International Journal of Epidemiology 2003;**32**:693–694
DOI: 10.1093/ije/dyg274

Commentary: Null points—has interpretation of significance tests improved?

Jonathan Sterne

Perhaps the most striking impression gained in reading Berkson's piece,[1] more than 60 years since its publication, is the author's struggle with questions of interpretation that still plague those conducting and interpreting statistical analyses today. Berkson seems to make little progress with solutions to the problems he presents, so it is of interest to see how statisticians today might deal with them.

A *P*-value (significance level) is used to assess evidence against a null hypothesis. If, as Berkson states, we do not 'find experimentalists typically engaged in disproving things' then why does the formulation of statistical questions in terms of null hypotheses and their falsification remain so pervasive? Of course, the idea of science as a process of falsification was articulated in detail by Popper, and remains an attractive explanation of why, for example, Newton's laws of mechanics were accepted until Einstein proved that there were circumstances in which they did not hold. Nonetheless Berkson argues forcefully that the usual discussion of evidence takes the form of positive statements ('Someone has been murdered') rather than negative ones ('… evidence against the null hypothesis that no one is dead').

Why, in medical and psychological statistics, do we remain so attached to the formulation of null hypotheses? In the context of randomized trials, it still seems reasonable to demand that those proposing that resources be spent on a particular treatment should, as a minimum, provide evidence against there being no treatment effect at all. Similarly, so many factors have been postulated over the years, in the pages of this and other epidemiology journals, to be associated with a multitude of disease outcomes that some quantification of the possible role of chance in explaining observed results is enduringly useful. Further, in choosing a statistical model it is inevitable that we make decisions about the inclusion or otherwise of different covariates, different forms of these covariates (linear, non-linear, categorical), and interactions between them. Such a process is difficult to conduct without some recourse to null hypotheses stating that certain parameter values in a more complex model are zero, and hence that a simpler model is appropriate.

It thus remains the case that there are genuine reasons for consideration, in the reporting of our statistical analyses, of the extent to which the data are compatible with particular null hypotheses. However, confusion still reigns over how this should be assessed: often manifesting itself in a confused mixture of Fisherian significance testing and Neyman-Pearson hypothesis testing.[2] As discussed in more detail in the commentary by Stone,[3] Fisher emphasized that research workers interpret significance levels in the light of their wider knowledge of the subject.[4] In contrast, Neyman and Pearson attempted to replace the subjective interpretation of *P*-values with an objective, decision-theoretic interpretation of results.[5] However, both methods are misused: Fisher's in that null hypotheses are mechanistically rejected if $P < 0.05$, and Neyman and Pearson's in that results are interpreted without consideration of the Type II error rate that should be used in defining the critical region within which values of the test statistic lead to rejection of the null hypothesis. As shown by Oakes,[6] interpretation of *P*-values depends on both the power of tests and on the proportion of null hypotheses that are truly false.[7]

How would modern medical statistics deal with the problems that Berkson raises? There has been a struggle since the 1970s in the pages of general medical journals against the misinterpretation of 'non-significant' differences (referred to by Berkson as 'middle *P*'s') as providing evidence *in favour* of the null hypothesis.[8] We now understand that *P*-values alone cannot be used to interpret statistical analyses: we need to consider the magnitude of estimated associations, and to examine confidence intervals in conjunction with *P*-values to prevent ourselves from being misled.[9] For example, well-reported analyses of the 'experiences' presented in Berkson's Table 1 might note that the odds of success in judging the sex of a fetus were 1.5 (95% CI: 0.42, 5.32) in Experience 1, compared with 1.02 (95% CI: 0.90, 1.15) in Experience 2. Examination of these results would lead most people to agree with Berkson's informally derived conclusions, though we might disagree with Berkson in noting that Experience 2 leaves open the possibility of the physician being able to discriminate modestly better (or worse) than by chance alone. A Bayesian statistician might present the posterior distribution for the odds of success, based on combining the data with her prior distribution—(s)he would reach similar conclusions unless using a strongly informative prior, in which case the small amount of data provided by Experience 1 would not change the prior distribution greatly.

Sadly, examples still abound of the misinterpretation of *P*-values in examples such as this. Many can be related to the issue raised by Berkson, that however carefully we teach the principles of (frequentist) statistical inference, those presenting statistical analyses wish to frame their conclusions in terms of positive statements (acceptance of the null hypothesis, evidence of no difference) rather than more convoluted but more appropriate statements such as 'no/little evidence that there was a difference between the groups' or 'no/little evidence against the null hypothesis of no association'.

A proponent of Bayesian statistics would have a simple response to Berkson's concerns: that rather than focusing on null hypotheses and significance tests we need to focus on the probability of the different parameter values (null and alternative hypotheses in the terms used by Berkson) given the data.

