MEDICAL PRACTICE

Occasional Review

Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976

SHEILA M GORE, IAN G JONES, EILIF C RYTTER

British Medical Journal, 1977, **1**, 85-87

Summary

Sixty-two reports that appeared as Papers and Originals (excluding short reports) in 13 consecutive issues of the British Medical journal included statistical analysis. Thirty-two had statistical errors of one kind or another; in 18 fairly serious faults were discovered. The summaries of five reports made some claim that was unsupportable on re-examination of the data. Medical investigators should consult with people who have a real understanding of statistical methods throughout their projects.

Introduction

The quality of statistical work in medical journals has improved dramatically in recent years. Even so, the Royal Colleges of Physicians expressed concern in 1973[1] that doctors were unfamiliar with medical statistics. The introduction into the MRCP examinations of questions on statistics initiated reform. The *British Medical Journal* has recently published a series of articles[2] on elementary statistics and includes in its instructions to authors[3] the requirement that "Any statistical procedure should be detailed in the methods section of the paper, and any not in common use should be either described in detail or supported by references."

As a tutorial exercise, diploma students in community medicine at the University of Edinburgh criticised the exposition and statistical analysis of reports in an issue of the *British Medical Journal* from April 1976. Several errors were discovered, which prompted a more detailed investigation.

Medical Computing and Statistics Group, University of Edinburgh
SHEILA M GORE, MA, (present address: Department of Statistics, University of Aberdeen)

Area Health Board, Dunfermline, Fife
IAN G JONES, MRCP, DCM

Health Directorate, Oslo, Norway
EILIF C RYTTER, CAND MED, DCM,

Methods

The Papers and Originals (excluding descriptive papers and short reports) in the 13 issues of the *British Medical Journal* published in January to March 1976 were independently read and annotated by a trainee community medicine specialist IGJ (7 issues) or ECR (6 issues)) and a medical statistician (SMG (13 issues)). Five types of error, which would hinder the reader's understanding, were defined.

"Errors of commission," which, broadly, are abuses of statistics, were distinguished (asterisked in table II) from "errors of omission," whereby incomplete information of inadequate explanation was given by researchers. Errors of omission are less serious. We arrived at a consensus for the presence or absence in an analytical paper of each category of error by reference to the two assessors' reports and the text. Presence of an error category was defined as the occurrence of at least one error from that category.

The appendix gives a more detailed description of the error categories, defined briefly as follows:

(I) *Inadequate description of basic data*—Measures of location—median, mode, mean—indicate the "centre" of sample values, but unless a measure of spread or dispersion—for example, range, standard deviation—is also quoted the reader is unable to visualise the data.

(II) *Disregard for statistical independence*—A serious error results when an investigator analyses multiple observations on one patient as though they represented single observations from distinct patients. Repeat observations vary less than independent assessments. It is then impossible to draw inferences from an analysis relying on statistically independent observations.

(III) *Errors related to randomisation*—Failure to randomise patients to treatment is seldom justified, and a clear explanation of the reasons that underlie non-random assignment is necessary to persuade the reader that biased allocation has not occurred. The researcher who states that "patients were randomised to treatment" without explaining even briefly the chosen method and strata for randomisation or how the scheme operates fails to inform adequately.

(IV) *Errors with Student's* t *test*—Researchers who want to compare the responses in two samples too readily use Student's t test without verifying that the assumptions underlying the test are even approximately satisfied. Three important assumptions are that: (*a*) the distribution of sample means is normal; (*b*) sample observations are statistically independent; (*c*) variances are equal. Transformation of observed responses that are considerably skewed often results in the sample mean of the transformed variable following more nearly a normal distribution. If the variance of observations is very different in the two samples (known as heteroscedascity) not only is a pooled estimate of variance inappropriate but a reduction in the degrees of freedom associated with Student's t test is also required. Failure to do this may lead to falsely significant results. It is erroneous to use Student's 2-sample t test on paired data.

(V) *Errors with* χ^2 *tests*—Many hypotheses are assessed with a χ^2 test statistic. The hypothesis under test should, therefore, be clearly explained, and the degrees of freedom associated with the test statistic noted. The χ^2 test is an approximate one. In 2×2 tables particularly, it is safer to use the so-called "continuity correction" to ensure that the approximation is more nearly correct. If the numbers in the table are small Fisher's exact test is required. Information is wasted when a χ^2 test is applied to data from samples which are matched, as the matching is ignored.

Results

Fifteen of the 77 Papers and Originals studied included no statistical analysis. Of the remaining 62 analytical reports, 32 (52%) included at least one error from one or more of the five categories.

