The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: https://www.tandfonline.com/loi/utas20

Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires

Stuart H. Hurlbert, Richard A. Levine & Jessica Utts

To cite this article: Stuart H. Hurlbert, Richard A. Levine & Jessica Utts (2019) Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires, The American Statistician, 73:sup1, 352-357, DOI: 10.1080/00031305.2018.1543616

To link to this article: https://doi.org/10.1080/00031305.2018.1543616

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 352–357: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2019.1543616

Coup de Grâce for a Tough Old Bull: "Statistically Significant" Expires

Stuart H. Hurlbert[a], Richard A. Levine[b], and Jessica Utts[c]

[a]Department of Biology, San Diego State University, San Diego, CA; [b]Department of Statistics, San Diego State University, San Diego, CA; [c]Department of Statistics, University of California, Irvine, CA

ABSTRACT

Many controversies in statistics are due primarily or solely to poor quality control in journals, bad statistical textbooks, bad teaching, unclear writing, and lack of knowledge of the historical literature. One way to improve the practice of statistics and resolve these issues is to do what initiators of the 2016 ASA statement did: take one issue at a time, have extensive discussions about the issue among statisticians of diverse backgrounds and perspectives and eventually develop and publish a broadly supported consensus on that issue. Upon completion of this task, we then move on to deal with another core issue in the same way. We propose as the next project a process that might lead quickly to a strong consensus that the term "statistically significant" and all its cognates and symbolic adjuncts be disallowed in the scientific literature except where focus is on the history of statistics and its philosophies and methodologies. Calculation and presentation of accurate *p*-values will often remain highly desirable though not obligatory. Supplementary materials for this article are available online in the form of an appendix listing the names and institutions of 48 other statisticians and scientists who endorse the principal propositions put forward here..

ARTICLE HISTORY
Received March 2018
Revised October 2018

KEYWORDS
NeoFisherian significance assessment; Statistical significance; Type I error; *p*-values; Dichotomized language; Teaching of statistics

An error does not become truth by reason of multiplied propagation, nor does truth become error because nobody will see it.

– Mahatma Gandhi

1. Introduction

Its cogency, clarity, and sharp focus on a limited set of important issues make the *ASA Statement on Statistical Significance and p-Values* (Wasserstein and Lazar 2016) of great potential value. It should be sent to the editor-in-chief of every journal in the natural, behavioral and social sciences for forwarding to their respective editorial boards and stables of manuscript reviewers. That would be a good way to quickly improve statistical understanding and practice.

The "call for papers" for this issue advises potential contributors to avoid "lengthy discussions of 'Don'ts,' which are already addressed effectively in the ASA statement and supplementary materials." We demur slightly. Principle 3 in the ASA statement reads: "Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold." That and accompanying explanatory text warrant being supplemented to make its operational imperatives clearer to statisticians, researchers, and editors. We believe those most knowledgeable about this issue may be close to a consensus that scientific conclusions and business or policy decisions should almost *never* be based on whether a *p*-value passes a specific threshold because in weighing the strength of evidence there is no need for arbitrarily defined thresholds.

We acknowledge with thanks the many useful suggestions on this manuscript provided by editors, referees, and several of the endorsers listed in Appendix A, especially Sander Greenland.

2. The Bull

An apt metaphor for the phrase "statistically significant" and its relatives is that of a *toro bravo*, a champion bull raised for bullfighting who is now on his last legs and awaiting only the coup de grâce. This particular bull was "bred" a century ago by British and Polish gentlemen to "fight," ostensibly, for more objectivity in data interpretation and decision-making. Since then the uncontrolled rampages of this toro bravo have damaged many on the arena floor, victims of bad statistical advice, reasoning, and decision-making. He remains a tough character having been periodically steroid-doped by his caretakers—influential statisticians, editors, textbook authors, and teachers. However, he also has been weakened by the logical critiques of generations of unsuccessful matadors and their assistants. For a swift and clean coup de grâce all that is required now is for that small portion of the scientific community who write the "instructions to authors" for journals to educate itself and translate those logical critiques into some gentle new prescriptions.

