Medicine and Mathematics

Statistics and ethics in medical research

V—Analysing data

DOUGLAS G ALTMAN

The incorrect analysis of data is probably the best known misuse of statistical methods, largely due to a series of reviews[1-3] that have shown how common such errors are in published papers. Nevertheless, these mistakes, which tend to be in the use of the simpler techniques, continue to proliferate. The mishandling of statistical analysis is as bad as the misuse of any laboratory technique. Both can lead to incorrect answers and conclusions and are thus unethical because they render research valueless.

In this article I will look briefly at problems associated with simple significance tests and will consider in more depth some less well-appreciated difficulties associated with two other common techniques—correlation and regression. I will then look at two specific medical problems that often result in incorrect analyses.

Errors in common statistical analyses

Nowadays some types of statistical analyses are seen so often in medical publications that their use is taken for granted. Everyone knows them, but the evidence suggests that many people do not know how to use them properly, or when *not* to use them. For example, Gore *et al*[2] found at least one such error in about half of the papers containing statistical analyses that they reviewed.

t TESTS AND χ^2 TESTS

The *t* tests to compare two groups of measurements are used extremely widely, but often incorrectly.[2-4] The problems usually relate to the data not complying with the underlying statistical assumption that the two sets of data come from populations that are Normal and have the same variance. Another serious error is to ignore the fact that the two sets of measurements relate to the same (or matched) individuals, in which case the paired *t* test is needed. These problems are fairly familiar and have been well illustrated by White[3] so I will not consider them further here.

Although generally posing fewer problems, χ^2 tests for comparing proportions also suffer some abuse, notably where there are too few observations. The sample size constraint also applies to the form of χ^2 test which simply entails comparing observed and expected frequencies. This method was used to compare observed numbers of deaths from five types of leukaemia (0, 1, 2, 4, 0) with their respective "expected" numbers (2, 1, 1, 3, 0),[5] but seven deaths is far too few for such an analysis to be valid.

CORRELATION

Perhaps one harmful side effect of the vast increase in availability of computing power is that the distinct statistical analyses of correlation and regression have become greatly confused. This is probably because of the close similarity between the mathematical calculations rather than for any logical reason, for it is relatively rare that one is truly interested in both analyses.

The correlation coefficient is a measure of the degree of linear (or "straight line") association between two continuous variables. If the relationship between the two variables is curved the correlation may be an artificially low measure of association. Alternatively, the correlation may be artificially high if a few observations are very different from the rest. For these reasons it is unwise to place any importance on the magnitude of the correlation without looking at a scatter plot of the data.

Misleading correlations can also be obtained if the data relate to different groups of subjects having different characteristics. Adam[6] looked at the relationship between body weight and the proportion of sleep that was rapid eye movement sleep in 16 adults, and found a rank correlation of 0·78. The original high correlation, however, was partly due to the men having higher values of both variables, for the correlations for men and women separately were 0·61 and 0·37 respectively. A further incorrect procedure is to use data comprising more than one observation per individual.

The main problem is that the test of significance of a correlation coefficient, which is a test of the null hypothesis of no association (zero correlation), is based on the assumption of joint Normality of the two variables. This is characterised by the data points having a roughly elliptical shape in the scatter diagram. If this is not so the correlation will be misleading and the test of significance invalid. The distributional assumption may be overcome either by transformation of the data, or by the calculation of "rank" correlation, which makes no important assumptions.

In medical research correlations are greatly overused, perhaps because they are easy to calculate and are measured on a scale that is independent of the data. Correlation ought really to be considered to be mainly an investigative analysis, suggesting

Division of Computing and Statistics, Clinical Research Centre, Harrow, Middx HA1 3UJ
DOUGLAS G ALTMAN, BSC, medical statistician (member of scientific staff)

areas for further research; for forming hypotheses rather than for testing them.

REGRESSION

The rationale for regression analysis is very different. In regression we are interested in describing mathematically the dependence of one variable on one or more other variables. In the simple linear case we are calculating the equation of the "best" straight line relating to the so-called "dependent" variable (Y) to the "independent" (or explanatory) variable (X).* For example, we might be interested in the dependence of lung function on height or of blood pressure on age. The appropriateness of a linear relationship can again best be verified by means of a scatter plot.

The most important underlying assumption in regression is that the Y variable is Normally distributed with the same variance for each value of X, and major departures from this condition can usually be detected by eye. There are no restrictions on X, so that it is perfectly valid, for example, to choose a wide range of X values to get a better estimate of the regression line. This would, however, artificially inflate the correlation coefficient, although correlations are often calculated from such data.

