ASA News
AMERICAN STATISTICAL ASSOCIATION
Promoting the Practice and Profession of Statistics®
732 North Washington Street, Alexandria, VA 22314 • (703) 684-1221 • Toll Free: (888) 231-3473 • www.amstat.org • www.twitter.com/AmstatNews

AMERICAN STATISTICAL ASSOCIATION RELEASES STATEMENT ON STATISTICAL SIGNIFICANCE AND *P*-VALUES

Provides Principles to Improve the Conduct and Interpretation of Quantitative Science

March 7, 2016

The American Statistical Association (ASA) has released a "Statement on Statistical Significance and *P*-Values" with six principles underlying the proper use and interpretation of the *p*-value [http://amstat.tandfonline.com/doi/abs/10.1080/00031305.2016.1154108#.Vt2XIOaE2MN]. The ASA releases this guidance on *p*-values to improve the conduct and interpretation of quantitative science and inform the growing emphasis on reproducibility of science research. The statement also notes that the increased quantification of scientific research and a proliferation of large, complex data sets has expanded the scope for statistics and the importance of appropriately chosen techniques, properly conducted analyses, and correct interpretation.

Good statistical practice is an essential component of good scientific practice, the statement observes, and such 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."

"The *p*-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, the ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post *p*<0.05 era.'"

"Over time it appears the *p*-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the 'file-drawer effect,' in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as '*p*-hacking' and 'data dredging' that emphasize the search for small *p*-values over other statistical and scientific reasoning."

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

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.*
5. *A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.*
6. *By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.*

The statement has short paragraphs elaborating on each principle.

In light of misuses of and misconceptions concerning *p*-values, the statement notes that statisticians often 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."

"The contents of the ASA statement and the reasoning behind it are not new—statisticians and other scientists have been writing on the topic for decades," Utts said. "But this is the first time that the community of statisticians, as represented by the ASA Board of Directors, has issued a statement to address these issues."

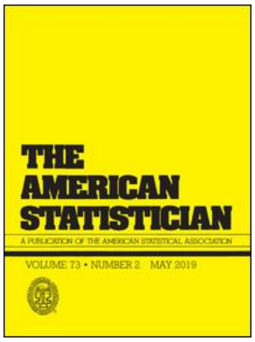
"The issues involved in statistical inference are difficult because inference itself is challenging," Wasserstein said. He noted that more than a dozen discussion papers are being published in the ASA journal *The American Statistician* with the statement to provide more perspective on this broad and complex topic. "What we hope will follow is a broad discussion across the scientific community that leads to a more nuanced approach to interpreting, communicating, and using the results of statistical methods in research."

About the American Statistical Association

The ASA is the world's largest community of statisticians and the oldest continuously operating professional science society in the United States. Its members serve in industry, government and academia in more than 90 countries, advancing research and promoting sound statistical practice to inform public policy and improve human welfare. For additional information, please visit the ASA website at www.amstat.org.

For more information:

Ron Wasserstein
(703) 302-1859
ron@amstat.org



THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 73 • NUMBER 2 MAY 2019

The American Statistician

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

The ASA Statement on *p*-Values: Context, Process, and Purpose

Ronald L. Wasserstein & Nicole A. Lazar

To cite this article: Ronald L. Wasserstein & Nicole A. Lazar (2016) The ASA Statement on *p*-Values: Context, Process, and Purpose, The American Statistician, 70:2, 129-133, DOI: 10.1080/00031305.2016.1154108

To link to this article: https://doi.org/10.1080/00031305.2016.1154108

View supplementary material

Published online: 09 Jun 2016.

Submit your article to this journal

Article views: 552097

View related articles

View Crossmark data

Citing articles: 1249 View citing articles

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=utas20

THE AMERICAN STATISTICIAN
2016, VOL. 70, NO. 2, 129–133
http://dx.doi.org/10.1080/00031305.2016.1154108

Taylor & Francis
Taylor & Francis Group

EDITORIAL

The ASA's Statement on *p*-Values: Context, Process, and Purpose

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 $p = 0.05$?
> 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 was not 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, provides 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 more than 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, 20 members of the group met at the ASA Office in Alexandria, Virginia. The 2-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 3 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

© 2016 American Statistical Association

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. 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 *The American Statistician* (*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, and a guide to those statements appears at the end of this editorial. 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.

Of special note is the following article, which is a significant contribution to the literature about p-values and statistical significance.

> Greenland, S., Senn, S.J., Rothman, K.J., Carlin, J.B., Poole, C., Goodman, S.N. and Altman, D.G.: "Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations."

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 us 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.

Guide to the Online Supplemental Material to the ASA Statement on *P*-Values and Statistical Significance

Many of the participants in the development of the ASA statement contributed commentary about the statement or matters related to it. Their comments are posted as online supplements to the statement. We provide here a list of the supplemental articles.

Supplemental Material to the ASA Statement on *P*-Values and Statistical Significance

- *Altman, Naomi*: Ideas from multiple testing of high dimensional data provide insights about reproducibility and false discovery rates of hypothesis supported by p-values
- *Benjamin, Daniel J, and Berger, James O:* A simple alternative to p-values
- *Benjamini, Yoav:* It's not the p-values' fault
- *Berry, Donald A:* P-values are not what they're cracked up to be
- *Carlin, John B:* Comment: Is reform possible without a paradigm shift?
- *Cobb, George:* ASA statement on p-values: Two consequences we can hope for
- *Gelman, Andrew:* The problems with p-values are not just with p-values
- *Goodman, Steven N:* The next questions: Who, what, when, where, and why?
- *Greenland, Sander:* The ASA guidelines and null bias in current teaching and practice
- *Ioannidis, John P.A.:* Fit-for-purpose inferential methods: abandoning/changing P-values versus abandoning/changing research
- *Johnson, Valen E.:* Comments on the "ASA Statement on Statistical Significance and P-values" and marginally significant p-values
- *Lavine, Michael, and Horowitz, Joseph:* Comment
- *Lew, Michael J:* Three inferential questions, two types of P-value
- *Little, Roderick J:* Discussion
- *Mayo, Deborah G:* Don't throw out the error control baby with the bad statistics bathwater
- *Millar, Michele:* ASA statement on p-values: some implications for education
- *Rothman, Kenneth J:* Disengaging from statistical significance
- *Senn, Stephen:* Are P-Values the Problem?
- *Stangl, Dalene:* Comment
- *Stark, P.B.:* The value of p-values
- *Ziliak, Stephen T:* The significance of the ASA statement on statistical significance and p-values

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*. [129]

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science [online]," *American Scientist*, 102. Available at *http://www.american-scientist.org/issues/feature/2014/6/the-statistical-crisis-in-science*. [129]

Johnson, V. E. (2013), "Uniformly Most Powerful Bayesian Tests," *Annals of Statistics*, 41, 1716–1741. [130]

