Fresh stirrings among statisticians: statistical commentary

Keith Godfrey

Orthodontic Department, Khon Kaen University, Thailand

For some years there has been unrest in the statistical world regarding the use of the *p*-value. It has been indicated that the significance of *p*-values is open to question, which therefore reduces the ability to measure the strength of evidence. This paper examines the use and misuse of the *p*-value and recommends consideration in its application.
(Aust Orthod J 2016; 32: 109–112)

Received for publication: February 2016
Accepted: March 2016

Keith Godfrey: keith_and_jill@yahoo.com.au

There has been some recent excitement among statisticians involving expressions of consternation and the publication of two policy statements.

One was from the editors of the journal Basic and Applied Social Psychology (BASP) in 2014 and amplified in their editorial in 2015.[1]

The expressions of consternation concerned the editor's advice that the consideration of articles that employed a null hypothesis significance testing procedure (NHSTP), and the accompanying inferential *p* statistic, would be banned. The 2015 editorial provided 'do's and don'ts' for researchers, with guidance and short explanations of what must have been numerous questions following their 2014 promulgation.

Briefly:

1. No inclusion of *p* values. They would no longer accept any statements about significant or non-significant differences.
2. Inferential statistics and procedures, such as confidence intervals (CIs) that were allied with NHSTP, would not be accepted. Inferences based on Bayesian methods may be accepted.
3. Rely on 'strong descriptive statistics, including effect sizes' as a requirement… 'also encourage the presentation of frequency or distributional data … [and] … use of larger sample sizes' (possibly more feasible among the subject base of psychology).

The Web provided numerous examples of antagonists, protagonists, and neutral commentaries addressing this apparent statistical 'earthquake'. Unlike an earthquake, this cannot be perceived as a sudden and unexpected change in the world of statistical analysis. To continue the metaphor, the problems concerning the use and abuse of significance testing have been generating heat for more than 70 years.[2]

A later policy statement (2016), the development of which was a triggered partial response to the above-noted editorial requirements, came from the American Statistical Association: 'ASA Statement on Statistical Significance and *P*-values.'[3]

To quote from the flyer[4] to the full ASA statement:

'The statement's six principles, many of which address misconceptions and misuse of the *p*-value, are the following:

1. *P*-values can indicate how incompatible the data are with a specified statistical model.
2. *P*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

The statement has short paragraphs elaborating on each principle.'

The full ASA statement provides a 'starter' reading list for individuals who would like to explore, in greater detail, the issue raised.

Baker[5] cites an interesting comment by the executive director and the senior author of this statement from the American Statistical Association (ASA): 'This is the first time that the 177-year-old ASA has made explicit recommendations on such a foundational matter in statistics.'

One 'trigger' for the preparation of the ASA statement was the statement from the editors of the BASP mentioned earlier.

An additional publication referenced was by Nuzzo, who cited 'A true story of what could have happened' to illustrate two aspects of the use and abuse of basing inferences on a *p*-values from a single study.[6] This story showed that eagerness and often acceptance for publication of a novel and 'significant' result ($p = 0.01$ in this case from a survey of 1,979 subjects) should be replaced by prudence and replicating the study. The result of repeating the study provided the authors with a salutary lesson in significance testing with a substantially different $p = 0.59$ (with 1,300 subjects). This was a cautionary tale of over-enthusiasm about a novel and 'significant' finding.

The possibilities of finding valid guidelines for clinical practice depend on the availability of research findings that meet the specific requirements for a systematic literature review and meta-analysis of data: [Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA)].[8,9] This depends on the availability of suitable statistical data or sometimes reconstruction of the data that is published. Bland[10] listed the statistical data available for group analysis for inclusion in research reports and which may be useful for the construction of meta-analyses, even from as few as two trials:

'We need to extract the information required from what is available.

1. standard errors – this is straightforward, as the formula is known for the standard error and so, provided the sample sizes are known, a standard deviation can be calculated,
2. confidence intervals – this is also straightforward, as we can work back to the standard error,
3. reference ranges – again straightforward, as the reference range is four standard deviations wide,
4. inter-quartile ranges – here an assumption is needed about distribution; provided this is normal we know how many standard deviations wide the IQR should be, but of course this is often not the case,
5. range – this is very difficult, as not only do we need to make an assumption about the distribution but the estimates are unstable and affected by outliers,
6. significance test – sometimes we can work back from a t-value to the standard error, but not from some other tests, such as the Mann Whitney U test,
7. *P* value – if we have a t-test we can work back to a t-value hence to the standard error, but not for other tests, and we need the exact *p* value.
8. 'Not significant' or '$p < 0.05$' – this is hopeless.'