Regression is used to estimate a dependence relationship. The resulting equation can be used to predict Y (say, lung function) from X (height) for an individual. The difference between an individual's actual and predicted lung functions can be used as a measure of lung function standardised for height.

Examples of improper practices are the use of the regression equation to predict the Y variable for values of the X variable outside the range of the original data set (called extrapolation); the fitting of a straight line where the data show curvature; the use of a Y on X regression equation to predict X from Y (except in certain circumstances); and the use of simple regression where there are heterogeneous subgroups (the correct technique being analysis of covariance). Unless there is a plot of the data most of these procedures may be undetectable in a published paper.

Method comparison studies

Some of the practical problems in analysing data, notably the choice of the correct analysis to match the relevant hypothesis, are well illustrated by the problems of method comparison studies.

In medical research it is quite common to carry out a study to compare two different methods of measuring something. This may be to compare measurements made with some new piece of equipment with the "true" measurements, but it is more often to compare two different measuring devices where neither can be said to give the truth. (A similar problem arises when comparing the same measurement on different occasions.)

The obvious first step in the analysis is to plot the values obtained by each method as a scatter diagram. To judge from publications, the apparently obvious second step is to calculate the correlation between the two measurements. This is, however, a completely misguided approach, stemming from the common failure to appreciate what information the correlation coefficient gives.

An example of the false reasoning that is very common in published work is given by a study[7] comparing two methods of assessing the gestational age of newborn babies; one was the much-used Dubowitz method based on neurological and physiological signs and the other the Robinson method, which is based on neurological signs only. The scatter diagram showed only moderate agreement. The correlation between the two methods, however, was 0·85, and the authors argued directly from this that the two methods agreed well and that it would be reasonable to use the simpler method.

To test an observed correlation coefficient for statistical significance is to test how likely the observed result would be under the "null hypothesis" that the two variables were not associated at all. This is patently ludicrous when the two variables are obviously associated by their very nature; we would be astonished to find that two methods of measurement were uncorrelated. In fact, it can be shown that in these circumstances the magnitude of the correlation largely reflects the spread of the measurements. As such, its use is completely erroneous in this context.

What we really want to know in these studies is how well the two measures agree. The simplest approach is to calculate the difference between two measurements for each subject. The mean of these differences will then be a measure of accuracy (or bias) and the standard deviation a measure of precision. Both bias and precision are necessary in order to assess agreement. The between-method differences may tend to increase as the measurements increase, in which case it may be necessary to transform the data before analysis. With more than two methods, or if repeat observations are made (which is desirable), the more general analysis of variance must be used.

Hunyor *et al*[8] did calculate the mean and standard deviation of paired differences when comparing various sphygmomanometric methods with intra-arterial blood pressures, but then based their statements about relative accuracy on the high correlations they found. They studied hypertensives only; had they studied some normotensives as well they would undoubtedly have observed higher correlations, but these would not have implied any better agreement between methods.

One last point about method comparison studies is that they are often carried out on such small numbers of subjects that the two methods will not be found significantly different unless there is an enormous difference between them. There is considerable potential here for incorrectly finding a new method acceptable, and for such methods to be recommended for widespread use without justification.

Reference ranges

Another area where simple statistical methods are often applied blindly is in the construction of reference (or normal) ranges against which to judge future observations. For example, some people believe that since a range is required, all that is needed is to obtain results from some "normal" subjects and quote the range of values. Apparent differences in reference ranges for the same index can often be attributed to one or more of them having been calculated incorrectly. Also, the sample size taken is often too small to get reliable answers. I have seen a reference range calculated from seven subjects, incorrectly at that, whereas at least 100 observations are needed to get a reliable range.

The usual calculation of a 95% reference range as the mean ± 2 standard deviations is yet again based on the assumption that the data follow a Gaussian or Normal distribution. Often this condition is not fulfilled and we see statements like "The mean ^{99m}Tc uptake in this group was 1·8% $\pm$ SD 1·1%, making the upper limit of normal (mean ± 2 SD) 4·0%."[9] The unstated lower limit is negative, however, which is nonsense. This type of calculation of a normal range on skew data results in considerably more than the nominal 5% of subjects being classified as "abnormal." The consequence of such a classification may be to perform further tests, so that there is a clear ethical aspect to the construction and interpretation of normal ranges. Even where the range is calculated sensibly there is a strong case for quoting the standard error of the limits, to emphasise the considerable uncertainty involved.

Whether or not the use of such ranges is sensible is beyond the scope of this article; the issues have been clearly discussed by Oldham[10] and Healy.[11]

*These terms simply denote which variable is considered to be dependent on the other.