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/*. [129]

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*. [129]

Nuzzo, R. (2014), "Scientific Method: Statistical Errors," *Nature*, 506, 150–152. Available at *http://www.nature.com/news/scientific-method-statistical-errors-1.14700*. [129]

Peng, R. (2015), "The Reproducibility Crisis in Science: A Statistical Counterattack," *Significance*, 12, 30–32. [129]

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. [129]

Siegfried, T. (2010), "Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics," *Science News*, 177, 26. Available at *https://www.sciencenews.org/article/odds-are-its-wrong*. [129]

Siegfried, T. (2014), "To Make Science Better, Watch out for Statistical Flaws," *Science News Context Blog*, February 7, 2014. Available at *https://www.sciencenews.org/blog/context/make-science-better-watch-out-statistical-flaws*. [129]

Trafimow, D., and Marks, M. (2015), "Editorial," *Basic and Applied Social Psychology* 37, 1–2. [129]

Ronald L. Wasserstein and Nicole A. Lazar
✉ *ron@amstat.org*
American Statistical Association, 732 North Washington Street,
Alexandria, VA 22314-1943.

ASA Statement on Statistical Significance and *P*-Values

1. 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 nontechnical terms a few select principles that could improve the conduct or interpretation of quantitative science, according to widespread consensus in the statistical community.

2. What is a *p*-Value?

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

3. 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**

   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.

4. 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.

5. 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.

Acknowledgments

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.

Edited by Ronald L.Wasserstein, Executive Director
On behalf of the American Statistical Association
Board of Directors

A Brief *p*-Values and Statistical Significance Reference List

Altman D.G., and 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, 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., and Stern, H.S. (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 Science*s, 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.

——— (2001), “Low *P*-values or Narrow Confidence Intervals: Which are More Durable?” *Epidemiology*, 12, 291–294.

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, 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, 1359–1366.

Stang, A., and Rothman, K.J. (2011), “That Confounded *P*-value Revisited,” *Journal of Clinical Epidemiology*, 64, 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, 225–230.

Sterne, J. A. C. (2002). “Teaching Hypothesis Tests—Time for Significant Change?” *Statistics in Medicine*, 21, 985–994.

Sterne, J. A. C., and Smith, G. D. (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, MI: University of Michigan Press.

Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations

Sander GREENLAND, Stephen J. SENN, Kenneth J. ROTHMAN, John B. CARLIN, Charles POOLE, Steven N. GOODMAN, and Douglas G. ALTMAN

Misinterpretation and abuse of statistical tests, confidence intervals, and statistical power have been decried for decades, yet remain rampant. A key problem is that there are no interpretations of these concepts that are at once simple, intuitive, correct, and foolproof. Instead, correct use and interpretation of these statistics requires an attention to detail which seems to tax the patience of working scientists. This high cognitive demand has led to an epidemic of shortcut definitions and interpretations that are simply wrong, sometimes disastrously so—and yet these misinterpretations dominate much of the scientific literature.

In light of this problem, we provide definitions and a discussion of basic statistics that are more general and critical than typically found in traditional introductory expositions. Our goal is to provide a resource for instructors, researchers, and consumers of statistics whose knowledge of statistical theory and technique may be limited but who wish to avoid and spot misinterpretations. We emphasize how violation of often unstated analysis protocols (such as selecting analyses for presentation based on the P-values they produce) can lead to small P-values even if the declared test hypothesis is correct, and can lead to large P-values even if that hypothesis is incorrect. We then provide an explanatory list of 25 misinterpretations of P-values, confidence intervals, and power. We conclude with guidelines for improving statistical interpretation and reporting.

KEY WORDS: Confidence intervals; Hypothesis testing; Null testing; P-value; Power; Significance tests; Statistical testing.

Online supplement to the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Sander Greenland (corresponding author), Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA (Email: lesdomes@ucla.edu). Stephen J. Senn, Competence Center for Methodology and Statistics, Luxembourg Institute of Health, Luxembourg (Email: stephen.senn@lih.lu). Kenneth J. Rothman, RTI Health Solutions, Research Triangle Institute, Research Triangle Park, NC. John B. Carlin, Clinical Epidemiology and Biostatistics Unit, Murdoch Children's Research Institute, School of Population Health, University of Melbourne, Victoria, Australia (Email: john.carlin@mcri.edu.au). Charles Poole, Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC (Email: cpoole@unc.edu). Steven N. Goodman, Meta-Research Innovation Center, Departments of Medicine and of Health Research and Policy, Stanford University School of Medicine, Stanford, CA (Email: steve.goodman@stanford.edu). Douglas G. Altman, Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology & Musculoskeletal Sciences, University of Oxford, Oxford, United Kingdom (Email: doug.altman@csm.ox.ac.uk). SJS receives funding from the IDEAL project supported by the European Union's Seventh Framework Programme for research, technological development and demonstration under Grant Agreement no 602552. We thank Stuart Hurlbert, Deborah Mayo, Keith O'Rourke, and Andreas Stang for helpful comments, and Ron Wasserstein for his invaluable encouragement on this project.

Introduction

Misinterpretation and abuse of statistical tests has been decried for decades, yet remains so rampant that some scientific journals discourage use of "statistical significance" (classifying results as "significant" or not based on a P-value) (Lang et al. 1998). One journal now bans all statistical tests and mathematically related procedures such as confidence intervals (Trafimow and Marks 2015), which has led to considerable discussion and debate about the merits of such bans (e.g., Ashworth 2015; Flanagan 2015).

Despite such bans, we expect that the statistical methods at issue will be with us for many years to come. We thus think it imperative that basic teaching as well as general understanding of these methods be improved. Toward that end, we attempt to explain the meaning of significance tests, confidence intervals, and statistical power in a more general and critical way than is traditionally done, and then review 25 common misconceptions in light of our explanations. We also discuss a few more subtle but nonetheless pervasive problems, explaining why it is important to examine and synthesize all results relating to a scientific question, rather than focus on individual findings. We further explain why statistical tests should never constitute the sole input to inferences or decisions about associations or effects. Among the many reasons are that, in most scientific settings, the arbitrary classification of results into "significant" and "nonsignificant" is unnecessary for and often damaging to valid interpretation of data; and that estimation of the size of effects and the uncertainty surrounding our estimates will be far more important for scientific inference and sound judgment than any such classification.