3. Some History, Briefly

The birth and career of "statistically significant" and his cousins ("significantly different," "nonsignificant," "critical *p*-values," "fixed alpha," "$p < \alpha$," etc.) have been reviewed, with greater or

CONTACT Stuart H. Hurlbert hurlbert@sdsu.edu Department of Biology Emeritus, San Diego State University, San Diego, CA 92812.
Supplementary materials for this article are available online. Please go to www.tanfonline.com/r/TAS.

lesser degrees of accuracy and clarity, in dozens if not hundreds of articles in the published literature and on social media. So we know in particular (1) how Fisher codified earlier ideas about tests of significance based on standardized critical p-values, (2) how Neyman and Pearson developed a superficially similar decision theoretic framework based on critical alphas or Type I error rates, (3) how others, especially publishers and early textbook writers, created awkward hybrids of those paradigms, and (4) how almost the whole statistical world "went dichotomous" and helped make "statistically significant" the *toro bravo* that he remains to this day. The 38-page review of Hurlbert and Lombardi (2009) describes that history in detail, also discussing *in extenso* several closely related issues and controversies beyond the scope of this note.

All the sound and fury, all the resulting confusion, errors, bad teaching, and poor advice from reviewers and editors have taken a large toll. Their effects are impossible to estimate. But many writers, starting more than half a century ago, have provided strong and widely accepted arguments for concluding that it is time to abandon "statistically significant" and dichotomization of the p-scale.

Perhaps we needed that dichotomy for a time. The proposition of concrete evidentiary standards such as critical p-values or fixed alphas (of, say, 0.05), combined with a powerful label ("statistically significant"), may have been a critical and "brilliant stroke of simplification that opened the arcane domain of statistical calculation to a world of experimenters and research workers" who were confronting the new statistical methodologies with some trepidation (Stigler 2008; Hurlbert and Lombardi 2009). Neophytes at the beginning of the 20th century likely were neither amenable to nor capable of nuanced interpretations of new statistical constructs. Nor did they have the computational power to churn out exact p-values or tables comprehensive enough to allow their estimation by interpolation.

But now statistical neophytes are a smaller portion of the scientific community, this and related issues have been thrashed out in the literature for half a century, and simple reason and logic should have a better chance of carrying the day.

4. Solution: A Proscription and a Prescription

For the narrowly defined problem at hand, we propose a two-part solution. It *could* be implemented quickly. It will not *require* any immediate large increase in the statistical sophistication of scientists collectively; but it will *promote* just such an increase over time.

Brian Caffo (pers. comm.) has suggested to us that our two propositions might be regarded as calls for community policing and editorial policing, respectively.

First, we propose that in research articles all use of the phrase "statistically significant" and closely related terms ("nonsignificant," "significant at $p = 0.xxx$," "marginally significant," etc.) be disallowed on the solid grounds long existing in the literature. Just present the p-values without labeling or categorizing them. Every professional statistician, every scientist who uses statistics, and every statistics instructor is in a position to help build momentum for this improvement and proscribe "statistically significant" in those situations where they have individual decision-making authority.

Second, we propose that direct formal requests be made to the editors and editorial boards of journals to modify their instructions to authors to include a disallowance of manuscripts that do not adhere to the above proscription. This task *might* be accomplished under the aegis of ASA, the Royal Statistical Society and perhaps other statistical societies, with a statement that is much simpler and briefer than the 2016 ASA statement, but that has a much larger number of endorsers. This process would build on the sendout of the 2016 statement recommended above.

For a journal an additional "instruction to authors" could read something like the following:

> *There is now wide agreement among many statisticians who have studied the issue that for reporting of statistical tests yielding p-values it is illogical and inappropriate to dichotomize the p-scale and describe results as "significant" and "nonsignificant." Authors are strongly discouraged from continuing this never justified practice that originated from confusions in the early history of modern statistics.*

These recommendations should be well received. Occasionally in the past, strong statements have been put out unilaterally by a few individual editors or editorial boards of journals as to what statistical procedures must or must not be used. In contrast, an instruction like that suggested above neither proscribes nor requires use of particular methodologies. It only stipulates a simple matter of language.

