Dear Prof Gelman,

I have been reading your papers and blogs on the “p-value threshold” with considerable interest. After retiring a few years ago, I lost access to the statistics and biomedical literature, and have found your blog extremely useful for learning about a variety of current statistical issues. So, many thanks. I have learnt a lot by reading your blogs.

For a couple of decades (from about 1988 to 2006) I was employed as a support statistician, and became very interested in the p-value issue; hence my interest in your contribution to this debate. (I am not familiar with the p-value 'reconciliation' literature, as published after about 2005.) I would hugely appreciate it, if you might find the time to comment further on some of the questions listed in the attachment to this email.

I realise you may not have the time to look at every one of these questions. In this case I would be particularly interested in learning more about your views on strict Neyman-Pearson hypothesis testing, based on critical values (critical regions), given an insistence on power calculations among research funding organisations (i.e., first section headed 'p-value thresholds'), and the long-standing recommendation that biomedical researchers should focus on confidence intervals instead of p-values (i.e., last section headed 'estimation and confidence intervals').

Yours sincerely,
Martin King
(Retired. Previously, MRI researcher and support statistician, London. UK)

p-value thresholds*

My main question is about 'dichotomous thinking' and p-value thresholds. McShane and Gal (2017, page 888) refers to “*dichotomous thinking and similar errors”*. Is it correct to say that dichotomous thinking is an error? It is my impression that this is central to the Neyman-Pearson (NP) hypothesis testing method. Many might prefer to adopt some alternative approach (a Bayesian analysis or Fisher's inductive method, for example), but is it correct to suggest that it is an error to adopt a strict NP decision analysis based on critical values?

To elaborate. In biomedical and clinical research, many funding bodies** insist on power calculations. Researchers applying for funding have no say in this. By definition this forces researchers to consider the notions of Type I and Type II error rates, even if some consider this to be a nonsense. (I am aware of your lack of enthusiasm for the concepts of Type 1 error, Type 2 error etc.) Surely, these error rates depend on adopting a very strict threshold. Having forced biomedical researchers to adopt this approach, does it not create a double nonsense to insist that they abandon the very rule on which the error rates depend (i.e., decision/action based on a strict threshold*). As I see the situation, this makes a nonsense of the very notion of error rates, even if one takes the view that the binary reject-accept hypothesis testing approach to many problems in biomedicine is a very bad idea. So, if funding bodies insist on strict hypothesis testing (otherwise why the insistence on power analysis, as opposed to some other assessment of adequate precision), is it fair to criticise researchers for obeying the rules dictated by the method? In summary, before banning p-value thresholds, do you have to persuade the funding bodies to abandon their insistence on power calculations, and allow applicants more flexibility in showing that a proposed study has sufficient precision.

(*A question about terminology. On reading various papers and website contributions to the p-value threshold discussion, I had assumed that the objection was to the Neyman-Pearson (NP) deductive method based on using a strict critical region, thus arriving at a decision (accept/reject). I realise that most formal treatments (Neyman and Pearson, 1933, for example) do not use p-values, and refer to the critical region for the test statistic. Thus, before computers became common place, data analysts would refer to tables of critical values. But, I assume that most data analysts now use software that lists both the value of the test statistic and associated probability, and that it is now common practice to compare this probably with alpha, equivalent to comparing the test statistic with the critical level. Am I correct to assume that it is this strict NP decision approach that you object to, regardless of using a p-value threshold, or the alternative based on the critical value for the test statistic? It occurs to me that some statisticians might reserve the term p-value for Fisher's inductive method, taking the view that p-values are not part of the NP method. (It is well known that Fisher hated the very idea of making decisions based on strict rules involving critical values [see Goodman, 1993, for example].) If those that object to a strict p-value threshold do so because thresholds have no place in Fisher's inductive method, and do not object to thresholds as a central component of NP hypothesis testing, then there should be no disagreement about thresholds, per se. The discussion shifts to one regarding the merits of inductive (Fisher) and deductive (NP) inference. So I am not entirely sure what is meant by 'abandon thresholds'. Is the suggestion that the NP decision approach should be banned?)

(**I retired a few years ago, but doubt that much has changed since then. It was my impression that, in the UK, a biomedical/clinical research funding application stands no chance of success, unless power calculation results are provided. Is this so in the USA?)

Statistics courses; what should be taught?

