Features Section

Guest Editorial

The Controversy Over How to Present Research Findings

This editorial from our statistical advisor has been stimulated by the letter from Malcolm Savage in this Journal.

Malcolm Savage draws attention to the resurgence of debate concerning the appropriate strategy for interpreting data from research studies. Although a working paper by Robert Matthews is cited, it is his shorter communication, published in the *Sunday Times* of 13 September 1998 and accessible at http://www.esef.org/matthews.html, which journal readers are likely to have seen. This short article uses emotive language to present a swingeing criticism of *hypothesis testing* as the primary tool for drawing inferences from research data. Aspersions are cast widely, on the integrity of clinical and other researchers, the statistical profession, including Sir Ronald Fisher (one of its founder fathers), and those involved in scientific publication and in the training of scientists. Matthews contrasts hypothesis testing with the inference system founded on the work of Reverend Thomas Bayes, which he portrays as both free of drawbacks, and also hard done by and widely rejected as subjective.

My purpose in responding to Malcolm Savage's letter is primarily to reassure journal readers that these contentions have not taken the journal editorial team unawares. Indeed, far from being a cover-up job, these issues have been debated at great length both by statisticians and those involved in publishing clinical research.

Problems with Statistical Significance

It must be conceded that, regrettably, for many decades research papers tended to present only summary statistics and expressions of statistical significance, and generally the emphasis fell on the latter as the basis for drawing inferences. Often effects were simply classed as 'significant' or 'not significant' at the 5 per cent level, which were widely interpreted to mean 'there is a real difference' and 'there is not'. Sometimes more extreme *P*-values were quoted as $P < 0{\cdot}01$ or $P < 0{\cdot}001$. It was difficult to avoid presenting *P*-values in this rounded form, when researchers had to rely on published tables. Now that statistical software is widely available, it is preferable that when *P*-values are quoted, about 2 significant figures are used. For example, $P = 0{\cdot}06$ and $P = 0{\cdot}85$ convey different information: the former indicates that the true effect size could be close to zero or could be twice the observed value, the latter suggests that the true effect size could be about 10 times the observed value, in either a positive or a negative direction.

The *P*-value is essentially a measure of surprise or coincidence: how surprising is the pattern shown by the data, pointing towards an effect in the direction shown, if in fact the null hypothesis is true? It does not indicate whether the effect is large and important, or small and unimportant. Occasionally, a large study yields strong evidence for the existence of a difference that is, however, too small to matter. Much more frequently, though, a study is carried out which is simply far too small and a potentially important difference fails to be detected. Among such small studies, it is those in which the play of chance conveniently exaggerates the effect size that tend to get submitted and often accepted for publication, the well-recognized phenomenon of publication bias. Furthermore, irrespective of study size, for every 20 hypothesis tests that are performed (or thought about doing), one 'significant' finding will emerge, just by chance—the multiple comparison problem. Positive 'effects' produced by this process will tend to attract attention and be hailed as 'breakthroughs'.

The Bayesian Paradigm

Alternative analytic approaches are available. The most radically different approach is '*Bayesian*' inference. In this paradigm, existing knowledge about the issue is first quantified, then updated to take account of the observed data, giving appropriate weightings to prior and new knowledge. it is scientifically accepted as the only valid approach when the issue is an inference about an *individual*, in situations such as computer-aided differential diagnosis and genetic counselling. It is also the only valid way to use evidence such as DNA testing in legal proceedings, although it has been unpalatable to the legal profession to concede this.

Where the controversy arises is in the use of Bayesian methods to draw inferences about population parameters, such as means or proportions or their differences. Here, the computational processes are generally complicated ones. Computer software is becoming more widely available, but not yet used to an appreciable degree by researchers other than statisticians. Hence, the presentation of results of Bayesian analyses will tend to look unfamiliar to many readers.

