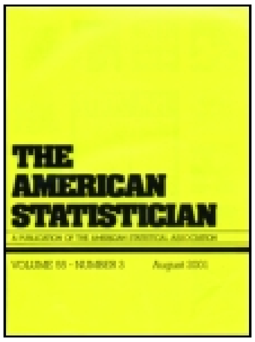

The American Statistician

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

The ASA's statement on p-values: context, process, and purpose

Ronald L. Wasserstein & Nicole A. Lazar

The ASA's statement on p-values: context, process, and purpose

Ronald L. Wasserstein and Nicole A. Lazar

In February, 2014, George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, posed these questions to an ASA discussion forum:

> Q: Why do so many colleges and grad schools teach $p = .05$?
>
> A: Because that's still what the scientific community and journal editors use.
>
> Q: Why do so many people still use $p = 0.05$?
>
> A: Because that's what they were taught in college or grad school.

Cobb's concern was a long-worrisome circularity in the sociology of science based on the use of bright lines such as $P < 0.05$: "We teach it because it's what we do; we do it because it's what we teach." This concern was brought to the attention of the ASA Board.

The ASA Board was also stimulated by highly visible discussions over the last few years. For example, ScienceNews (Siegfried, 2010) wrote: "It's science's dirtiest secret: The 'scientific method' of testing hypotheses by statistical analysis stands on a flimsy foundation." A November, 2013, article in Phys.org Science News Wire (2013) cited "numerous deep flaws" in null hypothesis significance testing. A ScienceNews article (Siegfried, 2014) on February 7, 2014, said "statistical techniques for testing hypotheses…have more flaws than Facebook's privacy policies." A week later, statistician and "Simply Statistics" blogger Jeff Leek responded. "The problem is not that people use P-values poorly," Leek wrote, "it is that the vast majority of data analysis is not performed by people properly trained to perform data analysis" (Leek, 2014).

That same week, statistician and science writer Regina Nuzzo published an article in *Nature* entitled “Scientific method: statistical errors” (Nuzzo, 2014). That article is now one of the most highly viewed *Nature* articles, as reported by altmetric.com (http://www.altmetric.com/details/2115792#score).

Of course, it wasn’t simply a matter of responding to some articles in print. The statistical community has been deeply concerned about issues of *reproducibility* and *replicability* of scientific conclusions. Without getting into definitions and distinctions of these terms, we observe that much confusion and even doubt about the validity of science is arising. Such doubt can lead to radical choices, such as the one taken by the editors of Basic and Applied Social Psychology, who decided to ban p-values (null hypothesis significance testing) (Trafimow and Marks, 2015). Misunderstanding or misuse of statistical inference is only one cause of the “reproducibility crisis” (Peng, 2015), but to our community, it is an important one.

When the ASA Board decided to take up the challenge of developing a policy statement on p-values and statistical significance, it did so recognizing this was not a lightly taken step. The ASA has not previously taken positions on specific matters of statistical practice. The closest the association has come to this is a statement on the use of value-added models (VAM) for educational assessment (Morganstein and Wasserstein, 2014) and a statement on risk-limiting post-election audits (American Statistical Association, 2010). However, these were truly policy-related statements. The VAM statement addressed a key educational policy issue, acknowledging the complexity of the issues involved, citing limitations of VAMs as effective performance models, and urging that they be developed and interpreted with the involvement of statisticians. The statement on election auditing was also in response to a major but specific policy issue (close elections in 2008), and said that statistically-based election audits should become a routine part of election processes.

By contrast, the Board envisioned that the ASA statement on p-values and statistical significance would shed light on an aspect of our field that is too often misunderstood and misused in the broader research community, and, in the process, provide the community a service. The intended audience would be researchers, practitioners and science writers who are not primarily statisticians. Thus, this statement would be quite different from anything previously attempted.

The Board tasked Wasserstein with assembling a group of experts representing a wide variety of points of view. On behalf of the Board, he reached out to over two dozen such people, all of whom said they would be happy to be involved. Several expressed doubt about whether agreement could be reached, but those who did said, in effect, that if there was going to be a discussion, they wanted to be involved.