This brings us to the second question regarding what should be taught in statistics courses, aimed at biomedical researchers. A teacher might want the freedom to design courses that assumes an ideal world in which statisticians and researchers are free to adopt a rational approach of their choice. Thus, a teacher might decide to drop frequentist methods (if she/he regards frequentist statistics a nonsense) and focus on the alternatives. But this creates a problem for the course recipients, if grant awarding bodies and journal editors insist on frequentist statistics?

In addition to these two issues (thresholds and course content), you have made a number of other suggestions and comments in your blog. I would be extremely interested to hear your opinion regarding some of the questions I have on these issues.

Is there a need for change?

The first thing that struck me about the McShane-Gal (2017) publication and related material is that the thresholds issue seems to be a re-run of the Fisher versus Neyman-Pearson debate. Despite decades of acrimony, it appears that the matter remains unresolved. Edwards (1972, p179-180) discussed what he referred to as "the *approved statistical techniques"* in hypothesis testing, and refers to a *"dangerous nonsense (dressed up as 'the scientific method')",* which *"will cause much trouble before it is widely appreciated as such"*. Where are we today? In 2014 the Lancet published several papers on biomedical research and how to reduce wasted resources. One paper (Macleod et al., 2014) referred to previous reports (Chalmers and Glasziou, 2009, for example) and suggested "*that about 85% of research investment - equating to $200 billion of the* *investment"* was wasted in a single year (presumably a global figure). This message was not new. Two decades previously Altman published a paper with the title "The scandal of poor medical research" stating that huge sums of money are spent annually on research that is seriously flawed (Altman, 1994). He referred to incorrect analyses and faulty interpretation as among the problems. Altman suggested that "*the system encourages poor research*" (pressure to publish etc). Your blog gives the strong impression that faulty/invalid statistical analysis and misinterpretation remain a major problem. I never understood why the funding bodies (and journal editors) appear to have so little concern for this huge waste of money.

p-value thresholds, continued

Regarding the comments that "*even expert statisticians*" "*engage in dichotomous thinking and thus misinterpret data*" (McShane and Gal, 2017, page 888) and that "*researchers who are primarily statisticians are also prone to misuse and misinterpret p-values*" (McShane and Gal, 2017, Abstract): are they saying that NP, and other tests based on critical values is definitively wrong? Are they saying that NP is discredited from a mathematical/theoretical perspective? If not, should individual researchers not have the freedom to adopt this approach? Referring to your statement (McShane et al., 2018, Appendix A): "*While this formalism allows for mathematical optimization under some restricted collection of distributions and testing problems, it is quite rudimentary from a decision-theoretic point of view, even to the extent of failing …*", I am not entirely sure whether you and your co-authors are saying that, Neyman-Pearson (1933), for example, is mathematically flawed.

For several decades, I worked in an experimental biomedical research environment, and in that setting I preferred Fisher's approach, as opposed to strict Neyman-Pearson hypothesis testing. I saw the latter as an unnecessary straight-jacket that prohibits the researcher from engaging in individual thought. It is well known that Fisher was extremely hostile to the "*concept that the scientific worker can regard himself as an inert item in a vast co-operative concern working according to accepted rules",* and to the notion of the "*supposed duty mechanically to make a succession of automatic 'decisions', deriving spurious authority from the ... mathematics of the theory of Decision Functions"* (Fisher, 1973, p104). (Similarly, you and your colleagues (McShane and Gal, 2017; McShane et al., 2018), refer to the "*rote and recipe-like application of NHST*".) I find Fisher's argument appealing. The notion that one should present a reasoned argument in which p-values might be used as a subjective measure of evidence, incorporating relevant additional evidence, to make the case for the plausibility of some proposition, makes sense. But, as stated above, the demand for power calculations, amounts to a rejection of this approach, and substitution of strict NP decisions, based on critical values (thresholds). Furthermore, the NP papers were published in respectable statistical journals. If some researchers dislike the subjectivity of Fisher's method, or Bayesian analysis, and wish to adopt strict NP, can this be dismissed as a misused statistical method? Similarly, if an individual researcher misinterprets the critical value (e.g., thinks that it provides a probability relating to a specific observation and hypothesis under investigation, as opposed to a long-run average, or takes rejection of the null as definitive evidence in favour of some preferred alternative (McShane et al., 2018) or makes mistaken statements alluding to zero effects, based on a failure to reject the null, is this a reason for forbidding others from undertaking a legitimate NP analysis? Restating the question, is it right to ban everybody from using a legitimate method because some individuals make a mistake? Are the mistakes made by some individuals a valid reason for disallowing NP in its entirety?