Arguments abound over the preferred statistical test for a particular trial. One example is given here of how to apply statistical testing to the frequent studies of reliability in measurement comparisons in orthodontics. Donatelli and Lee[11] caution orthodontists on the use of correlation coefficient and *t*-test for such studies in favour of the Bland-Altman limits of agreement method (LoA).[12] (Robert Grant, in his 2013 blog, reported from his trawl of Google Search statistical paper citations, that the 1986 Bland-Altman paper was third among 'The world's favourite stats papers'.) The LoA method arose from dissatisfaction among medical researchers with traditional frequentist methods of statistical analysis of medical measurement data. The logical form of the LoA and, importantly, its graphical representation, provide all the necessary statistical information for judging the importance (forget *p*-values and 'significance'!) of any mean difference comparing two sets of measurements.[11-13]

The researcher may have no concern about including *p* values in a paper if not intending to submit it to the BASP journal. However, there could be the editor (or reviewer) for a professional journal who prefers the demarcation point for the 'test of significance' as 0.005 or 0.001 as suggested by Johnson.[14]

The problem is two-fold. On the one hand, there is substantial evidence that use of NHST in research studies is not justified, while on the other hand teachers with students, some statisticians, numerous textbooks, and some statistical software, continue to accept and provide use of NHST and *p* values, sometimes termed 'frequentist statistics', and as part of what has been termed 'traditional statistics'.[15-18]

Readers are encouraged to look up the critique on '*P*' presented by Emeritus Professor Geoffrey Cumming of Latrobe University.[19]

One might be inclined to 'blame' Sir Ronald Fisher, a doyen among statisticians and the originator of the $p \le 0.05$, for the furor that has continued over a long time. Fisher should be allowed some final words on the topic:[20] 'It is usual and convenient for experimenters to take the 5 per cent level and are prepared to ignore all results which fail to reach this standard. This means an elimination, from further discussion, of the greater part of the fluctuations which chance causes have introduced into their experimental results. No such selection can eliminate all of the possible effects of chance coincidence, and if we accept this convenient convention, and agree that an event which could occur by chance only once in 70 trials is decidedly "significant", in the statistical sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon… (page 15).'

In a later publication,[21] Fisher sought to kill off the 'null hypothesis' (which wasn't his invention anyway): 'In relation to any experiment we may speak of this hypothesis as the "null Hypothesis", and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.'

Acknowledgment

To my friend Mark Cordato, with whom I've been fortunate to share academic interests, and not least in encouraging me to expose myself as a non-statistician.

Corresponding author

Keith Godfrey
12/30-36 Belmont St
Sutherland
NSW, 2232
Australia

Email: keith_and_jill@yahoo.com.au

References

1. Trafimow D, Marks M. Editorial, Basic Applied Soc Psychol 2015;37:1-2.
2. Berkson J. Tests of significance considered as evidence. J Am Statist Assoc 1942;37:325-35.
3. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Statistician. 2016. <http://amstat.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108>
4. American Statistical Association releases statement on statistical significance and P-values, media release, American Statistical Association, 2016. <https://www.amstat.org/newsroom/pressreleases/P-ValueStatement.pdf>
5. Baker M. Statisticians issue warning over misuse of P values. Nature 2016;531:151.
6. Nuzzo RL. Scientific method: Statistical errors. Nature 214;506:150-2.
7. Nosek BA, Spies JR, Motyl M. Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspect Psychol Sci 2012;7:615-31.
8. PRISMA Transparent reporting of systematic reviews and meta-analyses. <http://www.prisma-statement.org/>.
9. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015;349:g7647.
10. Bland M. Meta-analysis: methods for quantitative data synthesis. University of York, 2006, <https://www-users.york.ac.uk/~mb55/msc/systrev/week6/meta_text.pdf>.
11. Donatelli RE, Lee SJ. How to report reliability in orthodontic research: Part 1. Am J Orthod Dentofacial Orthop 2013;144:156-61.
12. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
13. Bland JM, Altman DG. Applying the right statistics: analyses of measurement studies. Ultrasound Obstet Gynecol 2003;22:85-93.
14. Johnson VE. Revised standards for statistical evidence. Proc Natl Acad Sci U S A 2013;110:19313-7.
15. Haller H, Krauss S. Misinterpretations of significance: A problem students share with their teachers? Methods Psychol Res Online 2002;7:1-20.
16. Lecoutre M-P, Poitevineau J, Lecoutre B. Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. Int J Psychol 2003;38:37-45.

17. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. Ann Intern Med 1999;130:995-1004.
18. Svensson E. Important considerations for the optimal communication between statisticians and medical researchers in consulting, teaching and collaborative research, with a focus on the analysis of ordered categorical data. In: Batanero C, ed. Training researchers in the use of statistics, IASE Round table Conference, Tokyo 2000. International Association for Statistical Education, 2001:23-36.
19. Cumming G. The problem with p values: how significant are they, really? The Conversation, November 12 2013, <http://theconversation.com/the-problem-with-p-values-how-significant-are-they-really-20029>.
20. Fisher RA. The design of experiments, 8th edn. New York: Hafner Publishing Co. 1971.
21. Fisher RA. Statistical methods and scientific induction. J Royal Stat Soc Series B (Methodology) 1955;17:69-78.