Over the course of many months, group members discussed what format the statement should take, tried to more concretely visualize the audience for the statement, and began to find points of agreement. That turned out to be relatively easy to do, but it was just as easy to find points of intense disagreement.

The time came for the group to sit down together to hash out these points, and so in October, 2015, twenty members of the group met at the ASA Office in Alexandria, Virginia. The two-day meeting was facilitated by Regina Nuzzo, and by the end of the meeting, a good set of points around which the statement could be built was developed.

The next three months saw multiple drafts of the statement, reviewed by group members, by Board members (in a lengthy discussion at the November 2015 ASA Board meeting), and by members of the target audience. Finally, on January 29, 2016, the Executive Committee of the ASA approved the statement.

The statement development process was lengthier and more controversial than anticipated. For example, there was considerable discussion about how best to address the issue of multiple *potential* comparisons (Gelman and Loken, 2014). We debated at some length the issues behind the words “a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis” (Johnson, 2013). There were differing perspectives about how to characterize various alternatives to the p-value and in how much detail to address them. In order to keep the statement reasonably simple, we did not address alternative hypotheses, error types, or power (among other things), and not everyone agreed with that approach.

As the end of the statement development process neared, Wasserstein contacted Lazar and asked if the policy statement might be appropriate for publication in TAS. After consideration, Lazar decided that TAS would provide a good platform to reach a broad and general statistical readership. Together, we decided that the addition of an online discussion would heighten the interest level for the TAS audience, giving an opportunity to reflect the aforementioned controversy.

To that end, a group of discussants was contacted to provide comments on the statement. You can read their statements in the online supplement. We thank Naomi Altman, Douglas Altman, Daniel J. Benjamin, Yoav Benjamini, Jim Berger, Don Berry, John Carlin, George Cobb,

Andrew Gelman, Steve Goodman, Sander Greenland, John Ioannidis, Joseph Horowitz, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Michele Millar, Charles Poole, Ken Rothman, Stephen Senn, Dalene Stangl, Philip Stark and Steve Ziliak for sharing their insightful perspectives.

Though there was disagreement on exactly what the statement should say, there was high agreement that the ASA should be speaking out about these matters.

Let's be clear. Nothing in the ASA statement is new. Statisticians and others have been sounding the alarm about these matters for decades, to little avail. We hoped that a statement from the world's largest professional association of statisticians would open a fresh discussion and draw renewed and vigorous attention to changing the practice of science with regards to the use of statistical inference.

References:

American Statistical Association (2010), "ASA Statement on Risk-Limiting Post Election Audits," available at http://www.amstat.org/policy/pdfs/Risk-Limiting_Endorsement.pdf

Siegfried, T. (2010), "Odds Are, It's Wrong: Science fails to face the shortcomings of statistics," ScienceNews, 177, 26, available at https://www.sciencenews.org/article/odds-are-its-wrong

Johnson, V.E. (2013), "Uniformly most powerful Bayesian tests," Annals of Statistics, 41, 1716-1741.

Phys.org Science News Wire (2013), “The problem with p values: how significant are they, really?” available at http://phys.org/wire-news/145707973/the-problem-with-p-values-how-significant-are-they-really.html

Gelman, A., and Loken, E. (2014), “The Statistical Crisis in Science [online],” *American Scientist*, 102. Available at http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science

Leek, J. (2014), “On the scalability of statistical procedures: why the p-value bashers just don't get it,” Simply Statistics blog, available at http://simplystatistics.org/2014/02/14/on-the-scalability-of-statistical-procedures-why-the-p-value-bashers-just-dont-get-it/

Nuzzo, R. (2014), “Scientific Method: statistical errors”, *Nature*, 506, 150-152, available at http://www.nature.com/news/scientific-method-statistical-errors-1.14700