More detailed discussion of the general issues can be found in many articles, chapters, and books on statistical methods and their interpretation (e.g., Altman et al. 2000; Atkins and Jarrett 1979; Cox 1977, 1982; Cox and Hinkley 1974; Freedman et al. 2007; Gibbons and Pratt 1975; Gigerenzer et al. 1990, Ch. 3; Harlow et al. 1997; Hogben 1957; Kaye and Freedman 2011; Morrison and Henkel 1970; Oakes 1986; Pratt 1965; Rothman et al. 2008, Ch. 10; Ware et al. 2009; Ziliak and McCloskey 2008). Specific issues are covered at length in these sources and in the many peer-reviewed articles that critique common misinterpretations of null-hypothesis testing and "statistical significance" (e.g., Altman and Bland 1995; Anscombe 1990; Bakan 1966; Bandt and Boen 1972; Berkson 1942; Bland and Altman 2015; Chia 1997; Cohen 1994; Evans et al. 1988; Fidler and Loftus 2009; Gardner and Altman 1986; Gelman 2013; Gelman and Loken 2014; Gelman and Stern 2006; Gigerenzer 2004;

©2016 S. Greenland, S. Senn, K. Rothman, J. Carlin, C. Poole, S. Goodman, and D. Altman

Gigerenzer and Marewski 2015; Goodman 1992, 1993, 1999, 2008; Greenland 2011, 2012ab; Greenland and Poole, 2011, 2013ab; Grieve 2015; Harlow et al. 1997; Hoekstra et al. 2006; Hurlbert and Lombardi 2009; Kaye 1986; Lambdin 2012; Lang et al. 1998; Langman 1986; LeCoutre et al. 2003; Lew 2012; Loftus 1996; Matthews and Altman 1996a; Pocock and Ware 2009; Pocock et al. 1987; Poole 1987ab, 2001; Rosnow and Rosenthal 1989; Rothman 1978, 1986; Rozeboom 1960; Salsburg 1985; Schmidt 1996; Schmidt and Hunter 2002; Sterne and Davey Smith 2001; Thompson 1987; Thompson 2004; Wagenmakers 2007; Walker 1986; Wood et al. 2014).

Statistical Tests, P-values, and Confidence Intervals: A Caustic Primer

Statistical Models, Hypotheses, and Tests

Every method of statistical inference depends on a complex web of assumptions about how data were collected and analyzed, and how the analysis results were selected for presentation. The full set of assumptions is embodied in a *statistical model* that underpins the method. This model is a mathematical representation of data variability, and thus ideally would capture accurately all sources of such variability. Many problems arise, however, because this statistical model often incorporates unrealistic or at best unjustified assumptions. This is true even for so-called "nonparametric" methods, which (like other methods) depend on assumptions of random sampling or randomization. These assumptions are often deceptively simple to write mathematically, yet in practice are difficult to satisfy and verify, as they may depend on successful completion of a long sequence of actions (such as identifying, contacting, obtaining consent from, obtaining cooperation of, and following up subjects, as well as adherence to study protocols for treatment allocation, masking, and data analysis).

There is also a serious problem of defining the scope of a model, in that it should allow not only for a good representation of the observed data but also of hypothetical alternative data that might have been observed. The reference frame for data that "might have been observed" is often unclear, for example if multiple outcome measures or multiple predictive factors have been measured, and many decisions surrounding analysis choices have been made after the data were collected—as is invariably the case (Gelman and Loken 2014).

The difficulty of understanding and assessing underlying assumptions is exacerbated by the fact that the statistical model is usually presented in a highly compressed and abstract form—if presented at all. As a result, many assumptions go unremarked and are often unrecognized by users as well as consumers of statistics. Nonetheless, all statistical methods and interpretations are premised on the model assumptions; that is, on an assumption that the model provides a valid representation of the variation we would expect to see across data sets, faithfully reflecting the circumstances surrounding the study and phenomena occurring within it.

In most applications of statistical testing, one assumption in the model is a hypothesis that a particular effect has a specific size, and has been targeted for statistical analysis. (For simplicity, we use the word "effect" when "association or effect" would arguably be better in allowing for noncausal studies such as most surveys.) This targeted assumption is called the study hypothesis or *test hypothesis*, and the statistical methods used to evaluate it are called *statistical hypothesis tests*. Most often, the targeted effect size is a "null" value representing *zero* effect (e.g., that the study treatment makes no difference in average outcome), in which case the test hypothesis is called the *null hypothesis*. Nonetheless, it is also possible to test other effect sizes. We may also test hypotheses that the effect does or does not fall within a specific range; for example, we may test the hypothesis that the effect is no greater than a particular amount, in which case the hypothesis is said to be a *one-sided* or *dividing* hypothesis (Cox 1977, 1982).

Much statistical teaching and practice has developed a strong (and unhealthy) focus on the idea that the main aim of a study should be to test null hypotheses. In fact most descriptions of statistical testing focus *only* on testing null hypotheses, and the entire topic has been called "Null Hypothesis Significance Testing" (NHST). This exclusive focus on null hypotheses contributes to misunderstanding of tests. Adding to the misunderstanding is that many authors (including R.A. Fisher) use "null hypothesis" to refer to any test hypothesis, even though this usage is at odds with other authors and with ordinary English definitions of "null"—as are statistical usages of "significance" and "confidence."

Uncertainty, Probability, and Statistical Significance

A more refined goal of statistical analysis is to provide an evaluation of certainty or uncertainty regarding the size of an effect. It is natural to express such certainty in terms of "probabilities" of hypotheses. In conventional statistical methods, however, "probability" refers not to hypotheses, but to quantities that are hypothetical frequencies of data patterns under an assumed statistical model. These methods are thus called *frequentist* methods, and the hypothetical frequencies they predict are called "frequency probabilities." Despite considerable training to the contrary, many statistically educated scientists revert to the habit of misinterpreting these frequency probabilities as hypothesis probabilities. (Even more confusingly, the term "likelihood of a parameter value" is reserved by statisticians to refer to the probability of the observed data *given* the parameter value; it does not refer to a probability of the parameter taking on the given value.)

Nowhere are these problems more rampant than in applications of a hypothetical frequency called the *P-value*, also known as the "observed significance level" for the test hypothesis. Statistical "significance tests" based on this concept have been a central part of statistical analyses for centuries (Stigler 1986). The focus of traditional definitions of P-values and statistical significance has been on null hypotheses, treating all other assumptions used to compute the P-value as if they were known to be correct. Recognizing that these other assumptions are often questionable if not unwarranted, we will adopt a more general view of the P-value as a statistical summary of the compatibility between the observed data and what we would predict or expect to see if we knew the entire statistical model (*all* the assumptions used to compute the P-value) were correct.

Specifically, the distance between the data and the model prediction is measured using a *test statistic* (such as a t-statistic or a chi-squared statistic). The P-value is then the probability that the chosen test statistic would have been *at least* as large as its observed value if *every* model assumption were correct, including the test hypothesis. This definition embodies a crucial point lost in traditional definitions: In logical terms, the P-value tests *all* the assumptions about how the data were generated (the entire model), not just the targeted hypothesis it is supposed to test (such as a null hypothesis). Furthermore, these assumptions include far more than what are traditionally presented as modeling or probability assumptions—they include assumptions about the conduct of the analysis, for example that intermediate analysis results were not used to determine which analyses would be presented.

