section. If we carry out a paired $t$ test with $\alpha = 0.05$ we get a probability $p$ less than 0.01. We interpret this as giving strong evidence that the drug is effective, but accept that there is a small probability ($p$) that our finding is a chance extreme occurrence and that the drug is not effective.

Because the sample size is considered in the calculation of $p$, even a very small effect (for example the difference between two groups) will be significant if the sample is large enough. Conversely, with a very small sample it may be impossible to obtain a significant result. It is advisable to specify the size of difference that would be considered important before the trial begins. Thus, the size of a trial can be calculated to give a specified probability ($\alpha$) that a difference of the specified size will give a statistically significant result.

**Presenting Results**

Visual methods, if properly presented, are often the best way of summarising results, especially where some trend or pattern is evident. But devices such as graphs or histograms are liable to be unintentionally misleading (Huff, 1973); graphs showing numerous lines or incorporating several different scales can be extremely confusing. Graphs should not be plotted on a logarithmic scale without some scientific justification, as this can greatly alter the visual impression.

In general, tables and figures should be able to be understood without reference to the text. Tables of frequencies are usually improved by including row or column percentages but it must be made clear which they are. Where mean values of measurements are given it is customary to include a measure of the variability of the data, usually the standard error or standard deviation — again, it must always be made clear which.

The tests should be specified. Feinstein (1974) found that 20 per cent of 757 statistical procedures mentioned in six leading medical journals over a period of six months were unidentified. As he observed: 'It is difficult enough for a clinician to interpret the meaning of a statistical procedure with which he is unfamiliar; it is much more difficult when he is not told what the procedure was'.

Schor and Karten (1966), in their study of statistical procedure used in 295 papers published in ten medical journals in 1964, found much the same and commented: 'Whenever a statistical test is applied or the word "significant" is used in the text, a description should be given of the test employed, how it was applied and what level of significance was used'. To these should be added the
probability $p$ associated with the result. If possible, it is also desirable to include the data, such as means or frequencies, upon which the test was carried out.

**Drawing Conclusions**

Populations of individuals being investigated can be medically, geographically, or demographically defined. In most studies inferences are made about a population from results obtained from a sample of individuals. The validity of this procedure depends on having a representative sample. Methods of obtaining these are well-documented (Armitage, 1971; Hill, 1971). If a sample is not representative of the population the validity of extrapolating results must be a subjective decision based on non-statistical criteria.

The significance of a result is a measure of how sure we are that the effect measured is a real one and not a chance occurrence. It does not help us decide whether such an effect is important; such a conclusion must be taken in the light of existing knowledge of the subject. No amount of statistics, however significant, will indicate the importance of a finding.

Too much reliance can be placed on non-significant results. These could be the result of a trial being too small and not necessarily because the null hypothesis is true. However, a significant result obtained from a small sample is no less reliable because of the sample size, as this is allowed for in the calculation.

The interpretation of associations can also present problems. What can be deduced from associations such as: 'More heroin addicts than non-addicts smoke cigarettes', or 'Bournemouth has a higher mortality rate than London'? Few would conclude that taking heroin leads people to smoke cigarettes (or vice versa) or that Bournemouth is a dangerous place to live, without consideration of other relevant information.

The effect of all such 'nuisance' variables has to be allowed for, or considered in some way, before deducing that a causal relationship may exist. Analysis of covariance (in which allowance is made for nuisance variables) or analysis within subgroups (for example, for each sex) can get round these difficulties. However, if important data are not available, it will be much harder to draw firm conclusions. Even if all known intervening variables have been allowed for, there can be no certainty that the association implies causation. Extreme caution is needed in the interpretation of association, and causation must be verified by more detailed investigation, preferably by experiment.

The drawing of conclusions without testing the data is a common fault in medical papers; for example, it is not acceptable to conclude from a drug trial on two groups of ten patients that a success rate of 90 per cent in one group and 60 per cent in the other means that one drug is superior to the other, as the difference is not significant. Even if the two drugs are equally effective such a result would be very likely because the sample is so small. Schor and Karten
showed that drawing conclusions about populations without reference to any statistical test was the most common published error, occurring in 25 per cent of the papers studied.

In attempting to consider some more general issues concerning the interpretation and presentation of results, and by concentrating on the problems that may be encountered, it is possible that I have given the impression that statistical principles are difficult to grasp or dangerous to use. This would be unfortunate. Used with care, statistical techniques provide a very powerful means of advancing the cause of rational and constructive medicine. Statistical methods provide rules for asking the right questions and producing reliable answers.

Acknowledgements
I would like to thank M. F. D'Souza, C. du V. Florey and A. V. Swan for their helpful criticism during the preparation of this article.

References
Armitage, P. (1971) Statistical Methods in Medical Research. Oxford: Blackwell.
Feinstein, A. R. (1974) Clinical Pharmacology and Therapeutics, 15, 97.
Finney, D. J. (1955) Experimental Design and its Statistical Basis. University of Chicago.
Gifford, R. H. and Feinstein, A. R. (1969) New England Journal of Medicine, 280, 351.
Hill, A. B. (1971) Principles of Medical Statistics, 9th Edition, London: Lancet.
Huff, D. (1973) How to Lie with Statistics. Harmondsworth: Penguin.
Jellinek, E. M. (1946) Biometrics, 2, 87.
Schor, S. and Karten, I. (1966) Journal of the American Medical Association, 195, 1123.
'Student' (1931) Biometrika, 23, 398.

Requests for reprints of this article should be sent to the author at the Clinical Research Centre, Watford Road, Harrow, Middlesex HA1 3UJ.