Morganstein, D., and Wasserstein, R. (2014), “ASA Statement on Value-Added Models,” Statistics and Public Policy, 1, 108-110, available at http://amstat.tandfonline.com/doi/full/10.1080/2330443X.2014.956906

Siegfried, T. (2014), “To make science better, watch out for statistical flaws,” ScienceNews, available at https://www.sciencenews.org/blog/context/make-science-better-watch-out-statistical-flaws

Peng, R. (2015), “The reproducibility crisis in science: A statistical counterattack,” Significance, 12 (3), 30–32

Trafimow, D., and Marks, M. (2015), editorial in Basic and Applied Social Psychology, 37, 1-2

ASA Statement on Statistical Significance and P-values

February 5, 2016

Edited by Ronald L. Wasserstein, Executive Director

On behalf of the American Statistical Association Board of Directors

Introduction

Increased quantification of scientific research and a proliferation of large, complex datasets in recent years have expanded the scope of applications of statistical methods. This has created new avenues for scientific progress, but it also brings concerns about conclusions drawn from research data. The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly.

Underpinning many published scientific conclusions is the concept of "statistical significance," typically assessed with an index called the p-value. While the p-value can be a useful statistical measure, it is commonly misused and misinterpreted. This has led to some scientific journals discouraging the use of p-values, and some scientists and statisticians recommending their abandonment, with some arguments essentially unchanged since p-values were first introduced.

In this context, the American Statistical Association (ASA) believes that the scientific community could benefit from a formal statement clarifying several widely agreed upon principles underlying the proper use and interpretation of the p-value. The issues touched on here affect not only research, but research funding, journal practices, career advancement, scientific education, public policy, journalism, and law. This statement does not seek to resolve all the issues relating to sound statistical practice, nor to settle foundational controversies. Rather, the statement articulates in non-technical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

What is a p-value?

Informally, a p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.

Principles

1. *P-values can indicate how incompatible the data are with a specified statistical model.*

A p-value provides one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data. The most common context is a model, constructed under a set of assumptions, together with a so-called "null hypothesis." Often the null hypothesis postulates the absence of an effect, such as no difference between two groups, or the absence of a relationship between a factor and an outcome. The smaller the p-value, the greater the statistical incompatibility of the data with the null hypothesis, if the underlying assumptions used to calculate the p-value hold. This incompatibility can be interpreted as casting doubt on or providing evidence against the null hypothesis or the underlying assumptions.

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*.

Researchers often wish to turn a p-value into a statement about the truth of a null hypothesis, or about the probability that random chance produced the observed data. The p-value is neither. It is a statement about data in relation to a specified hypothetical explanation, and is not a statement about the explanation itself.

3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*

Practices that reduce data analysis or scientific inference to mechanical “bright-line” rules (such as “$p < 0.05$”) for justifying scientific claims or conclusions can lead to erroneous beliefs and poor decision-making. A conclusion does not immediately become “true” on one side of the divide and “false” on the other. Researchers should bring many contextual factors into play to derive scientific inferences, including the design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis. 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. The widespread use of “statistical significance” (generally interpreted as “$p \leq 0.05$”) as a license for making a claim of a scientific finding (or implied truth) leads to considerable distortion of the scientific process.

4. *Proper inference requires full reporting and transparency*

 ACCEPTED MANUSCRIPT

P-values and related analyses should not be reported selectively. Conducting multiple analyses of the data and reporting only those with certain p-values (typically those passing a significance threshold) renders the reported p-values essentially uninterpretable. Cherry-picking promising findings, also known by such terms as data dredging, significance chasing, significance questing, selective inference and “p-hacking,” leads to a spurious excess of statistically significant results in the published literature and should be vigorously avoided. One need not formally carry out multiple statistical tests for this problem to arise: Whenever a researcher chooses what to present based on statistical results, valid interpretation of those results is severely compromised if the reader is not informed of the choice and its basis. Researchers should disclose the number of hypotheses explored during the study, all data collection decisions, all statistical analyses conducted and all p-values computed. Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted, and how those analyses (including p-values) were selected for reporting.