The need to incorporate prior knowledge is widely criticized as leading to *subjectivity*, but perhaps the real obstacle to the widespread use of Bayesian inference is the *complexity* introduced by having to put together all relevant prior knowledge. If the Bayesian principle is applied strictly, the word *all* should be taken quite literally here, including journals and books in totally unfamiliar languages and those which have, for no good reason, not found a place in widely accessible bibliographic databases. The literature search process is accepted as an important

0301-228X/99/030000+00$02.00 © 1999 British Orthodontic Society

part of developing a paper for publication, but the Bayesian paradigm requires it to be exhaustive. It is then necessary to assess the validity of each part of the evidence base, and quantify just what this tells us about the issue at hand.

Thus, the major issue is as follows. Currently, a paper submitted for publication presents the results of a single study. Other related studies are described in a background section, and the discussion section seeks to fit together findings of the present study and others informally. Should we, instead, require a research article to aim to identify, appraise, and incorporate all prior knowledge in a highly systematic and formal way? While this whole process is integral to the production of *systematic reviews* of research findings, it is a major imposition; it does not seem reasonable for this degree of depth to be required of every researcher before their paper can be regarded as valid as a contribution to the literature. Furthermore, it would be prohibitive for the editorial peer review process to have to validate that this had been done adequately.

In practice, most Bayesian analyses that are performed do not truly conform to the Bayes principle, but start by assuming nothing explicit is known. The model is initialized by using 'vague prior knowledge', which is given very little weight. This approach is computationally relatively simple, but logically often not really satisfactory from a truly Bayesian viewpoint.

Estimation and Confidence Intervals

The Bayesian paradigm is not the only alternative to the traditional over-emphasis on statistical significance. It is widely accepted that a very helpful way to present the results of a study is to give *point and interval estimates* of relevant quantities. This is done primarily for a single study, but results of several studies may also be fitted together. In a descriptive study, we might simply estimate a population parameter such as a mean bond strength or percentage of bonds failing, together with a confidence interval. As with all statistical calculations, the important assumption that the sample is representative of the relevant population needs to be substantiated. Usually, a 95 per cent confidence interval is calculated. The width of the confidence interval expresses the precision to which we can claim to have estimated the parameter, in view of the sample size studied. A four-fold increase in sample size is required to halve the width of the confidence interval. The lower and upper limits can be interpreted directly for clinical importance; in the one case, they are in bond strength units, in the other, percentage failure rates.

In a comparative study, a measure of *effect size* can be calculated, together with a confidence interval. Effect size measures include absolute ones such as the difference between two means or two proportions, and relative ones such as the relative risk and the odds ratio. This approach is complementary to hypothesis testing, but much more satisfactory. It is more directly informative and also the criticisms that apply to hypothesis tests do not apply nearly so strongly to confidence intervals—in fact, Matthews' short article makes no mention of them. A decade ago, leading journals such as the *British Medical Journal* adopted a policy that confidence intervals are preferable to hypothesis tests in presenting research findings; while hypothesis tests are not outlawed, they are to be regarded as complementary and secondary, the primary emphasis in interpretation is on the estimate and its confidence interval. The *BJO* fully endorses this policy. This principle is also incorporated in widely adopted standards for the reporting of clinical trials (Begg *et al*., 1996; Altman, 1996). The *BMJ*'s policy is explained in a booklet entitled *Statistics with Confidence* (Gardner & Altman, 1989), which also presents methods for calculating confidence intervals for a range of situations. Software (Confidence Interval Analysis) implementing these methods is also available. A second edition of the book and software are scheduled for publication this year, incorporating major improvements in methods for proportions and differences between proportions. The *BJO* intends to publish a review.

Dr R. G. NEWCOMBE
Senior Lecturer in Medical Statistics
University of Wales College of Medicine
Heath Park, Cardiff

References

Altman, D. G. (1996)
Better reporting of randomised controlled trials: the CONSORT statement,
British Medical Journal, **313**, 570–571.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., Pitkin, R., Rennie, D., Schultz, K. F., Simel, D. and Stroup, D. F. (1996)
Improving the quality of reporting of randomized controlled trials. The CONSORT statement,
Journal of the American Medical Association, **276**, 637–639.

Gardner, M. J. and Altman, D. G. (1989)
Statistics with Confidence—confidence intervals and statistical guidelines,
British Medical Journal.