Department of Social Medicine, University of Bristol, Canynge Hall, Whiteladies Road, Bristol BS8 2PR, UK. E-mail: jonathan.sterne@bristol.ac.uk

Indeed, Berkson's paper hints at such an approach, in the section discussing the distinction between rejecting the null hypothesis H_0 and accepting the alternative hypothesis H_1. Certainly, the Bayesian approach can be used to demonstrate that conventionally statistically significant *P*-values do not necessarily correspond to strong evidence against the null hypothesis, in the sense that the when testing a normal mean a *P*-value of 0.05 implies that the posterior probability of the null is at least 0.3 in very general circumstances.[10] Whether the more widespread use of the Bayesian approach to statistical inference would itself reduce the misinterpretation of tests of evidence against null hypotheses remains unclear: for the reasons outlined above researchers will continue to require some form of probability statement quantifying the evidence regarding particular null hypotheses of interest. The development of Bayesian methods for model choice is an area of active research interest.[11]

Other examples used by Berkson are of situations in which confusion is caused by non-standard hypothesis tests: these would not be controversial in modern medical statistics but rather illustrate problems of interpretation familiar to anyone attempting to teach students how to fit linear terms in regression models and to check whether their models are appropriate. Such students often share Berkson's apparent confusion over the significant departure from linearity in his Chart 1. A first null hypothesis, of no linear association between facet number and temperature, corresponds to the slope of the regression line being equal to zero. Visual inspection of Chart 1 suggests to us, as it did to Berkson, that the data provide strong evidence against this null hypothesis. A second null hypothesis states that there is no *extra-linear* variation. These days, we could display the data rather better by including confidence intervals around the closed circles representing the mean facet number at each temperature, hence avoiding the misleading impression gained by Berkson that 'it appears as straight a line as one can expect to find in biological material'. We might proceed to test the second null hypothesis by including powers of temperature in our regression model, and would no doubt reach the same conclusion as Fisher, that the data provide evidence of extra-linear variation. Finally, we might test a number of further null hypotheses to check model assumptions, such as that the errors are not heteroscedastic.

Berkson concludes by calling for 'investigation into the finding of middle *P*'s', although he is 'not ready to say what this should be or just what it might mean'. It is tempting to imagine a contemporary commission of inquiry whose focus is not on the meaning of middle *P*'s, since as illustrated earlier this is well understood, but on whether blame for the continuing misinterpretation of *P*-values and significance tests lies with teachers of statistics,[12] its students and practitioners, or the counter-intuitive and difficult nature of the subject matter. I think we should be told.

References

1 Berkson J. Tests of significance considered as evidence. *J Am Statist Assoc* 1942;**37:**325–35.

2 Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *J Am Stat Assoc* 1993;**88:**1242–49.

3 Stone M. Commentary: Worthwhile polemic or transatlantic storm-in-a-teacup? *Int J Epidemiol* 2003;**32:**694–98.

4 Fisher RA. *Statistical Methods and Scientific Inference*. London: Collins Macmillan, 1973.

5 Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Phil Trans Roy Soc Ser A* 1933;**231:**289–337.

6 Oakes M. *Statistical Inference*. Chichester: Wiley, 1986.

7 Sterne JA, Davey Smith G. Sifting the evidence—what's wrong with significance tests? *BMJ* 2001;**322:**226–31.

8 Altman DG, Gore SM, Gardner MJ, Pocock, SJ. Statistical guidelines for contributors to medical journals. *BMJ Clin Res Ed* 1983;**286:** 1489–93.

9 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *BMJ Clin Res Ed* 1986; **292:**746–50.

10 Berger JO, Sellke T. Testing a point null hypothesis: The irreconcilability of *P* values and evidence. *J Am Stat Assoc* 1987;**82:** 112–22.

11 Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J Roy Statist Soc Ser B* 2002; **64:**583–616.

12 Sterne JA. Teaching hypothesis tests—time for significant change? *Stat Med* 2002;**21:**985–94.

International Journal of Epidemiology 2003;**32**:694–698
DOI: 10.1093/ije/dyg273

Commentary: Worthwhile polemic or transatlantic storm-in-a-teacup?

M Stone

Wouldn't it be wonderful if Berkson's quotation of Karl Pearson were true for all applications of the *higher statistics*? Readers will know from experience that it is not, and that journals such as this must keep alive the search for that elusive *common sense*—by letting voices from the past speak again and provoke responses that may help reduce the number of misapplications.

University College London, Department of Statistical Science, Gower Street, London WC1E 6BT, UK. E-mail: mervyn@stats.ucl.ac.uk