A community grassroots effort might advance implementation of both propositions. A large contingent of reputable statisticians and other scientists endorsing them could approach the editors and editorial boards of journals directly urging them to develop new guidelines internally via their own established procedures for doing so. That process by itself might cause some individual editors and editorial board members to join the "community policing" movement, even if their journals or societies decline to formally modify their "instructions to authors."

Curious about how much support our propositions would have, we decided to test reaction to them while this article was under review. We forwarded the manuscript to about 100 scientists who have authored statistics textbooks or reference books, who have served as lead editors for journals (statistical or otherwise), who have served as society presidents, or who have published significant critiques of statistical practice. We asked each whether they were willing to publicly endorse the first sentences of the "proscription" and "prescription" above. We were pleasantly surprised to receive endorsements from 47 scientists (in addition to ourselves), from 10 countries and a wide variety of disciplines (e.g., medicine, psychology, sociology, economics, environmental sciences, etc.). Several endorsers also prompted useful changes in the manuscript itself. The list of endorsers is given in supplementary file Appendix A.

Let us consider briefly some objections to our two propositions that might be raised.

The text under principle #3 in the 2016 ASA statement includes, "Pragmatic considerations often require binary, 'yes–no' decisions but this does not mean that p-values alone can ensure that a decision is correct or incorrect." We would say that the "often" is unwarranted and that situations requiring

binary decisions solely on the basis of individual *p*-values are vanishingly rare in both basic and applied research. And even where some sort of concrete or physical action is to be taken, the terms "statistically significant" and "nonsignificant" will remain unneeded and actions taken by any decision-makers will not be determined solely by the results of a single test or *p*-value any more than they are now.

Others may be concerned about how we can justify and determine or fix set-wise or family-wise Type I error rates when multiple tests or comparisons are being conducted if we abandon critical *p*-values and fixed α's for individual tests. The short and happy answer is: "You can't. And shouldn't try!" Backed up by a 20-page review of the topic, Hurlbert and Lombardi (2012) recommended: "Whatever statistical tests are dictated by the objectives and design of a study are best carried out one-by-one without any adjustments for multiplicities, whether the latter derive from there being multiple treatments, multiple monitoring dates or multiple response variables. Clarity and interpretability of results will be favoured." Despite the obsession over set-wise Type I error rates in certain quarters, for example, some statisticians advising clinical trials, other statisticians have been pointing out their ponderous and arbitrary nature for half a century (e.g., Wilson 1962; Cox 1965; J. A. Nelder [in O'Neill and Wetherill 1971]; Carmer and Walker 1982; Perry 1986; Finney 1988; Mead 1988; Rothman 1990; Keppel 1991; Pearce 1993; Stewart-Oaten 1995; Nakagawa 2004; Schulz and Grimes 2005).

It would seem that any real problems created by our propositions would be so rare or unique as to not constitute any counter argument of weight. Benefits should surely outweigh any potential cost. And authors are always free to make "special-case" pleas to editors.

The statistical use of the word "significant" will never be understood by those without statistical training to mean anything less than its synonyms, "important" and "influential." According to the Merriam Webster dictionary (*https://www.merriam-webster.com/dictionary/significant*), the English word dates back to at least 1579, and it is among the 10% most frequently used words in the English language. So it is time for the discipline of statistics to give it up and return it to its natural roots. Even if statisticians and scientists understand that the statistical meaning of significance differs from its synonyms, scientific findings often are presented to the public by the media. Consumers of them almost surely are misled by equating "significant" with its synonyms, "important," and "influential," even when a journalist is careful to say "a small but significant effect…." In data analysis contexts use of the term should be minimized lest scientist readers automatically infer that it refers to *p*-values lower than some alpha.

5. Earlier Matadors

Many matadors over many decades have pointed out the desirability of doing away with critical *p*-values and fixed alphas and the attendant dichotomized terminology. Key excerpts from a few of them will show that our proposals are not revolutionary but are in fact quite moderate and easily defended ones. Here are nine sets of authors, selected in part for the explicitness with which they advocate abandoning the terms "significant" and "nonsignificant" and the dichotomized thinking they reflect. Historical readings must allow for the fact that continuous *p*-values were traditionally called "significance levels" by Fisher and those he instructed (like D. R. Cox), yet Neyman and his successors adopted the same term for the fixed alpha level against which a calculated *p*-value would be compared.