TABLE I—*Numbers of categories of error found in 62 papers*

No of categories, n, from which errors occurred	Frequency of papers with n category errors	Expected frequency—based on Poisson distribution, rate r = 0·76 error categories per paper
0	30	29·1
1	19	22·0
2	11	8·3
3	2	2·1
4		0·4
5		0·1
Total	62	62·0

Goodness of fit test: $\chi^2_1 = 0{\cdot}84$.

TABLE II—*Number of analytical papers containing errors from categories I-V. Within categories III-V frequency of papers with each subdivision of error is shown*

Category	Description	No of papers with suberror	No of papers containing at least one error from category
I	Inadequate description of basic data		10
II	Disregard for statistical independence*		6
III	Errors related to randomisation:		8
	(*a*) incomplete explanation of procedure	6	
	(*b*) failure to randomise	2	
IV	Errors with Student's t test:		11
	(*a*) underlying discontinuity or non-normality*	6	
	(*b*) heteroscedascity*	6	
	(*c*) 2-sample test applied to paired data	0	
V	Errors with χ^2_1 tests:		12
	(*a*) continuity correction omitted when critical*	3	
	(*b*) no indication of hypothesis under test, or degress of freedom	8	
	(*c*) loss of information on matched data	2	
			47

*Errors of commission.

Table I shows the number of categories from which errors were discovered in the 62 papers. The mean number of error categories per paper was 0·76 (47 errors in 62 papers). There was good agreement between the observed distribution of error categories in the 62 papers and the distribution that would be expected if error categories occurred at random in papers at the constant rate of 0·76 categories per paper. If the occurrence rate were quartered only one paper in 62 would include errors from two or more categories.

Table II shows the number of papers which included errors from the individual categories I-V. Papers may include errors from more than one subdivision within a category. The subdivisions of error within categories III-V are listed, together with the number of papers that showed the suberror. Errors of commission are denoted by an asterisk. Ten papers made errors of commission only, 14 made simply errors of omission. A further eight papers included both types of error.

The seriousness of errors of commission is put into perspective by examining the validity of investigators' conclusions, as stated in the summary. On reanalysis we found that five of the 62 analytical reports (8%) made some claim in the summary that was not supported by the data presented.

Discussion

We criticised reports within the framework selected by the investigators; we made no criticism of design. We restricted our attention to errors from five categories. A further limitation resulted from the ease with which some researchers disguised the data they analysed, so that our suspicion of error could not be proved and was discounted. Certainly, not all suspicions were unfounded (see table II, category IV*c*). The true error rate was thus underestimated.

We have not identified the papers with errors. We do not want to pillory the investigators concerned, rather to draw attention to shortcomings which, plausibly, occurred at random in the 62 reports. It is clearly not sufficient to digest the summary of a paper; critical perusal may discover statistical defects. The results presented should not be generalised. The *British Medical Journal* is highly rated; less respected journals probably show a greater frequency of errors.

Schor and Karten[4] found that only 28% of a random sample of 149 analytical articles in American medical periodicals had sufficient statistical support for drawn conclusions. Comparison between their results and ours, however, is inappropriate both because medical statistics has advanced since 1966 and because the 12 errors listed by Schor and Karten included criticism of study design.

Although errors of commission would be avoided by introducing more extensive biostatistical refereeing, undue delay before publication may result. One practical solution is that investigators should consult medical statisticians. It is, however, some comfort that only five papers drew a false conclusion.

Appendix

A useful reference is Armitage.[5]

(I) *Inadequate description of basic data*

The semi-interquartile range is a more informative measure of spread than the range of observations. A crude indication of skewness is given by the difference between the mean and median. Scattergrams are fully descriptive.

(II) *Disregard for statistical independence*

The following example illustrates this disregard.

The spleens from a random sample of 10 patients undergoing splenectomy to stage Hodgkin's disease are studied to estimate the concentration of B cells. Four sections are cut from each spleen and the B-cell concentration is measured; the four measurements are *independent* observations of the B-cell concentration in a *given spleen*. The pathologist derives a "best" estimate for the concentration of B cells in that spleen from the four independent measurements. The 10 best estimates are *independent* assessments for splenic concentration of B cells in the patients. There is no value of which the 40 measurements are independent estimates.

(III) *Errors related to randomisation*

Skilful randomisation improves the precision and balance of experiments. Simple randomisation is only an elementary method.

(IV) *Errors with Student's* t *test*

The variance of the distribution of sample means is estimated from the data, and the precision of the estimate is reflected by the degrees of freedom associated with Student's *t* distribution.