Regarding the comments that "*the p-value was never intended to be a substitute for scientific reasoning*" (McShane and Gal, 2017), and that "*no single index should substitute for scientific reasoning*" (McShane and Gal, 2017; McShane et al., 2018, and references therein); presumably, both Fisher and NP would agree with this. For example, Fisher argued that "*mathematical probability is inadequate to express the nature and extent of our uncertainty in the face of certain types of observational material*" (Fisher's correspondence with D.J. Finney, 1954; taken from Bennett, 1990, p92)*,* and that "*although some uncertain inferences can be rigorously expressed in terms of mathematical probability, it does not follow that mathematical probability is an adequate concept for the rigorous expression of uncertain inferences of every kind*" (Fisher, 1935, p40)*.* Similarly*,* Neyman and Pearson (1933, p291) suggest that, "*as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis*". Presumably, the only circumstance where evidence can be definitive is in those special cases were falsification is possible. Otherwise, as has been suggested on numerous occasions (including, for example, Laber and Shedded, 2017; McShane et al., 2018) the best that can be done is a formal analysis that considers the costs and benefits associated with alternative decisions and associated errors although, as acknowledged in McShane et al., (2018), this would be very difficult to apply in a basic sciences research study undertaken to improve our understanding of some biological/pathophysiological process. From this perspective, presumably one must distinguish between studies in which the objective is to present evidence with a view to gaining knowledge, as opposed to public health policy and decision making (large population questions at one extreme, and treatment planning for individual patients, at the other extreme).

In one of your blogs you say that researchers, including statisticians, routinely misinterpret p-values "*as some sort of replication probability*". I assume you are referring to data-specific p-values as opposed to pre-study critical values (alpha). Neyman and Pearson, 1933 (page 291) refer to "*rules*" to ensure "*that, in the long run of experience, we shall not be too often wrong*". They suggest that "*according to such a rule, then in the long run we shall reject H when it is true no more, say, than once in a hundred times ...*" Their legend to Fig 1 includes "*if the process of sampling was repeated many times ...*" It is my understanding that the concept of hypothetical long-run/repeated sampling that is fundamental to the frequentist method, and the resulting long-run error rates, remain valid providing analysts obey the strict NP accept-reject decision rule, based on a critical region (threshold). I realise, of course, that there is no consensus among statisticians regarding the utility of the NP method; but this simply returns us to the frequentist-Fisher-Bayesian debate. Although a number of researchers have made a convincing case for adopting a Fisherian approach in preference to NP, I doubt that NP enthusiasts will take the view that these translate into NP being declared invalid. I understand the reason why many statisticians, you included, reject the NP method as an inferior mode of inference compared with some of the alternatives. But I am unsure whether you reject the very notion of long-run error frequencies, regardless of the context. For example, do you reject as invalid the frequentist/NP approach in one-sided hypothesis testing? (The argument that estimation is preferable to hypothesis testing in many one-sided problems (see section on p-value reconciliation and the point null) is convincing but, presumably, that does not mean that frequentist one-sided hypothesis tests are invalid.)

Point null hypotheses

In McShane et al. (2018) you and your co-authors say that p-values are, in a majority of applications, "*defined relative to the generally implausible and uninteresting sharp point* *null*". Below I reference some of the 1980s 'p-value reconciliation' literature where it is suggested that, in many instances, one-sided tests are more appropriate. From that perspective, do some statisticians argue that 2-sided tests based on a point null are more conservative than the corresponding more plausible one-sided (composite) test? Is this the reason why few data analysts worry about the plausibility of the point null?

It is suggested (McShane et al. 2018) that researchers often fail to provide sufficient information on *currently subordinate factors*. I spent many years working in an experimental biomedical environment, and it is my impression that most experimental biomedical researchers do present this kind of information. (They do not spend time doing experiments that are not expected to work or collecting data that are not expected to yield useful and substantial information. It is my impression that some authors go to the extreme in attempting to present an argument for relevance and plausibility.) Do you have a specific literature in mind where it is common to see results offered with no regard for motivation, relevance, mechanism, plausibility etc. (apart from data dredging/data mining studies in which mechanism and plausibility might be elusive)?