It is true that the smaller the P-value, the more unusual the data would be *if* every single assumption were correct; but a very small P-value does *not* tell us which assumption is incorrect. For example, the P-value may be very small because the targeted hypothesis is false; but it may instead (or in addition) be very small because the study protocols were violated, or because it was selected for presentation based on its small size. Conversely, a large P-value indicates only that the data are not unusual under the model, but does not imply that the model or any aspect of it (such as the targeted hypothesis) is correct; it may instead (or in addition) be large because (again) the study protocols were violated, or because it was selected for presentation based on its large size.

The general definition of a P-value may help one to understand why statistical tests tell us much less than what many think they do: Not only does a P-value *not* tell us whether the hypothesis targeted for testing is true or not; it says nothing specifically related to that hypothesis unless we can be completely assured that every other assumption used for its computation is correct—an assurance that is lacking in far too many studies.

Nonetheless, the P-value can be viewed as a continuous measure of the compatibility between the data and the entire model used to compute it, ranging from 0 for complete incompatibility to 1 for perfect compatibility, and in this sense may be viewed as measuring the fit of the model to the data. Too often, however, the P-value is degraded into a dichotomy in which results are declared "statistically significant" if P falls on or below a cut-off (usually 0.05) and declared "nonsignificant" otherwise. The terms "significance level" and "alpha level" (α) are often used to refer to the cut-off; however, the term "significance level" invites confusion of the cut-off with the P-value itself. Their difference is profound: the cut-off value α is supposed to be fixed in advance and is thus part of the study design, unchanged in light of the data. In contrast, the P-value is a number computed from the data and thus an analysis result, unknown until it is computed.

Moving From Tests to Estimates

We can vary the test hypothesis while leaving other assumptions unchanged, to see how the P-value differs across competing test hypotheses. Usually, these test hypotheses specify different sizes for a targeted effect; for example, we may test the hypothesis that the average difference between two treatment groups is zero (the null hypothesis), or that it is 20 or –10 or any size of interest. The effect size whose test produced $P = 1$ is the size most compatible with the data (in the sense of predicting what was in fact observed) *if* all the other assumptions used in the test (the statistical model) were correct, and provides a *point estimate* of the effect under those assumptions. The effect sizes whose test produced $P > 0.05$ will typically define a range of sizes (e.g., from 11.0 to 19.5) that would be considered more compatible with the data (in the sense of the observations being closer to what the model predicted) than sizes outside the range—again, if the statistical model were correct. This range corresponds to a $1 - 0.05 = 0.95$ or 95% *confidence interval*, and provides a convenient way of summarizing the results of hypothesis tests for many effect sizes. Confidence intervals are examples of *interval estimates*.

Neyman (1937) proposed the construction of confidence intervals in this way because they have the following property: If one calculates, say, 95% confidence intervals repeatedly *in valid applications*, 95% of them, on average, will contain (i.e., include or cover) the true effect size. Hence, the specified confidence level is called the *coverage probability*. As Neyman stressed repeatedly, this coverage probability is a property of a long sequence of confidence intervals computed from valid models, rather than a property of any single confidence interval.

Many journals now require confidence intervals, but most textbooks and studies discuss P-values only for the null hypothesis of no effect. This exclusive focus on null hypotheses in testing not only contributes to misunderstanding of tests and underappreciation of estimation, but also obscures the close relationship between P-values and confidence intervals, as well as the weaknesses they share.

What P-values, Confidence Intervals, and Power Calculations Don't Tell Us

Much distortion arises from basic misunderstanding of what P-values and their relatives (such as confidence intervals) do *not* tell us. Therefore, based on the articles in our reference list, we review prevalent P-value misinterpretations as a way of moving toward defensible interpretations and presentations. We adopt the format of Goodman (2008) in providing a list of misinterpretations that can be used to critically evaluate conclusions offered by research reports and reviews. Every one of the italicized statements in our list has contributed to statistical distortion of the scientific literature, and we add the emphatic "No!" to underscore statements that are not only fallacious but also not "true enough for practical purposes."

Common Misinterpretations of Single P-values

1. *The P-value is the probability that the test hypothesis is true; for example, if a test of the null hypothesis gave $P = 0.01$, the null hypothesis has only a 1% chance of being true; if instead it gave $P = 0.40$, the null hypothesis has a 40% chance of being true.*—No! The P-value *assumes* the test hypothesis is true—it is *not* a hypothesis probability and may be far from any

reasonable probability for the test hypothesis. The P-value simply indicates the degree to which the data conform to the pattern predicted by the test hypothesis and all the other assumptions used in the test (the underlying statistical model). Thus $P = 0.01$ would indicate that the data are not very close to what the statistical model (including the test hypothesis) predicted they should be, while $P = 0.40$ would indicate that the data are much closer to the model prediction, allowing for chance variation.

2. The P-value for the null hypothesis is the probability that chance alone produced the observed association; for example, if the P-value for the null hypothesis is 0.08, there is an 8% probability that chance alone produced the association.—No! This is a common variation of the first fallacy and it is just as false. To say that chance *alone* produced the observed association is logically equivalent to asserting that every assumption used to compute the P-value is correct, *including the null hypothesis*. Thus to claim that the null P-value is the probability that chance alone produced the observed association is completely backwards: The P-value is a probability computed *assuming* chance was operating alone. The absurdity of the common backwards interpretation might be appreciated by pondering how the P-value, which is a probability deduced *from* a set of assumptions (the statistical model), can possibly refer to the probability *of* those assumptions.

Note: One often sees "alone" dropped from this description (becoming "the P-value for the null hypothesis is the probability that chance produced the observed association"), so that the statement is more ambiguous, but just as wrong.

3. A significant test result ($P \leq 0.05$) means that the test hypothesis is false or should be rejected.—No! A small P-value simply flags the data as being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; it may be small because there was a large random error or because some assumption other than the test hypothesis was violated (for example, the assumption that this P-value was not selected for presentation because it was below 0.05). $P \leq 0.05$ only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large *or larger than* that observed no more than 5% of the time if *only* chance were creating the discrepancy (as opposed to a violation of the test hypothesis or a mistaken assumption).

4. A nonsignificant test result ($P > 0.05$) means that the test hypothesis is true or should be accepted.—No! A large P-value only suggests that the data are *not* unusual if all the assumptions used to compute the P-value (including the test hypothesis) were correct. The same data would also not be unusual under many other hypotheses. Furthermore, even if the test hypothesis is wrong, the P-value may be large because it was inflated by a large random error or because of some other erroneous assumption (e.g., the assumption that this P-value was not selected for presentation because it was above 0.05). $P > 0.05$ only means that a discrepancy from the hypothesis prediction (e.g., no difference between treatment groups) would be as large *or larger than* that observed more than 5% of the time if *only* chance were creating the discrepancy.