5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*

Statistical significance is not equivalent to scientific, human, or economic significance. Smaller p-values do not necessarily imply the presence of larger or more important effects, and larger p-values do not imply a lack of importance or even lack of effect. Any effect, no matter how tiny, can produce a small p-value if the sample size or measurement precision is high enough, and large effects may produce unimpressive p-values if the sample size is small or measurements are imprecise. Similarly, identical estimated effects will have different p-values if the precision of the estimates differs.

6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

Researchers should recognize that a p-value without context or other evidence provides limited information. For example, a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis. Likewise, a relatively large p-value does not imply evidence in favor of the null hypothesis; many other hypotheses may be equally or more consistent with the observed data. For these reasons, data analysis should not end with the calculation of a p-value when other approaches are appropriate and feasible.

Other approaches

In view of the prevalent misuses of and misconceptions concerning p-values, some statisticians prefer to supplement or even replace p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence, credibility, or prediction intervals; Bayesian methods; alternative measures of evidence, such as likelihood ratios or Bayes Factors; and other approaches such as decision-theoretic modeling and false discovery rates. All these measures and approaches rely on further assumptions, but they may more directly address the size of an effect (and its associated uncertainty) or whether the hypothesis is correct.

Conclusion

Good statistical practice, as an essential component of good scientific practice, emphasizes principles of good study design and conduct, a variety of numerical and graphical summaries of data, understanding of the phenomenon under study, interpretation of results in context, complete reporting and proper logical and quantitative understanding of what data summaries mean. No single index should substitute for scientific reasoning.

Acknowledgment: The ASA Board of Directors thanks the following people for sharing their expertise and perspectives during the development of the statement. The statement does not necessarily reflect the viewpoint of all these people, and in fact some have views that are in opposition to all or part of the statement. Nonetheless, we are deeply grateful for their contributions. *Naomi Altman, Jim Berger, Yoav Benjamini, Don Berry, Brad Carlin, John Carlin, George Cobb, Marie Davidian, Steve Fienberg, Andrew Gelman, Steve Goodman, Sander Greenland, Guido Imbens, John Ioannidis, Valen Johnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, Regina Nuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, Dalene Stangl, Philip Stark, Steve Ziliak.*

.

A Brief P-Values and Statistical Significance Reference List

To Accompany the *ASA Statement on P-Values and Statistical Significance*

The list below is not comprehensive, but provides a good starting point for individuals who would like to explore in greater detail the issues raised in the *ASA Statement on P-Values and Statistical Significance*. The items are listed alphabetically.

Altman D.G., Bland J.M. (1995), "Absence of evidence is not evidence of absence," *British Medical Journal*, 311:485.

Altman, D.G., Machin, D., Bryant, T.N., and Gardner, M.J., eds. (2000), Statistics with Confidence, 2nd ed., London: BMJ Books.

Berger, J.O., and Delampady, M. (1987), "Testing precise hypotheses," *Statistical Science*, 2, 317–335

Berry, D. (2012), "Multiplicities in Cancer Research: Ubiquitous and Necessary Evils," *Journal of the National Cancer Institute*, 104, 1124–1132

Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59, 2, 121-126

Cox, D.R. (1982), "Statistical Significance Tests," *British Journal of Clinical Pharmacology*, 14, 325-331

Edwards, W., Lindman, H., and Savage, L.J. (1963), "Bayesian statistical inference for psychological research," *Psychological Review*, 70, 193–242.

Gelman, A., and Loken, E. (2014), “The Statistical Crisis in Science [online],” *American Scientist*, 102. Available at http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science

Gelman A, Stern HS. (2006), “The difference between ‘significant’ and ‘not significant’ is not itself statistically significant,” *The American Statistician,* 60:328–331.

Gigerenzer G (2004), “Mindless statistics,” *Journal of Socioeconomics,* 33:567–606.

Goodman, S.N. (1999a), “Toward Evidence-Based Medical Statistics 1: The P Value Fallacy,” *Annals of Internal Medicine*, 130, 995-1004.