Fisher-Neyman-Pearson debate, cont

As I have indicated above, I find the Fisher argument that p-values should be used together with reasoned argument very appealing, but am under the impression that statisticians were among those that had argued that this had led to a dangerous lack of rigour, with too much of the literature filled with unreliable work, heavily based on subjectivity and judgement. The reasoned argument approach was not rigorous. An alternative is to insist that the standard scientific method should be free from subjectivity and bias (allowing the data to speak). As you say in McShane et al. (2018), an approach based on thresholds (critical regions) provides "*a valuable brake*" on subjectivity. Thus the key was scientific objectivity although, in many instances, this might be an illusion (Berger and Berry, 1988). I note that you and your co-authors (McShane et al., 2018) have also pointed out that a "*p-value is not a purely objective standard*", and that "*no single number … is capable of eliminating subjectivity and personal* biases". Nevertheless, NP was supposed to provide this objectivity, and it is my impression that, for this reason, many funding bodies and editors want to see frequentist/NP statistics. This does not preclude the provision of convincing statements on motivation, together with reasoned argument.

Institutional change

This takes us back to the question of institutional structure and the role of funding agencies, editors and reviewers, a matter that you and your co-authors have discussed (McShane et al., 2018, page 17), and suggested the need for reform. If the suggestion is that researchers should be released from those institutional constraints that demand analytical/statistical methods that are counter-productive and irrational (you refer to a "*rote and recipe-like application of NHST*"), then this reversal takes up back to the days when researchers were accused of being too subjective in their approach. As I have said, I get the impression that a perceived lack of rigour and objectivity was the reason why strict NP was adopted so widely. The suggestion that this should be abandoned takes us full circle.

p-value reconciliation and the point null. Bayes factors

You have suggested (McShane et al. 2018) that "*it seldom makes sense to calibrate evidence as a function of p-values*". This is a topic that has interested me in the past, and I would be very pleased to hear your views on the following.

I notice that Casella and Berger (1987b, page 134) make a similar statement, namely, that "*it is hopeless to calibrate p-values to posterior probabilities, but we were not calibrating*". They suggest that, although p(x) and p(H0|x) are "*seemingly related measures of statistical evidence",* since "*they are based on different assumptions*, …. *a general attempt at calibration is doomed to fail*". Is this the reason why you reject p-value calibration? My interest in the literature on p-value reconciliation was two-fold. Firstly, it was my first introduction to analyses showing that the weight of evidence associated with a given p-value could be surprisingly low, something I had not previously appreciated. I thought this was an important message. Secondly, I worked in a clinical/biomedical environment where busy clinicians were engaged in research, some of whom did not have statistical support, and needed to undertake their own analyses. In this environment there is a need for methods that can be adopted by busy clinicians who cannot be expected always to take on demanding methods. Hence my interest in papers aimed at presenting 'accessible statistical methods' for biomedical researchers (for example, the Altman-Bland BMJ statistical notes). As an example, I was interested in the suggestion that minimum Bayes factors might be used as an alternative to hypothesis testing (Goodman, 1999, 2001). At least, the minimum Bayes factor appears to offer a less exaggerated measure of strength of evidence, relative to the p-value obtained for some tests. I understand that you are not enthusiastic about using Bayes factors to evaluate strength of evidence (Gelman and Carlin, 2017). Apart from the technical difficulties discussed in Gelman and Carlin (2017, page 3), is it the classification component of the approach, alone, that you find objectionable, or Bayes factors (minimum Bayes factors) per se? Restated, if one resists the "*temptation to discretize continuous evidence*" (obviously the concept of frequentist error rates do not apply), do you regard minimum Bayes factors (Goodman, 1999, 2001), for example, an acceptable measure of evidence for those wanting to perform a relatively simple analysis?

Several prominent papers on p-value reconciliation deal with the point null hypothesis (Berger and Sellke, 1987a; Berger and Delampady 1987a, 1987b; Sellke et al., 2001; etc.) but you and co-authors are among many researchers who have pointed out the unsatisfactory nature or implausibility of the point null in most applications (McShane et al., 2018, and references therein; Casella and Berger, 1987a). In fact, several authors that contributed to the 1987 collection of papers on p-value reconciliation suggested that many important problems are one-sided with no strong prior belief in a point null (Casella and Berger, 1987a, p106; Casella and Berger, 1987c, p345). For example, with reference to point null hypotheses, Casella and Berger (1987a, p106) state that "*there is a direction of interest in many experiments, and saddling an experimenter with a two-sided test would not be appropriate*", a point that was discussed at length by others (see, for example, Berger and Sellke, 1987b, p136; Berger and Delampady 1987a, p326-7). The reason for mentioning this is that, having argued that the one-sided test applies to the majority of practical problems, Casella and Berger (1987a) go on to provide several formal results, involving the infimum of $Pr(H_0|data)$, that appears to vindicate using p-values as a measure of evidence in one-sided problems. I notice the comment from Berger and Mortera (1999, p543), which questions the validity of using a lower bound in isolation, thus taking the result that is least favourable to the null. That said, and referring to your lack of enthusiasm for attempting to calibrate evidence as a function of p-values, are you saying that you dismiss this literature on p-value reconciliation as unhelpful or irrelevant? Is there nothing to learn from these results?