5. A large P-value is evidence in favor of the test hypothesis.—No! In fact, any P-value less than 1 implies that the test hypothesis is *not* the hypothesis most compatible with the data, because any other hypothesis with a larger P-value would be even more compatible with the data. A P-value cannot be said to favor the test hypothesis except in relation to those hypotheses with smaller P-values. Furthermore, a large P-value often indicates only that the data are incapable of discriminating among many competing hypotheses (as would be seen immediately by examining the range of the confidence interval). For example, many authors will misinterpret $P = 0.70$ from a test of the null hypothesis as evidence for no effect, when in fact it indicates that, even though the null hypothesis is compatible with the data under the assumptions used to compute the P-value, it is *not* the hypothesis most compatible with the data—that honor would belong to a hypothesis with $P = 1$. But even if $P = 1$, there will be many other hypotheses that are highly consistent with the data, so that a definitive conclusion of "no association" cannot be deduced from a P-value, no matter how large.

6. A null-hypothesis P-value greater than 0.05 means that no effect was observed, or that absence of an effect was shown or demonstrated.—No! Observing $P > 0.05$ for the null hypothesis only means that the null is one among the many hypotheses that have $P > 0.05$. Thus, unless the point estimate (observed association) equals the null value exactly, it is a mistake to conclude from $P > 0.05$ that a study found "no association" or "no evidence" of an effect. If the null P-value is less than 1 some association must be present in the data, and one must look at the point estimate to determine the effect size most compatible with the data under the assumed model.

7. Statistical significance indicates a scientifically or substantively important relation has been detected.—No! Especially when a study is large, very minor effects or small assumption violations can lead to statistically significant tests of the null hypothesis. Again, a small null P-value simply flags the data as being unusual if all the assumptions used to compute it (including the null hypothesis) were correct; but the way the data are unusual might be of no clinical interest. One must look at the confidence interval to determine which effect sizes of scientific or other substantive (e.g., clinical) importance are relatively compatible with the data, given the model.

8. Lack of statistical significance indicates that the effect size is small.—No! Especially when a study is small, even large effects may be "drowned in noise" and thus fail to be detected as statistically significant by a statistical test. A large null P-value simply flags the data as *not* being unusual if all the assumptions used to compute it (including the test hypothesis) were correct; but the same data will also not be unusual under many other models and hypotheses besides the null. Again, one must look at the confidence interval to determine whether it includes effect sizes of importance.

9. The P-value is the chance of our data occurring if the test hypothesis is true; for example, $P = 0.05$ means that the observed association would occur only 5% of the time under the test hypothesis.—No! The P-value refers not only to what we observed, but also observations *more extreme* than what we observed (where "extremity" is measured in a particular way). And again, the P-value refers to a data frequency when all the assumptions used to compute it are correct. In addition to the test hypothesis, these assumptions include randomness in sampling, treatment assignment, loss, and missingness, as well as an assumption that the P-value was not selected for presentation based on its size or some other aspect of the results.

10. If you reject the test hypothesis because $P \leq 0.05$, the chance you are in error (the chance your "significant finding" is a false positive) is 5%.—No! To see why this description is false, suppose the test hypothesis is in fact true. Then, if you reject it, the chance you are in error is 100%, not 5%. The 5% refers only to how often you would reject it, and therefore be in error, over very many uses of the test across different studies when the test hypothesis and all other assumptions used for the test are true. It does not refer to your single use of the test, which may have been thrown off by assumption violations as well as random errors. This is yet another version of misinterpretation #1.

11. $P = 0.05$ and $P \leq 0.05$ mean the same thing.—No! This is like saying reported height = 2 meters and reported height ≤ 2 meters are the same thing: "height = 2 meters" would include few people and those people would be considered tall, whereas "height ≤ 2 meters" would include most people including small children. Similarly, $P = 0.05$ would be considered a borderline result in terms of statistical significance, whereas $P \leq 0.05$ lumps borderline results together with results very incompatible with the model (e.g., $P = 0.0001$) thus rendering its meaning vague, for no good purpose.

12. P-values are properly reported as inequalities (e.g., report "$P < 0.02$" when $P = 0.015$ or report $P > 0.05$ when $P = 0.06$ or $P = 0.70$).—No! This is bad practice because it makes it difficult or impossible for the reader to accurately interpret the statistical result. Only when the P-value is very small (e.g., under 0.001) does an inequality become justifiable: There is little practical difference among very small P-values when the assumptions used to compute P-values are not known with enough certainty to justify such precision, and most methods for computing P-values are not numerically accurate below a certain point.

13. Statistical significance is a property of the phenomenon being studied, and thus statistical tests detect significance.—No! This misinterpretation is promoted when researchers state that they have or have not found "evidence of" a statistically significant effect. The effect being tested either exists or does not exist. "Statistical significance" is a dichotomous description of a P-value (that it is below the chosen cut-off) and thus is a property of a result of a statistical test; it is not a property of the effect or population being studied.

14. One should always use two-sided P-values.—No! Two-sided P-values are designed to test hypotheses that the targeted effect measure equals a specific value (e.g., zero), and is neither above nor below this value. When however the test hypothesis of scientific or practical interest is a one-sided (dividing) hypothesis, a one-sided P-value is appropriate. For example, consider the practical question of whether a new drug is *at least* as good as the standard drug for increasing survival time. This question is one-sided, so testing this hypothesis calls for a one-sided P-value. Nonetheless, because two-sided P-values are the usual default, it will be important to note when and why a one-sided P-value is being used instead.

There are other interpretations of P values that are controversial, in that whether a categorical "No!" is warranted depends on one's philosophy of statistics and the precise meaning given to the terms involved. The disputed claims deserve recognition if one wishes to avoid such controversy.

For example, it has been argued that P-values overstate evidence against test hypotheses, based on directly comparing P-values against certain quantities (likelihood ratios and Bayes factors) that play a central role as evidence measures in Bayesian analysis (Edwards et al. 1963; Berger and Sellke 1987; Edwards 1992; Goodman and Royall 1988; Royall 1997; Sellke et al. 2001; Goodman 1992, 2005; Wagenmakers 2007). Nonetheless, many other statisticians do not accept these quantities as gold standards, and instead point out that P-values summarize crucial evidence needed to gauge the error rates of decisions based on statistical tests (even though they are far from sufficient for making those decisions). Thus, from this frequentist perspective, P-values do not overstate evidence and may even be considered as measuring one aspect of evidence (Cox 1977, 1982; Lehmann 1986; Senn 2001, 2002a; Mayo and Cox 2006), with $1 - P$ measuring evidence against the model used to compute the P-value. See also Murtaugh (2014) and its accompanying discussion.