> "It is customary to take arbitrary *p* values, such as .05 and .01 and use them to dichotomize this continuum into a *significant* and an *insignificant* portion. This habit has no obvious advantage, if what is intended is merely a restatement of the probability values these are already given in any case and are far more precise than a simple dichotomous statement. … If the verbal dichotomous scale is not satisfactory—as it clearly is not—the answer surely is to keep to the continuous *p* scale, rather than subdivide the verbal scale." (Eysenck 1960)

> "Tradition notwithstanding, there seems to be little justifiable reason [for dichotomizing our interpretation of the *P* scale, so scientists should]… do away with arbitrary levels of significance, and the calling of one test result 'significant' and another 'not significant'." (Skipper, Guenther, and Nass 1967)

> "In this note we confine ourselves to one of the simplest models, which already calls for the exercise of judgment in ways which will not appeal to those anxious to reduce the practice of statistics to the mechanical application of mathematical rules. … The statistician analyzing results of this kind should evaluate the *P* value. It is not for him, nor for the individual experimenter, alone to impose an α value on other persons. … The use of fixed significance levels, $\alpha = 0.05$ or 0.01, was introduced by Fisher for largely accidental reasons connected with Pearson's copyright in the tables of X^2. Although it has many advantages, especially the dangerously seductive one of saving us the effort of thinking, for the reasons indicated above and below we now ought to abandon it." (Barnard 1982).

> "It is ridiculous to interpret the results of a study differently according to whether the *P* value obtained was, say, 0.055 or 0.045. These *P* values should lead to very similar conclusions, not diametrically opposed ones… In recent years there has been a welcome move away from regarding the *P* value as significant or not significant, according to which side of the arbitrary 0.05 value it is, towards quoting the actual *P* value… Forcing a choice between significant and non-significant obscures the uncertainty present whenever we draw inferences from a sample." (Altman 1991)

> "The decision-theoretic approach to hypothesis testing suggested by Neyman and Pearson is disappearing from use in major medical journals, and the practice of dividing results of hypothesis tests into 'significant' and 'non-significant' is outdated and unhelpful. … It used to be the convention to say that there was 'significant' evidence against the null hypothesis if $P < 0.05$, and to categorize results as 'significant' or 'non-significant'. This is outdated: it is much better to report the precise *P*-value." (Sterne 2002)

> "In Fisherian testing, the *p* value is actually a more fundamental concept than the α level. … A reasonable view would be that an α level should never be chosen; that a scientist should simply evaluate the evidence embodied in the *p* value." (Christensen 2005)

> "We will advocate discarding this [Neyman-Pearsonian] framework for most significance testing situations and replacing it with an explicitly neoFisherian one that 1) does not fix α, 2) does not describe P values as 'significant' or 'non-significant,' 3) does not accept null hypotheses on the basis of high P values but only suspends judgment, 4) recognizes the obvious, near universal need to present effect size information in conjunction with significance tests, and 5) acknowledges the frequent utility of confidence intervals (and other adjunct statistics helpful to interpretation) as well as the fact that they are often unneeded. …. From discussion of this issue with other scientists, it seems the biggest psychological impediment to the acceptance of the neoFisherian paradigm is a reluctance to throw out that deceptive crutch, the phrase 'statistically significant.' As Stoehr (1999) points out, we all would like 'a quick, objective and automatic way' to evaluate our results, but there is none that also meets the additional requirements of 'logical' and 'useful.' We must simply apply the same sorts of nuanced thinking and nuanced language we use in other contexts involving gradations in strength of evidence." (Hurlbert and Lombardi 2009).