____________ (1999b), “Toward Evidence-Based Medical Statistics. 2: The Bayes Factor,” *Annals of Internal Medicine*, 130, 1005-1013.

____________ (2008), “A Dirty Dozen: Twelve P-Value Misconceptions,” *Seminars in Hematology*, 45, 135-140.

Greenland, S. (2011), “Null misinterpretation in statistical testing and its impact on health risk assessment,” *Preventive Medicine*, 53, 225–228.

___________ (2012). Nonsignificance plus high power does not imply support for the null over the alternative. *Annals of Epidemiology,* 22:364–368.

Greenland, S., and Poole C (2011), “Problems in common interpretations of statistics in scientific articles, expert reports, and testimony,” *Jurimetrics*, 51, 113–129.

Hoenig J.M., and Heisey D.M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician,* 55:19–24.

Ioannidis, J.P. (2005), "Contradicted and initially stronger effects in highly cited clinical research." *Journal of the American Medical Association*, 294, 218-228.

___________ (2008), "Why most discovered true associations are inflated (with discussion)," *Epidemiology* 19: 640-658.

Johnson, V.E. (2013), "Revised standards for statistical evidence," *Proceedings of the National Academy of Scienc*es, 110(48), 19313–19317.

__________ (2013), "Uniformly most powerful Bayesian tests," Annals of Statistics, 41, 1716-1741.

Lang, J., Rothman K.J., and Cann, C.I. (1998), "That confounded P-value. (Editorial)," *Epidemiology*, 9, 7-8.

Lavine, M. (1999), "What is Bayesian Statistics and Why Everything Else is Wrong," *UMAP Journal*, 20:2

Lew, M.J. (2012), "Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P," *British Journal of Pharmacology*, 166:5, 1559-1567.

Phillips, C.V. (2004), "Publication bias in situ," *BMC Medical Research Methodology*, 4:20.

Poole C. (1987), "Beyond the confidence interval," *American Journal of Public Health*, 77, 195–199.

Poole, C. (2001). Low P-values or narrow confidence intervals: Which are more durable? *Epidemiology*, 12, 291–294.

 ACCEPTED MANUSCRIPT

Rothman, K.J. (1978), "A show of confidence (Editorial)," *New England Journal of Medicine*, 299, 1362-1363.

___________ (1986), "Significance questing (Editorial)," *Annals of Internal Medicine*, 105, 445-447.

___________ (2010), "Curbing type I and type II errors," *European Journal of Epidemiology*, 25, 223-224.

Rothman, K.J., Weiss, N.S., Robins, J., Neutra, R., and Stellman, S. (1992), "Amicus Curiae brief for the U. S. Supreme Court, Daubert v. Merrell Dow Pharmaceuticals, Petition for Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit," No. 92-102, October Term, 1992

Rozeboom, W.M. (1960), "The fallacy of the null-hypothesis significance test," *Psychological Bulletin*, 57:416–428.

Schervish, M.J. (1996), "P Values: What They Are and What They Are Not," *The American Statistician*, 50:3, 203-206

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant," *Psychological Science*, 22(11), 1359-1366.

Stang, A., and Rothman, K.J. (2011), "That confounded P-value revisited," *Journal of Clinical Epidemiology*, 64(9), 1047-1048

Stang, A., Poole, C., and Kuss, O. (2010), "The ongoing tyranny of statistical significance testing in biomedical research," *European Journal of Epidemiology*, 25(4), 225-30.

Sterne, J. A. C. (2002). "Teaching hypothesis tests – time for significant change?" *Statistics in Medicine*, 21, 985-994.

Sterne, J. A. C. and G. D. Smith (2001). "Sifting the evidence – what's wrong with significance tests?" *British Medical Journal*, 322, 226-231.

Ziliak, S.T. (2010), "The Validus Medicus and a New Gold Standard," *The Lancet,* 376, 9738, 324-325.

Ziliak, S.T., and McCloskey, D.N. (2008), The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives, Ann Arbor: University of Michigan Press