Common Misinterpretations of P-Value Comparisons and Predictions

Some of the most severe distortions of the scientific literature produced by statistical testing involve erroneous comparison and synthesis of results from different studies or study subgroups. Among the worst are:

15. When the same hypothesis is tested in different studies and none or a minority of the tests are statistically significant (all $P > 0.05$), the overall evidence supports the hypothesis.—No! This belief is often used to claim that a literature supports no effect when the opposite is case. It reflects a tendency of researchers to "overestimate the power of most research" (Hedges and Olkin 1980). In reality, every study could fail to reach statistical significance and yet when combined show a statistically significant association and persuasive evidence of an effect. For example, if there were five studies each with $P = 0.10$, none would be significant at 0.05 level; but when these P-values are combined using the Fisher formula (Cox and Hinkley 1974, p. 80), the overall P-value would be 0.01. There are many real examples of persuasive evidence for important effects when few studies or even no study reported "statistically significant" associations (e.g., Chalmers and Lau 1996; Maheshwari et al. 2007). Thus, lack of statistical significance of individual studies should not be taken as implying that the totality of evidence supports no effect.

16. When the same hypothesis is tested in two different populations and the resulting P-values are on opposite sides of 0.05, the results are conflicting.—No! Statistical tests are sensitive to many differences between study populations that are irrelevant to whether their results are in agreement, such as the sizes of compared groups in each population. As a consequence, two studies may provide very different P-values for the same test hypothesis and yet be in perfect agreement (e.g., may show identical observed associations). For example, suppose we had two randomized trials A and B of a treatment, identical except that trial A had a known standard error of 2 for the mean difference between treatment groups whereas trial B had a known standard error of 1 for the difference. If both trials observed a difference between treatment groups of exactly 3, the usual normal test would produce $P = 0.13$ in A but $P = 0.003$ in B. Despite their difference in P-values, the test of the hypothesis of no difference in effect across studies would have $P = 1$, reflecting the perfect agreement of the observed mean differences from the studies. Differences between results must be evaluated by directly, for example by estimating and testing those differences to produce a confidence interval and a P-value comparing the results (often called analysis of heterogeneity, interaction, or modification).

17. When the same hypothesis is tested in two different populations and the same P-values are obtained, the results are in agreement.—No! Again, tests are sensitive to many differences between populations that are irrelevant to whether their results are in agreement. Two different studies may even exhibit identical P-values for testing the same hypothesis yet also exhibit clearly different observed associations. For example, suppose randomized experiment A observed a mean difference between treatment groups of 3.00 with standard error 1.00, while B observed a mean difference of 12.00 with standard error 4.00. Then the standard normal test would produce $P = 0.003$ in both; yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$, reflecting the large difference $(12.00 - 3.00 = 9.00)$ between the mean differences.

18. If one observes a small P-value, there is a good chance that the next study will produce a P-value at least as small for the same hypothesis.—No! This is false even under the ideal condition that both studies are independent and all assumptions including the test hypothesis are correct in both studies. In that case, if (say) one observes $P = 0.03$, the chance that the new study will show $P \leq 0.03$ is only 3%; thus the chance the new study will show a P-value as small or smaller (the "replication probability") is exactly the observed P-value! If on the other hand the small P-value arose solely because the true effect exactly equaled its observed estimate, there would be a 50% chance that a repeat experiment of identical design would have a larger P-value (Goodman 1992). In general, the size of the new P-value will be extremely sensitive to the study size and the extent to which the test hypothesis or other assumptions are violated in the new study (Senn 2002a); in particular, P may be very small or very large depending on whether the study and the violations are large or small.

Finally, although it is (we hope obviously) wrong to do so, one sometimes sees the null hypothesis compared with another (alternative) hypothesis using a two-sided P-value for the null and a one-sided P-value for the alternative. This comparison is biased in favor of the null in that the two-sided test will falsely reject the null only half as often as the one-sided test will falsely reject the alternative (again, under all the assumptions used for testing).

Common Misinterpretations of Confidence Intervals

Most of the above misinterpretations translate into an analogous misinterpretation for confidence intervals. For example, another misinterpretation of $P > 0.05$ is that it means the test hypothesis has only a 5% chance of being false, which in terms of a confidence interval becomes the common fallacy:

19. The specific 95% confidence interval presented by a study has a 95% chance of containing the true effect size.—No! A reported confidence interval is a range between two numbers. The frequency with which an observed interval (e.g., 0.72 to 2.88) contains the true effect is either 100% if the true effect is within the interval or 0% if not; the 95% refers only to how often 95% confidence intervals computed from very many studies would contain the true size *if all the assumptions used to compute the intervals were correct*. It is possible to compute an interval that can be interpreted as having 95% probability of containing the true value; nonetheless, such computations require not only the assumptions used to compute the confidence interval, but also further assumptions about the size of effects in the model. These further assumptions are summarized in what is called a *prior distribution*, and the resulting intervals are usually called *Bayesian posterior (or credible) intervals* to distinguish them from confidence intervals (e.g., see Rothman et al. 2008, Ch. 13 and 18).

Symmetrically, the misinterpretation of a small P-value as disproving the test hypothesis could be translated into:

20. An effect size outside the 95% confidence interval has been refuted (or excluded) by the data.—No! As with the P-value, the confidence interval is computed from many assumptions, the violation of which may have led to the results. Thus it is the combination of the data with the assumptions, along with the arbitrary 95% criterion, that are needed to declare an effect size outside the interval is in some way incompatible with the observations. Even then, judgements as extreme as saying the effect size has been refuted or excluded will require even stronger conditions.

As with P-values, nave comparison of confidence intervals can be highly misleading:

21. If two confidence intervals overlap, the difference between two estimates or studies is not significant.—No! The 95% confidence intervals from two subgroups or studies may overlap substantially and yet the test for difference between them may still produce $P < 0.05$. Suppose for example, two 95% confidence intervals for means from normal populations with known variances are (1.04, 4.96) and (4.16, 19.84); these intervals overlap, yet the test of the hypothesis of no difference in effect across studies gives $P = 0.03$. As with P-values, comparison between groups requires statistics that directly test and estimate the differences across groups. It can, however, be noted that if the two 95% confidence intervals fail to overlap, then when using the same assumptions used to compute the confidence intervals we will find $P < 0.05$ for the difference; and if one of the 95% intervals contains the point estimate from the other group or study, we will find $P > 0.05$ for the difference.

Finally, as with P-values, the replication properties of confidence intervals are usually misunderstood:

22. An observed 95% confidence interval predicts that 95% of the estimates from future studies will fall inside the observed interval.—No! This statement is wrong in several ways. Most importantly, under the model, 95% is the frequency with which *other unobserved* intervals will contain the *true effect*, not how frequently the one interval being presented will contain future estimates. In fact, even under ideal conditions the chance that a future estimate will fall within the current interval will usually be much less than 95%. For example, if two independent studies of the same quantity provide unbiased normal point estimates with the same standard errors, the chance that the 95% confidence interval for the first study contains the point estimate from the second is 83% (which is the chance that the difference between the two estimates is less than 1.96 standard errors). Again, an observed interval either does or does not contain the true effect; the 95% refers only to how often 95% confidence intervals computed from very many studies would contain the true effect if all the assumptions used to compute the intervals were correct.