> "Thus, it is a mark of good practice to present the P-value itself rather than to report whether or not the result was statistically significant. Because significance and nonsignificance are simply degraded descriptions of a P-value, they do not have most of the meanings implied or ascribed to them by some experts." (Greenland and Poole 2011)

> "Never, ever, use the word 'significant' in a paper. It is arbitrary, and, as we have seen, deeply misleading. Still less should you use 'almost significant', 'tendency to significant' or any of the hundreds of similar circumlocutions listed by Matthew Hankins [2013] on his *Still not Significant* blog." (Colquhoun 2014)

Many more recent papers have been published that strongly support these authors and the neoFisherian paradigm generally and warn of other problems in interpreting statistical analyses. Fortunately, a recent authoritative, comprehensive and detailed review by Greenland et al. (2016) makes it easy for readers to catch up with the best thinking on these topics. It is important reading for any person drawing statistical inferences or trying to understand the literature.

Given the illustrious string of authorities who have taken to task "statistically significant" and the distortions it entails, why does use of the term persist? Indeed, why did the ASA statement (Wasserstein and Lazar 2016) not recommend the phrase be abandoned? Some authors ascribe the problem to human cognitive biases that have been created or aggravated by elementary statistical training (McShane and Gal 2017; Greenland 2017). More prosaically, few statisticians have a deep knowledge of the historical literature of statistics or realize how much of the most cogent criticism of statistical practice is found in the journals of other fields they rarely consult. And few have time to seek it. Hurlbert and Lombardi (2009) thoroughly reviewed the history of "statistically significant," explicitly recommended its abandonment, and anticipated virtually all the conclusions of the ASA statement. The drafters of the ASA statement did not cite that paper perhaps feeling its blunt recommendations were too radical. But readers who consult it will find a wealth of information that supports and expands the conclusions in the ASA statement, including additional justification for the recommendations in this article.

6. Nuanced Reporting

How then does a mere researcher account for all these lamentations in practice? How should we write up analyses if we can no longer say what effects or correlations are "statistically significant"? What do Hurlbert and Lombardi (2009) mean by "nuanced thinking and nuanced language"? A new set of terminological baggage? Heavens no.

We continue to recommend the use of p-values, confidence intervals, Bayes factors and all of the other tools available. We simply wish to remove the dichotomous use of the term "significant" as an accompaniment to them. Two examples will help illustrate how well this can work.

Kristen Reifel and her colleagues provided one good example with their paper on the spatial distribution of plankton in the Salton Sea, California's largest lake (Reifel et al. 2007). The results (p-values, coefficients) of many regression analyses are fully incorporated into Table 2 and Figure 6 of the paper. There is no verbal characterization of any p-value, nor mention of a p-value in the text. The text includes only discussions of apparent trends or effects. Readers (including editors and reviewers) are expected to look at the graphs, p-values, and coefficients and make their own judgments about whether the authors over-interpreted their results, under-interpreted them, or got it just right.

In a quite different type of study, Pocock et al. (2016) reported how results of a clinical trial on the effects of renal denervation on patients with high blood pressure were clarified by use of ANCOVA. In their tables they report mean effect sizes, and p-values, 95% confidence intervals and standard errors for them with no recourse, with one minor exception, in the paper to labels such as "significant" or "nonsignificant."

We encourage the reader to download copies of these papers and confirm that it really is that simple. A few quotes from these papers will not suffice: it is examination of the entirety of their results and discussion sections that will confirm our claim most clearly. Studies conducting statistical analyses are too diverse in type, size, scope, objective, diversity of statistical procedures used, and other summary statistics accompanying p-values for generally applicable *specific* guidelines to be possible. One exception concerns the content of abstracts: in reporting the main subject matter findings, emphasis should always be on effect sizes, broadly defined. Especially in the past, many journals have been willing to accept abstracts, and even entire manuscripts, where no explicit information on effect sizes was given so long as sufficiently low p-values were cited.

Abstracts for the two articles cited above show the wide range of possibilities for focusing on effect sizes. In their abstract, Reifel et al. (2007) state, very partially, "Several diatom species increased up to 800-fold in abundance by ca. 20 km downcurrent from inflow points in September. …. Zooplankton abundances did not show regular trends downcurrent of river inflows

except for the larvae of [the barnacle], which increased in density ca. 100-fold." That was sufficient for an abstract; readers can consult the body of the paper for the statistical methods, graphs, precise effect size estimates and *p*-values forming the basis for those conclusions.