Selecting which data to analyse

A rather more subtle problem that can occur in any study is the selection of which data to analyse. Errors may occur when analyses are carried out as a direct result of having seen the data. In a comparison of several groups of subjects it is not valid to select those groups with the highest and lowest values and apply the usual significance test to the means purely on that basis, because the null hypothesis of no difference is inappropriate when the largest difference is being examined. More generally, selection of comparisons to test because they "look interesting" will in the long run result in more than the nominal (say 5%) proportion of falsely positive results.

A second form of selection is to analyse only a subset of the subjects on the basis of their results. In a recent study 30 patients with idiopathic hypercalciuria were given a dietary supplement of unprocessed bran.[12] Only 22 patients "achieved a reduction in urinary calcium," and only these 22 patients were analysed. No data were provided on the other eight subjects, so we can not tell whether they really were a different group or just one end of a distribution of differing responses to the bran, which seems more likely. This procedure is completely unacceptable without justification—anyone can show significant results by analysing only those subjects with the greatest response.

The basic principle is to analyse according to the original hypothesis and experimental design. Other results that look interesting are pointers for further research.

Summary

It is of no value collecting good data if the analysis is inadequate or invalid. The results obtained may then be worthless, or at best they will fail to realise the true potential of the data. Either way, the value of the whole experiment is diminished to a point where the ethics of the investigation must be called into question.

References

1 Schor S, Karten I. Statistical evaluation of medical journal manuscripts. *JAMA* 1966;**195**:1123-8.
2 Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in *BMJ* from January to March 1976. *Br Med J* 1977;i:85-7.
3 White SJ. Statistical errors in papers in the *British Journal of Psychiatry*. *Br J Psychiatry* 1979;**135**:336-42.
4 Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 1980;**61**:1-7.
5 Tabershaw IR, Lamm SH. Benzene and leukaemia. *Lancet* 1977;ii:867-8.
6 Adam K. Bodyweight correlates with REM sleep. *Br Med J* 1977;i:813-4.
7 Serfontein GL, Jaroszewicz AM. Estimation of gestational age at birth. *Arch Dis Child* 1978;**53**:509-11.
8 Hunyor SN, Flynn JM, Cochineas C. Comparison of performance of various sphygmomanometers with intra-arterial blood pressure readings. *Br Med J* 1978;ii:159-62.
9 Van 'T Hoff W, Pover GG, Eiser NM. Technetium-99^{m} in the diagnosis of thyrotoxicosis. *Br Med J* 1972;iv:203-6.
10 Oldham PD. The uselessness of normal values. In: Arcangeli P, Cotes JE, Cournand A, eds. *Introduction to the definition of normal values for respiratory function in man*. Turin: Panminerva Medica, 1969:49-56.
11 Healy MJR. Normal values from a statistical viewpoint. *Bullétin de l'Académie Royale de Médecine de Belgique* 1969;**9**:703-18.
12 Shah PJR, Green NA, Williams G. Unprocessed bran and its effect on urinary calcium excretion in idiopathic hypercalciuria. *Br Med J* 1980;**281**:426.

This is the fifth in a series of eight articles. No reprints will be available from the author.

The Drug Industry

Drug famine: possible solutions

TONY SMITH

Inside and outside the pharmaceutical industry informed observers are concerned with current trends. Increasingly the major innovative companies seem to be concentrating their efforts on the relatively few drugs that have large international markets—tranquillisers, cardiovascular drugs, antidepressants—rather than broadening the range of their research objectives. The explanation is straightforward enough. A company has to regain its development costs by obtaining profits on sales during the period of patent protection (after which competitors can produce generic equivalents sold simply at the cost of manufacture). Recent trends have all been in one direction: development costs have risen, prices have been cut by government diktat, and the effective patent life has been reduced.

British Medical Journal, London WC1H 9JR
TONY SMITH, BM, BCH, deputy editor

In consequence only those drugs with vast sales markets have the potential to make the profits required to sustain research. The factors responsible for these changes have been examined in my earlier articles—consumer campaigns for drug safety; improvements and refinements made by the pharmaceutical industry in its methods for development and testing of new compounds; and the proliferation of bureaucracy within the drug regulatory agencies.

Fortunately, there are signs that the public and politicians are beginning to recognise that the pendulum has swung too far in the direction of safety first. "The climate of opinion is changing," said Dr Dunne (Senior Medical Officer Pharmaceuticals, World Health Organisation, Geneva). "This has been exposed to some extent by the development of free trade areas and at the European Economic Community, when the civil servants responsible for drug regulation in individual countries came together to argue out a policy. Those from the United Kingdom soon realised that their rules were much more stringent than anywhere else in the EEC—and yet people were not dying like flies on the other side of the channel."