When means are based on small samples the central limit theorem may not guarantee that the distribution of sample means is normal especially if observations are skewed. An example follows.

The length of stay in hospital is recorded to the nearest half day. Fig 1 shows the recorded length of stay for 20 patients in hospital A. The distribution is positively skewed. In fig 2 the distribution of the transformed variable log (length of stay), which follows more nearly a normal distribution, is shown. Assumption *a* (category IV in text) is approximately satisfied by the transformed response. Fig 3 shows log (length of stay) for 20 patients in hospital B. Within each hospital length of stay was recorded for 20 distinct patients and so assumption *b* is essentially satisfied. It is evident from figs 2 and 3 that the variance or spread of the observations on log (length of stay) is very different in the samples from hospitals A and B. Correction for heteroscedascity should not be ignored (assumption *c*).

The term "paired data" is explained in a further example.

Suppose that interest lies in comparing intraocular pressure before and after treatment in a random sample of patients with glaucoma. For each patient, the difference in pressure before and after treatment is calculated from the *pair* of observations on that patient. Student's 1-sample *t* test makes the relevant comparison between the *sample mean difference* and the zero difference that is expected if treatment has no effect on intraocular pressure.

(V) *Errors with* χ^2 *tests*

In a clinical trial 32 patients are assigned by simple randomisation to drugs A and B. Treatment success, defined as "discharge from hospital within three days of admission" is recorded for seven of the 15 patients randomised to drug A and for 14 of the 17 patients given drug B. The experimenter hypothesises that success rate is identical for both drugs and assesses by a χ^2 test whether the observed 14 successes on drug B out of 21 successes are consistent with that assumption. A correction for continuity is critical. The hypothesis that success rate is identical on both drugs is rejected at the 5% level, based on the uncorrected statistic ($\chi^2_1=4{\cdot}5$), but only at the 10% level when the corrected statistic ($\chi^2_1=3{\cdot}1$) is evaluated.

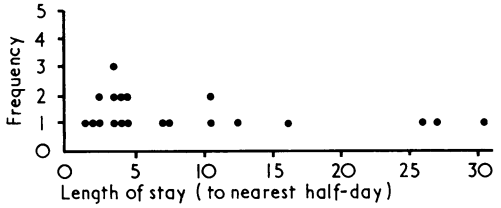

FIG 1—Hospital A. Skewed distribution of lengths of stay.

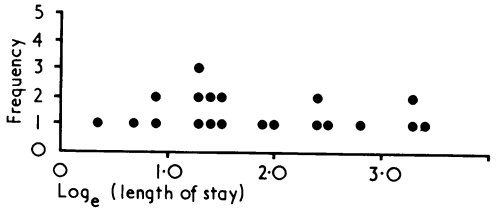

FIG 2—Hospital A. Distribution of transformed variable log$_e$ (length of stay).

FIG 3—Hospital B. Distribution of log$_e$ (length of stay).

References

1 Stokes, J F, *British Medical Journal*, 1973, **4**, 421.
2 Swinscow, T D V, *British Medical Journal*, 1976, **1**, 1262, 1325, 1393, 1458, 1513, 1585; **2**, 33, 94.
3 *British Medical Journal*, 1976, **1**, 6.
4 Schor, S, and Karten, I, *Journal of the American Medical Association*, 1966, **195**, 1123.
5 Armitage, P, *Statistical Methods in Medical Research*. Oxford, Blackwell Scientific, 1974.

(*Accepted 5 November 1976*)

Bone and Joint Diseases

Arthroscopy in practice

ADRIAN N HENRY

British Medical Journal, 1977, **1**, 87-88

The endoscopic examination of joints has been surprisingly slow to develop. The knee joint is particularly suitable for this technique because of the numerous mechanical derangements to which it is subject and also because arthrotomy does not always give adequate exposure owing to the intricacies of the joint's anatomy. Tagaki of Japan was the first to use an arthroscope to examine the knee, in 1920. It was not until 1931, however, that a satisfactory instrument was finally developed. Since then there have been several reports of the use of the arthroscope, notably by Finklestein and Meyer in America and Hurter in France. Japanese workers have, however, shown the greatest ingenuity in the use of the arthroscope, culminating in the instrument devised by Watanabe in 1959. Jackson of Toronto, having worked with Watanabe, introduced the technique to North America and did much to stimulate interest in this technique in the Western world.

In Europe over the past decade there has been considerable interest in the procedure, resulting in many new designs of

Guy's Hospital, London SE1 9RT
ADRIAN N HENRY, MCH, FRCS, senior consultant orthopaedic surgeon