23. If one 95% confidence interval includes the null value and another excludes that value, the interval excluding the null is the more precise one.—No! When the model is correct, precision of statistical estimation is measured directly by confidence interval *width* (measured on the appropriate scale). It is not a matter of inclusion or exclusion of the null or any other value. Consider two 95% confidence intervals for a difference in means, one with limits of 5 and 40, the other with limits of -5 and 10. The first interval excludes the null value of 0, but is 30 units wide. The second includes the null value, but is half as wide and therefore much more precise.

In addition to the above misinterpretations, 95% confidence intervals force the 0.05-level cutoff on the reader, lumping together all effect sizes with $P > 0.05$, and in this way are as bad as presenting P-values as dichotomies. Nonetheless, many authors agree that confidence intervals are superior to tests and P-values because they allow one to shift focus away from the null hypothesis, toward the full range of effect sizes compatible with the data—a shift recommended by many authors and a growing number of journals. Another way to bring attention to nonnull hypotheses is to present their P-values; for example, one could provide or demand P-values for those effect sizes that are recognized as scientifically reasonable alternatives to the null.

As with P-values, further cautions are needed to avoid misinterpreting confidence intervals as providing sharp answers when none are warranted. The hypothesis which says the point estimate is the correct effect will have the largest P-value ($P = 1$ in most cases), and hypotheses inside a confidence interval will have higher P-values than hypotheses outside the interval. The P-values will vary greatly, however, among hypotheses inside the interval, as well as among hypotheses on the outside. Also, two hypotheses may have nearly equal P-values even though one of the hypotheses is inside the interval and the other is outside. Thus, if we use P-values to measure compatibility of hypotheses with data and wish to compare hypotheses with this measure, we need to examine their P-values directly, not simply ask whether the hypotheses are inside or outside the interval. This need is particularly acute when (as usual) one of the hypotheses under scrutiny is a null hypothesis.

Common Misinterpretations of Power

The *power* of a test to detect a correct alternative hypothesis is the pre-study probability that the test will reject the test hypothesis (e.g., the probability that P will not exceed a prespecified cut-off such as 0.05). (The corresponding prestudy probability of failing to reject the test hypothesis when the alternative is correct is one minus the power, also known as the Type-II or beta error rate; see Lehmann 1986.) As with P-values and confidence intervals, this probability is defined over repetitions of the same study design and so is a frequency probability. One source of reasonable alternative hypotheses are the effect sizes that were used to compute power in the study proposal. Prestudy power calculations do not, however, measure the compatibility of these alternatives with the data actually observed, while power calculated from the observed data is a direct (if obscure) transformation of the null P-value and so provides no test of the alternatives. Thus, presentation of power does not obviate the need to provide interval estimates and direct tests of the alternatives.

For these reasons, many authors have condemned use of power to interpret estimates and statistical tests (e.g., Cox 1958; Smith and Bates 1992; Goodman 1994; Goodman and Berlin 1994; Hoenig and Heisey 2001; Senn 2002b; Greenland 2012a), arguing that (in contrast to confidence intervals) it distracts attention from direct comparisons of hypotheses and introduces new misinterpretations, such as:

24. If you accept the null hypothesis because the null P-value exceeds 0.05 and the power of your test is 90%, the chance you are in error (the chance that your finding is a false negative) is 10%.—No! If the null hypothesis is false and you accept it, the chance you are in error is 100%, not 10%. Conversely, if the null hypothesis is true and you accept it, the chance you are in error is 0%. The 10% refers only to how often you would be in error over very many uses of the test across different studies when the particular alternative used to compute power is correct *and* all other assumptions used for the test are correct in all the studies. It does not refer to your single use of the test or your error rate under any alternative effect size other than the one used to compute power.

It can be especially misleading to compare results for two hypotheses by presenting a test or P-value for one and power for the other. For example, testing the null by seeing whether $P \leq 0.05$ with a power less than $1-0.05 = 0.95$ for the alternative (as done routinely) will bias the comparison in favor of the null because it entails a lower probability of incorrectly rejecting the null (0.05) than of incorrectly accepting the null when the alternative is correct. Thus, claims about relative support or evidence need to be based on direct and comparable measures of support or evidence for both hypotheses, otherwise mistakes like the following will occur:

25. If the null P-value exceeds 0.05 and the power of this test is 90% at an alternative, the results support the null over the alternative.—This claim seems intuitive to many, but counterexamples are easy to construct in which the null P-value is between 0.05 and 0.10, and yet there are alternatives whose own P-value exceeds 0.10 and for which the power is 0.90. Parallel results ensue for other accepted measures of compatibility, evidence, and support, indicating that the data show lower compatibility with and more evidence against the null than the alternative, despite the fact that the null P-value is "not significant" at the 0.05 alpha level and the power against the alternative is "very high" (Greenland, 2012a).

Despite its shortcomings for interpreting current data, power can be useful for designing studies and for understanding why replication of "statistical significance" will often fail even under ideal conditions. Studies are often designed or claimed to have 80% power against a key alternative when using a 0.05 significance level, although in execution often have less power due to unanticipated problems such as low subject recruitment. Thus, if the alternative is correct and the actual power of two studies is 80%, the chance that the studies will both show $P \leq 0.05$ will at best be only $0.80(0.80) = 64\%$; furthermore, the chance that one study shows $P \leq 0.05$ and the other does not (and thus will be misinterpreted as showing conflicting results) is $2(0.80)0.20 = 32\%$ or about 1 chance in 3. Similar calculations taking account of typical problems suggest that one could anticipate a "replication crisis" even if there were no publication or reporting bias, simply because current design and testing conventions treat individual study results as dichotomous outputs of "significant"/"nonsignificant" or "reject"/"accept."

A Statistical Model is Much More Than an Equation with Greek Letters

The above list could be expanded by reviewing the research literature. We will however turn to direct discussion of an issue that has been receiving more attention of late, yet is still widely overlooked or interpreted too narrowly in statistical teaching and presentations: That the statistical model used to obtain the results is correct.

Too often, the full statistical model is treated as a simple regression or structural equation in which effects are represented by parameters denoted by Greek letters. "Model checking" is then limited to tests of fit or testing additional terms for the model. Yet these tests of fit themselves make further assumptions that should be seen as part of the full model. For example, all common tests and confidence intervals depend on assumptions of random selection for observation or treatment and random loss or missingness within levels of controlled covariates. These assumptions have gradually come under scrutiny via sensitivity and bias analysis (e.g., Lash et al. 2014), but such methods remain far removed from the basic statistical training given to most researchers.