Estimation and confidence intervals

As suggested above, many analysts might agree that a majority of biomedical problems are one-sided, with no strong prior belief in the point null, but some have gone further and suggested that most one-sided problems are concerned with estimation (discussed in Berger and Sellke, 1987b, p136; Berger and Delampady 1987a, pages 326-8; Casella and Berger, 1987b, p134, for example) in which the investigator wishes to determine the magnitude of an effect. This brings me to my next question, which is concerned with confidence intervals (CIs) and the recommendation that CIs be used in an assessment of practical/application significance, as opposed to statistical significance.

Altman is a well-known medical statistician in the UK, partly due to a series of statistical notes, written with Martin Bland and others, and published over a period of many years in the British Medical Journal (BMJ). Predating this series, Altman, together with Gardner and others, used the BMJ as a platform to encourage biomedical researchers to engage in estimation as an alternative to hypothesis testing, and the use of confidence intervals rather than p-values (Gardner and Altman, 1886, 1988, 1989; also see Sterne, 2001, who also suggest that CIs should be interpreted in terms of the clinical implications). I notice that you and your co-authors (McShane, et al., 2018) suggest that confidence intervals cannot provide a quick fix, but I assume this reservation applies only to those who assess the interval in terms of some critical value (typically zero). I had taken the view that CIs assessed in terms of clinical objectives (marked effect versus insignificant effect, based on clinical criteria) is a huge improvement over the rote NP accept/reject approach. But it does occur to me that, without careful wording, this might be invalid and a misuse of confidence intervals given their frequentist definition.

To restate the question for the sake of clarity; just as Berger and Delampady (1987a, p329), say that alpha has a valid frequentist interpretation, but "*data-dependent p-values have no such interpretation*", presumably the same can be said of CIs. As they say, "*any type of repetitive 'error rate' can be accused of addressing the wrong question"*. So, is it wrong to examine CIs with the clinical objective in mind (worthwhile clinical effect v ignorable clinical effect)? Is this valid in a discussion regarding estimation and precision, as it applies to the specific study, given the frequentist nature of CIs? I realise the answer is to perform a Bayesian analysis and use the posterior distribution, but for those who wish to stick to frequentist methods, should they use CIs as suggested by Altman, Bland and others, adopting the NP stance that they will not be too wrong, too often? While writing this I am not entirely sure. I notice the Carlin and Louis (1996, p 2-3) statement that using CIs "*in any single data-analytical setting is somewhat difficult*". But I had previously taken the view that it is valid to examine the confidence limits in terms of relevant clinical criteria, provided the underlying frequentist interpretation is acknowledged. Presumably, the fact is that nothing (Bayesian or frequentist) can be done to overcome the unavoidable uncertainty that arises in most studies. I would be very interested to hear your opinion on this matter, and whether it is wrong to deviate from a strict (but unhelpful) frequentist interpretation of CIs.

Final question on p-values as a measure of evidence

For many years it had not occurred to me that there is a distinction between looking at p-values (or any other measure of evidence) obtained as a participant in a research study, versus looking at third-party results given in some publication, because the latter have been through several unknown filters (researcher selection, significance filter etc). Although others had commented on this problem, it was your discussions on the significance filter that prompted me to fully realise the importance of this issue. Is it a fact that there is no mechanism by which readers can evaluate the strength of evidence in many published studies? I realise that pre-registration has been proposed as a partial solution to this problem. But it is my impression that, of necessity, much experimental and basic biomedical science research takes the form of an iterative and adaptive learning process, as outlined by Box and Tiao (pages 4-5), for example. I assume that many would find It difficult to see how pre-registration (with constant revision) would work in this context, without imposing a massive obstacle to making progress.

