Medicine and Mathematics

Statistics and ethics in medical research

VII—Interpreting results

DOUGLAS G ALTMAN

". . . it is a function of statistical method to emphasise that precise conclusions cannot be drawn from inadequate data."
E S PEARSON AND H O HARTLEY[1]

The problems of interpretation have already appeared several times in the preceding articles. Obviously the sorts of error already discussed, most likely in design or analysis, may lead to incorrect results and thus erroneous conclusions. But some errors are specific to the interpretation of results, and these I will consider in this article. Most emphasis will be given to tests of significance, since these quite clearly cause great difficulty.

Significance tests

Before tackling some of the trickier issues it is worth making the general point that the sensible interpretation of statistical analysis cannot be independent of the knowledge of what the data are (and how they were obtained).

Table I, for example, shows the results of the comparison of two groups of subjects given different treatments with the outcome for each subject recorded as positive or negative. A

TABLE I—*Comparison of outcomes for two treatment groups*

		Outcome A	Outcome B	Total
Treatment	1	4	4	8
	2	8	24	32
Total		12	28	40

χ^2 test $p<0.05$.

χ^2 test on these data shows a significant association between the grouping and the outcome, but in the absence of further information we are unable to interpret these results. Knowing that the subjects were all pregnant and the outcomes were male and female babies is likely to aid interpretation and increase interest, but the further knowledge that the subjects were all cows will probably lessen interest again, unless you are a farmer. Yet you may be curious to know what the "treatments" were—perhaps there is some relevance for people. Well, all the cows were artificially inseminated; those in group 1 were facing north at the time and those in group 2 were facing south.[2]

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

Given all the information, most people would probably dismiss this as a chance finding, rather than accept it as evidence of an association between the direction the cows were facing and the sex of their calves. This is quite reasonable behaviour if we consider the meaning of statistical significance.

INTERPRETING SIGNIFICANT RESULTS

Like several statistical terms, "significant" is perhaps an ill-chosen one. It should be realised that the level of significance is just an indication of the degree of plausibility of the "null hypothesis," which in the above example was that the outcomes of the two groups were really the same. If the null hypothesis is deemed too implausible we reject it and accept the "alternative hypothesis" that the treatments differ in their effect.

It is ridiculous to lay down rigid rules for something so subjective, especially as interpretation will be greatly influenced by other evidence—few studies are carried out in isolation. As Box *et al*[3] have said: "If the alternative hypothesis were plausible a priori, the experimenter would feel much more confident of a result significant at the 0·05 level than if it seemed to contradict all previous experience." Indeed, in the long run one in 20 comparisons of equally effective treatments will be significant at the 5% level (by definition), so to accept all significant results as real[4] is extremely unwise, as the above data illustrate.

Conventional significance levels (5%, 1%, 0·1%) are useful, but only as guides to interpretation, not as strict rules. To describe a result of $p=0.05$ as "probably significant"[5] implies that the interpretation depends on which side of 0·05 p really is. On the contrary, values of p of, say, 0·06 and 0·04 should not lead to opposite conclusions, but to closely similar ones.

One prevalent misconception relates to the precise meaning of p, the significance level; p is the probability of obtaining a result at least as unlikely as the observed one, if the null hypothesis of no effect is true. The last part of this definition is essential; to omit it leads to the common error of believing that p is also the probability so that we make a mistake by accepting the significant result as a real finding. This is just not so, and it is sad to see this view in a paper trying to explain the meaning of significance.[6] All we can say is that p is the probability of such a result arising if the null hypothesis is true. We obviously do not actually know whether the null hypothesis *is* true, so the probability of rejecting it in error is also unknown, although this clearly reduces as p gets smaller.

INTERPRETING NON-SIGNIFICANT RESULTS

Every significance test measures the credibility of a null hypothesis—for example, that two treatments are equally

effective. A non-significant result just means that the results were not strong enough to reject the null hypothesis; "not significant" does not imply either "not important" or "non-existent." To consider all non-significant results as indicating no effect of importance is clearly wrong. Conversely, to believe that an observed difference is a real one with an insufficient degree of certainty is to run a large risk of chasing shadows. Thus when reporting "negative" results, it is especially important to give a confidence interval around the observed effect[7 8]—for example, around the difference between two means.

In the third article I discussed at length the idea of the power of a significance test. It is appropriate to return to the topic of power here. Studies with low power (as a result of inadequate sample size) will often yield results showing effects which, if real, would be of clinical importance, but which are not statistically significant. In general it is safest to consider such non-significant results as being inconclusive (or "not proven"), preferably backed up with a recommendation that further data be collected. When this is not feasible and there are ethical implications, as in the following example, the problem of interpretation is particularly great.