In their abstract, Pocock et al. (2016) summarize the key finding in this more precise way: "Analysis of covariance was performed on the 6-month change in systolic blood pressure, estimating a mean treatment difference of −4.11 mm Hg (95% confidence interval: −8.44 to 0.22 mm Hg; $p = 0.064$), which was similar to the unadjusted difference but with a smaller confidence interval." More precise information is given than Reifel et al. (2007) provided, but that is feasible in the abstract only because Pocock et al. were dealing with a single dependent variable. Reifel et al. were dealing with about thirty on each of two sampling dates.

Few actions would more encourage authors to give primary emphasis to effect sizes in their abstracts than would editors disallowing the use of "statistically significant." Especially for large studies involving many different statistical analyses, decisions as to which effect sizes are sufficiently important and conclusively demonstrated as to merit mention in an abstract will necessarily be subjective ones, as will be decisions about how to describe and discuss results in the body of a paper.

7. Conclusion

Just as the initiators of the 2016 ASA statement achieved success by keeping the objective narrow and the focus sharp, so we believe that the next most feasible and concrete step is to implement the two steps outlined.

What could be simpler or more productive than persuading statisticians and other scholars to analyze and write up their data while forgoing "statistically significant" and related terms? Many benefits will flow from this practice, as so many other statistical controversies have been driven by the ancient dichotomy (e.g., misuse of one-tailed tests: Lombardi and Hurlbert 2009; irrational advocacy of set-wise Type I error rates: Hurlbert and Lombardi 2012). And how much work would it be to start getting editorial boards "with the program?" All of us can only try and see.

We are heartened that the three editors of this major special issue of The American Statistician, after thrashing all sorts of matters out with 45 sets of authors, have come out in full support of our thesis. Their introductory editorial states: "The ASA Statement on *P*-values and Statistical Significance stopped just short of recommending that declarations of 'statistical significance' be abandoned. We take that step here. We conclude, based on our review of the articles in this special issue and the broader literature, that it is time to stop using the term 'statistically significant' entirely" (Wasserstein et al. 2019).

Supplementary Materials

Appendix A: Statisticians and other scientists endorsing the propositions (1) that in research articles all use of the phrase "statistically significant" and closely related terms ("nonsignificant," "significant at $p = 0.xxx$," "marginally significant," etc.) be disallowed on the solid grounds long existing in the literature; and (2) that direct formal requests be made to the editors and editorial boards of journals to modify their instructions to authors to include a disallowance of manuscripts that do not adhere to the above proscription.

References

Altman, D. G. (1991), *Practical Statistics for Medical Research*, London: Chapman and Hall. [354]

Barnard, G. A. (1982), "Conditionality Versus Similarity in the Analysis of 2 × 2 Tables," in *Statistics and Probability: Essays in Honor of C.R. Rao*, eds. G. Kallianpur, P. R. Krishnaiah, and J. K. Ghosh, New York: North Holland Publishing, pp. 59–65. [354]

Carmer, S. G., and Walker, W. M. (1982), "Baby Bear's Dilemma: A Statistical Tale," *Agronomy Journal*, 74, 122–124. [354]

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59, 121–126. [354]

Colquhoun, D. (2014), "An Investigation of the False Discovery Rate and the Misinterpretation of *p*-Values," *Royal Society Open Science*, 1, 140216. [355]

Cox, D. R. (1965), "A Remark on Multiple Comparison Methods," *Technometrics*, 7, 223–224. [354]

Eysenck, H. J. (1960), "The Concept of Statistical Significance and the Controversy About One-Tailed Tests," *Psychological Review*, 67, 269–271. [354]

Finney, D. J. (1988), "Was This in Your Statistics Textbook? III. Design and Analysis," *Experimental Agriculture*, 24, 421–432. [354]

Greenland, S. (2017), "The Need for Cognitive Science in Methodology," *American Journal of Epidemiology*, 186, 639–645. [355]

Greenland, S., and Poole, C. (2011), "Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony," *Jurimetrics Journal*, 51, 113–129. [355]

Greenland, S., Senn, S. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016), "Statistical Tests, *P* Values, Confidence Intervals, and Power: A Guide to Misinterpretations," *European Journal of Epidemiology*, 31, 337–350. [355]