References

Altman, DG. (1994) The scandal of poor medical research. Br. Med. J. 308:283-284.

Bennett, JH. (1990) Statistical inference and analysis. Selected correspondence of R.A. Fisher. Clarendon Press. Oxford.

Berger, JO. and Berry, DA. (1988) Statistical analysis and the illusion of objectivity. American Scientist 76: 159-165.

Berger, JO. and Delampady, M. (1987a) Testing precise hypotheses. Statistical Sci. 2: 317-335.

Berger, JO. and Delampady, M. (1987b) Testing precise hypotheses. Rejoinder. Statistical Sci. 2: 348-352.

Berger, JO. and Mortera, J. (1999) Default Bayes factors for nonnested hypothesis testing. J. Am. Stat. Assoc. 94: 542-554.

Berger, JO. and Sellke, T. (1987a) Testing a point null hypothesis: The irreconcilability of *P* values and evidence. J. Am. Stat. Assoc. 82: 112-122.

Berger, JO. and Sellke, T. (1987b) Testing a point null hypothesis: The irreconcilability of *P* values and evidence: Rejoinder. J. Am. Stat. Assoc. 82: 135-139.

Box, GEP. and Tiao, GC. (1973) Bayesian inference in statistical analysis. Addison-Wesley; Massachusetts.

Carlin, BP. and Louis, TA. (1996) Bayes and empirical Bayes methods for data analysis. Chapman and Hall; London.

Casella, G. and Berger, RL. (1987a) Reconciling Bayesian and frequentist evidence in the one-sided testing problem. J. Am. Stat. Assoc. 82: 106-111.

Casella, G. and Berger, RL. (1987b) Testing a point null hypothesis: The irreconcilability of *P* values and evidence: Rejoinder. J. Am. Stat. Assoc. 82: 133-135.

Casella, G. and Berger, RL. (1987c) Testing precise hypotheses. Comment. Statistical Sci. 2: 344-347.

Chalmers, I. and Glasziou, P. (2009) Avoidable waste in the production and reporting of research evidence. Lancet 374: 86–89.

Edwards, AWF. (1972) Likelihood: an account of the statistical concept of likelihood and its application to scientific inference. Cambridge University Press. London.

Fisher, RA. (1935) The logic of inductive inference. J. R. Statist. Soc. 98: 39-82.

Fisher, RA. (1973) Statistical methods and scientific inference. 3rd ed. Macmillan; New

York.

Gardner, MJ. and Altman, DG. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. Br. Med. J. 292: 746-750.

Gardner, MJ. and Altman, DG. (1988) Estimating with confidence. Br. Med. J. 296: 1210-1211.

Gardner, MJ. and Altman, DG. (Eds) (1989) Statistics with confidence. British. Medical. Journal. London.

Gelman, A. and Carlin, J. (2017) Some natural solutions to the p-value communication problem - and why they won't work http://www.stat.columbia.edu/~gelman/research/published/jasa_signif_2.pdf

Goodman SN. (1993) p Values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. Am. J. Epidemiol. 137: 485-496.

Goodman, SN. (1999) Towards evidence-based medical statistics. 2: The Bayes factor. Ann. Intern. Med. 130: 1005-1013.

Goodman, SN. (2001) Of p-values and Bayes: A modest proposal. Epidemiology. 12:295-297.

Laber, EB. and Shedden, K. (2017), Statistical Significance and the Dichotomization of Evidence: The Relevance of the ASA Statement on Statistical Significance and p-values for Statisticians. J. Am. Stat. Assoc. 112:902-904.

Macleod, MR., Michie, S., Roberts, I. et al. (2014) Biomedical research: increasing value, reducing waste. *Lancet* 383: 101-104.

McShane, BB. and Gal, D. (2017) Statistical Significance and the Dichotomization of Evidence: *J. Am. Stat. Assoc.* 112: 885-908.

McShane, BB., Gal, D., Gelman, A., Robert, C., Tackett, JL. (2018) Abandon statistical significance. https://arxiv.org/pdf/1709.07588.pdf

Neyman J. and Pearson, ES. (1933) On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. Roy. Soc. A. 231: 289-337.

Sellke, T., Bayarri, MJ., Berger, JO. (2001) Calibration of *p* values for testing precise null hypotheses. The American Statistician 55: 62-71.

Sterne, JAC. and Smith, GD. (2001) Sifting the evidence-what's wrong with significance tests? Br. Med. J. 322: 226-231.