Carpenter and Emery[9] investigated the possible effect on the incidence of sudden unexpected infant death of an increase in the number of visits by the health visitor to high-risk babies. They found fewer unexplained deaths in the "treatment" group (five out of 837) than in the control group (nine out of 922), but the difference is not nearly statistically significant ($p>0.5$). From a statistical point of view, the results are inconclusive. Because such deaths are rare, the power of the study was very low; it would have needed a much larger sample to get a clear answer.[10] The authors asked: "Can we reasonably withhold increased surveillance from all high-risk infants?"[11] More dispassionately we might ask whether the evidence is really strong enough to justify a change in policy that would presumably necessitate withdrawing health visitors from other activities.

MULTIPLE TESTS OF SIGNIFICANCE

A further difficulty arises when several tests of significance are carried out on one set of data. This may, for example, take the form of looking to see which pairs of a number of groups are significantly different from each other, or which of a number of different factors are related to a variable of interest. Unfortunately, the greater the number of tests carried out, the higher the overall risk of a "false-positive" result. As Meier has pointed out,[12] it is not reasonable to restrict the number of aspects of the data that are investigated purely to relieve the statistician's problems of interpretation. He suggested a good compromise, which is to treat a small number of tests as being of primary importance, "and to regard other findings as tentative, subject to confirmation in future experiments." The level of significance will have some bearing here, since we will be more ready to accept a highly significant finding (say, $p<0.001$) even in the context of numerous tests.

Association and causation

It is widely believed that "you can prove *anything* with statistics," but it is much more realistic to say that you can establish *nothing* by statistics alone. This is especially true when considering the interpretation of observed associations between two variables. It is easy and often tempting to assume that the underlying relationship is a causal one, even in the absence of any supporting evidence, but many associations are not causal. In particular, misleading associations appear when each of the variables is correlated with a third "hidden" variable. A simple example of this phenomenon is when two variables that change with time display an association in the complete absence of any causal relationship—for example, the divorce rate and the price of petrol.

The deduction of a causal relationship from an observed association can rarely be justified from the data alone. Support is needed from prior knowledge, including other experimental or observational data. Sometimes, however, such information is not obtainable, and the causal hypothesis can be supported only by allowing for the most likely hidden variables. There are several examples of epidemiological studies producing associations that are not unanimously believed to be causal, such as that between water hardness and cardiovascular mortality. A few people do not even accept that the association between smoking and lung cancer is causal despite the great volume of collateral evidence.

A recent paper[13] concerning the failure to show a relationship between diet and serum cholesterol concentration gave a salutary reminder that variables may falsely appear to be unrelated. Although a strong relationship between dietary cholesterol and serum cholesterol has been shown in closely controlled dietary studies, the authors showed that a straightforward population study would be likely to miss such an association because of several sources of variability in both variables.

Another difficulty that can beset the interpretation of observed associations is where two possibly causal factors are inseparable. A simple example is where two alternative methods of measurement are compared with only one experimenter using each method.[14] Any observed differences may be due either to differences between the methods or between the experimenters, or both. The two effects are *confounded*. A much more complex version of the same problem arises when trying to explain different mortality rates for the same disease in different countries.

Prediction

The use of observed relationships to make predictions about individuals is another area with many pitfalls. Just as it is dangerous to generalise from the particular, we must be very careful about particularising from the general.

For continuous variables, relationships are usually described by regression equations. It must be remembered that such fitted equations are approximate, both because they are calculated from a sample of data, and also because the imposition of an exact relationship (straight-line or curved) may be more convenient than realistic. The degree of scatter of the observations around the fitted line indicates the closeness of the relationship between the variables, and thus the uncertainty associated with predicting one from the other for specific cases. For example, a regression of height on weight for adult men would show a clear positive relationship with a large amount of scatter.

Regression equations should be used for prediction only within their limitations, so the regression line described above would be inappropriate for either boys or women. Such extrapolation is completely invalid. Also the prediction of height would be more certain for someone of average weight than for a very light or very heavy man, and this is borne out by the correct 95% confidence intervals for prediction which become wider further from the mean. It is very common to see a single figure quoted for the precision of any possible estimate; this is quite wrong.

Prediction also poses problems where the data are categorical. Table II shows the relationship between two diagnostic tests and the presence or absence of two diseases. Data such as these are usually described by the sensitivity and specificity, confusing terms for the proportions of correctly diagnosed positives and negatives. In both cases the sensitivity and specificity are high at 0·9 (maximum 1·0). These do not, however, measure the value of such tests for predictive purposes; in fact they become more misleading the lower the prevalence of the disease. The best approach is to consider what proportions of the diagnosed

positives and negatives were true positives and negatives respectively. In table IIa, where the prevalence is 50%, these figures are also both 0·9, indicating high predictive ability. In table IIb, the prevalence is 2%. Although virtually all of those with a negative test were truly negative (4410/4420), only 16% (90/580) of those diagnosed as positive were true positives. So the value of the test is low, even though 90% of the true positives give positive results. The usefulness of such a test depends on the cost of a false-positive finding. This is the problem when deciding whether or not screening for rare conditions (such as breast cancer) is worth while. For such purposes, the sensitivity is of no use at all—a high sensitivity is a necessary but not sufficient condition for a good predictive test.