Hankins, M. C. (2013), "Still not significant," available at *http://mchankins.wordpress.com/2013/04/21/still-notsignificant-2/*. [355]

Hurlbert, S. H., and Lombardi, C. M. (2009), "Final Collapse of the Neyman-Pearson Decision-Theoretic Framework and Rise of the NeoFisherian," *Annales Zoologici Fennici*, 46, 311–349. [353,355]

——— (2012), "Lopsided Reasoning on Lopsided Tests and Multiple Comparisons," *Australian & New Zealand Journal of Statistics*, 54, 23–42. [354,356]

Keppel, G. (1991), *Design and Analysis: A Researcher's Handbook* (3rd ed.), Englewood Cliffs, NJ: Prentice Hall. [354]

Lombardi, C. M., and Hurlbert, S. H. (2009), "Misprescription and Misuse of One-Tailed Tests," *Austral Ecology*, 34, 447–468. [356]

McShane, B. B., and Gal, D. (2017), "Statistical Significance and the Dichotomization of Evidence," *Journal of the American Statistical Association*, 112, 885–908. [355]

Mead, R. (1988), *The Design of Experiments*, Cambridge: Cambridge University Press. [354]

Nakagawa, S. (2004), "A Farewell to Bonferroni: The Problems of Low Statistical Power and Publication Bias," *Behavioral Ecology*, 15, 1044–1045. [354]

O'Neill, R., and Wetherill, G. B. (1971), "The Present State of Multiple Comparison Methods," *Journal of the Royal Statistical Society*, Series B, 33, 218–250. [354]

Pearce, S. C. (1993), "Data Analysis in Agricultural Experimentation. III. Multiple Comparisons," *Experimental Agriculture*, 29, 1–8. [354]

Perry, J. N. (1986), "Multiple-Comparison Procedures: A Dissenting View," *Journal of Economic Entomology*, 79, 1149–1155. [354]

Pocock, S. J., Bakris, G., Rhatt, D. L., Brar, S., Fahy, M., and Gersh, B. J. (2016), "Regression to the Mean in SYMPLICITY HTN-3: Implications for Design and Reporting of Future Trials," *Journal of the American College of Cardiology*, 68, 2016–2025. [355,356]

Reifel, K. M., Trees, C. C., Olivo, E., Swan, B. K., Watts, J. M., and Hurlbert, S. H. (2007), "Influence of River Inflows on Spatial Variation of Phytoplankton Around the Southern End of the Salton Sea, California," *Hydrobiologia*, 576, 167–183. [355,356]

Rothman, K. J. (1990), "No Adjustments Are Needed for Multiple Comparisons," *Epidemiology*, 1, 43–46. [354]
Schulz, K. F., and Grimes, D. A. (2005), "Multiplicity in Randomized Trials I: Endpoints and Treatments," *Lancet*, 365, 1591–1595. [354]
Skipper, K. S. Jr., Guenther, A. L., and Nass, G. (1967), "The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science," *American Sociologist*, 2, 16–18. [354]
Sterne, J. A. C. (2002), "Teaching Hypothesis Tests—Time for Significant Change?," *Statistics in Medicine*, 21, 985–994. [354]
Stewart-Oaten, A. (1995), "Rules and Judgments in Statistics: Three Examples," *Ecology*, 76, 2001–2009. [354]
Stigler, S. (2008), "Fisher and the 5% Level," *Chance*, 21, 12. [353]
Stoehr, A. M. (1999), "Are Significance Thresholds Appropriate for the Study of Animal Behaviour?," *Animal Behaviour*, 57, F22–F25. [355]
Wasserstein, R., and Lazar, N. (2016), "ASA Statement on Statistical Significance and *p*-Values," *The American Statistician*, 70, 131–133. [352,355]
Wasserstein, R., Lazar, N., and Schirm, A. (2019), "Editorial: Moving to a World Beyond $p < 0.05$," *The American Statistician*, 73(1). [356]
Wilson, W. (1962), "A Note on the Inconsistency Inherent in the Necessity to Perform Multiple Comparisons," *Psychological Bulletin*, 59, 296–300. [354]