Exactly the same considerations apply to the interpretation of a value exceeding a reference (or normal) range as automatically indicating abnormality without consideration of the prevalence of abnormality. Indeed, this is equivalent to looking at only the top row in table IIb. Such a procedure can lead to ludicrous interpretations of data—for example, that it is safer to drive very fast as few accidents are caused by cars travelling at more than 100 miles an hour.

TABLE II—*Relation between diagnostic test and disease state with prevalence of disease (a) 50% and (b) 2%*

(a)

		Test +	Test −	
Disease	+	180	20	200
	−	20	180	200
		200	200	400

(b)

		Test +	Test −	
Disease	+	90	10	100
	−	490	4410	4900
		580	4420	5000

Conclusions

The enormous amount of published research makes it inevitable that papers will often be judged, in the first instance at least, by the authors' own conclusions or summary. It is thus vitally important that these contain valid interpretations of the results of the study, since the publication of misleading conclusions may both nullify the research in question and falsely influence medical practice and further research.

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

References

1 Pearson ES, Hartley HO. *Biometrika tables for statisticians.* Vol 1. 3rd ed. Cambridge: University Press, 1970: 83.
2 Wood PDP. On the importance of correct orientation to sex in cattle. *Statistician* 1977;**26**:304-6.
3 Box GEP, Hunter WG, Hunter JS. *Statistics for experimenters.* New York: Wiley, 1978:109.
4 Dudley H. When is significant not significant? *Br Med J* 1977;ii:47.
5 Newton J, Illingworth R, Elias J, McEwan J. Continuous intrauterine copper contraception for three years: comparison of replacement at two years with continuation of use. *Br Med J* 1977;i:197-9.
6 Glantz SA. Biostatistics: how to detect, correct and prevent errors in the medical literature. *Circulation* 1980;**61**:1-7.
7 Rose G. Beta-blockers in immediate treatment of myocardial infarction. *Br Med J* 1980;**280**:1088.
8 Chalmers TC, Matta RJ, Smith H, Kunzler A-M. Evidence favoring the use of anticoagulants in the hospital phase of acute myocardial infarction. *N Engl J Med* 1977;**297**:1091-6.
9 Carpenter RG, Emery JL. Final results of study of infants at risk of sudden death. *Nature* 1977;**268**:724-5.
10 Bland JM. Assessment of risk of sudden death in infants. *Nature* 1978; **273**:74.
11 Carpenter RG, Emery JL. Reply to J M Bland. *Nature* 1978;**273**:74-5.
12 Meier P. Statistics and medical experimentation. *Biometrics* 1975;**31**: 511-29.
13 Jacobs DR, Anderson JT, Blackburn H. Diet and serum cholesterol: do zero correlations negate the relationship? *Am J Epidemiol* 1979;**110**: 77-87.
14 Serfontein GL, Jaroszewicz AM. Estimation of gestational age at birth. *Arch Dis Child* 1978;**53**:509-11.

Today's Treatment

Clinical pharmacology

The Committee on Review of Medicines

T B BINNS

In retrospect it is astonishing that until 1964 there was virtually no control over the marketing of drugs in Britain. In that year the voluntary Committee on Safety of Drugs (Dunlop Committee) came into operation, and when the Medicines Act of 1968 was finally implemented in September 1971 this was succeeded by the Committee on Safety of Medicines (CSM). The CSM advises the licensing authority, which is part of the medicines division of the DHSS, on the granting of licences inter alia for clinical trials and new products. Products already on the market in September 1971 were given product licences of right. The data sheets on these products were prepared by the manufacturers, and very few of them had undergone independent evaluation.

Late in 1975 the Committee on Review of Medicines (CRM) was created to undertake this task and also to comply with an EEC directive that imposed a deadline for review of May 1990.[1] In status it corresponds to the CSM, being one of several committees set up under section 4 of the Medicines Act, and it operates in much the same way. Its members are appointed by the health ministers with the advice of the Medicines Commission. They are broadly representative of the interests concerned but are appointed as individuals and not as delegates. They are paid a nominal honorarium that is not remotely commensurate with the time taken up, so their decisions are in no way influenced by considerations of finance or job security. The first chairman was Sir Eric Scowen, and he was succeeded by Professor Owen Wade in 1979. Technically these committees report directly to the

Department of Pharmacology and Therapeutics, London Hospital Medical College, London E1 2AD

T B BINNS, FRCP, honorary senior lecturer; member of the CRM