Less often stated is the even more crucial assumption that the analyses themselves were not guided toward finding nonsignificance or significance (analysis bias), and that the analysis results were not reported based on their nonsignificance or significance (reporting bias and publication bias). Selective reporting renders false even the limited ideal meanings of statistical significance, P-values, and confidence intervals. Because author decisions to report and editorial decisions to publish results often depend on whether the P-value is above or below 0.05, selective reporting has been identified as a major problem in large segments of the scientific literature (Dwan et al. 2013; Page et al. 2014; You et al. 2012).

Although this selection problem has also been subject to sensitivity analysis, there has been a bias in studies of reporting and publication bias: It is usually assumed that these biases favor significance. This assumption is of course correct when (as is often the case) researchers select results for presentation when $P \leq 0.05$, a practice that tends to exaggerate associations (Button et al. 2013; Eyding et al. 2010; Land 1980; Land 1981). Nonetheless, bias in favor of reporting $P \leq 0.05$ is not always plausible let alone supported by evidence or common sense. For example, one might expect selection for $P > 0.05$ in publications funded by those with stakes in acceptance of the null hypothesis (a practice which tends to understate associations); in accord with that expectation, some empirical studies have observed smaller estimates and "nonsignificance" more often in such publications than in other studies (Eyding et al. 2010; Greenland 2009; Xu et al. 2013).

Addressing such problems would require far more political will and effort than addressing misinterpretation of statistics, such as enforcing registration of trials, along with open data and analysis code from all completed studies (as in the AllTrials initiative, *http://www.alltrials.net/*). In the meantime, readers are advised to consider the entire context in which research reports are produced and appear when interpreting the statistics and conclusions offered by the reports.

Conclusions

Upon realizing that statistical tests are usually misinterpreted, one may wonder what if anything these tests do for science. They were originally intended to account for random variability as a source of error, thereby sounding a note of caution against overinterpretation of observed associations as true effects or as stronger evidence against null hypotheses than was warranted. But before long that use was turned on its head to provide fallacious support for null hypotheses in the form of "failure to achieve" or "failure to attain" statistical significance.

We have no doubt that the founders of modern statistical testing would be horrified by common treatments of their invention. In their first paper describing their binary approach to statistical testing, Neyman and Pearson (1928) wrote that "it is doubtful whether the knowledge that [a P-value] was really 0.03 (or 0.06), rather than 0.05 ... would in fact ever modify our judgment" and that "The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision." Pearson (1955) later added, "No doubt we could more aptly have said, 'his final or provisional decision'." Fisher (1956, p. 42) went further, saying "No scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas." Yet fallacious and ritualistic use of tests continued to spread, including beliefs that whether P was above or below 0.05 was a universal arbiter of discovery. Thus by 1965, Hill (1965) lamented that "too often we weaken our capacity to interpret data and to take reasonable decisions whatever the value of P. And far too often we deduce 'no difference' from 'no significant difference'."

In response, it has been argued that some misinterpretations are harmless in tightly controlled experiments on well-understood systems, where the test hypothesis may have special support from established theories (e.g., Mendelian genetics) and in which every other assumption (such as random allocation) is forced to hold by careful design and execution of the study. But it has long been asserted that the harms of statistical testing in more uncontrollable and amorphous research settings (such as social-science, health, and medical fields) have far outweighed its benefits, leading to calls for banning such tests in research reports—again, with one journal banning confidence intervals as well as P-values (Trafimow and Marks 2015).

Given, however, the deep entrenchment of statistical testing, as well as the absence of generally accepted alternative methods, there have been many attempts to salvage P-values by detaching them from their use in significance tests. One approach is to focus on P-values as continuous measures of compatibility, as described earlier. Although this approach has its own limitations (as described in points 1, 2, 5, 9, 17, and 18), it avoids misconceptions arising from comparison of P-values with arbitrary cutoffs such as 0.05 (as described in points 3, 4, 6–8, 10–13, 15, 16, 21, and 23–25). Another approach is to teach and use correct relations of P-values to hypothesis probabilities. For example, under common statistical models, one-sided P-values can provide lower bounds on probabilities for hypotheses about effect directions (Casella and Berger 1987ab; Greenland and Poole 2013ab). Whether such reinterpretations can eventually replace common misinterpretations to good effect remains to be seen.

A shift in emphasis from hypothesis testing to estimation has been promoted as a simple and relatively safe way to improve practice (Yates 1951; Rothman 1978; Altman et al. 2000; Poole 2001; Cumming 2011), resulting in increasing use of confidence intervals and editorial demands for them; nonetheless, this shift has brought to the fore misinterpretations of intervals such as 19–23 above (Morey et al. 2015). Other approaches combine tests of the null with further calculations involving both null and alternative hypotheses (Rosenthal and Rubin 1994; Mayo and Spanos 2006); such calculations may, however, may bring with them further misinterpretations similar to those described above for power, as well as greater complexity.

Meanwhile, in the hopes of minimizing harms of current practice, we can offer several guidelines for users and readers of statistics, and re-emphasize some key warnings from our list of misinterpretations:

a) Correct and careful interpretation of statistical tests demands examining the sizes of effect estimates and confidence limits, as well as precise P-values (not just whether P-values are above or below 0.05 or some other threshold).

b) Careful interpretation also demands critical examination of the assumptions and conventions used for the statistical analysis—not just the usual statistical assumptions, but also the hidden assumptions about how results were generated and chosen for presentation.

c) It is simply false to claim that statistically nonsignificant results support a test hypothesis, because the same results may be even more compatible with alternative hypotheses—even if the power of the test is high for those alternatives.

d) Interval estimates aid in evaluating whether the data are capable of discriminating among various hypotheses about effect sizes, or whether statistical results have been misrepresented as supporting one hypothesis when those results are better explained by other hypotheses (see points 4–6). We caution however that confidence intervals are often only a first step in these tasks. To compare hypotheses in light of the data and the statistical model it may be necessary to calculate the P-value (or relative likelihood) of each hypothesis. We further caution that confidence intervals provide only a best-case measure of the uncertainty or ambiguity left by the data, insofar as they depend on an uncertain statistical model.

e) Correct statistical evaluation of multiple studies requires a pooled analysis or meta-analysis that deals correctly with study biases (Whitehead 2002; Borenstein et al. 2009; Chen and Peace 2013; Cooper et al. 2009; Greenland and O'Rourke 2008; Petitti 2000; Schmidt and Hunter 2014; Sterne 2009). Even when this is done, however, all the earlier cautions apply. Furthermore, the outcome of any statistical procedure is but one of many considerations that must be evaluated when examining the totality of evidence. In particular, statistical significance is neither necessary nor sufficient for determining the scientific or practical significance of a set of observations. This view was affirmed unanimously by the U.S. Supreme Court, (Matrixx Initiatives, Inc., et al. v. Siracusano et al. No. 091156. Argued January 10, 2011, Decided March 22, 2011), and can be seen in our earlier quotes from Neyman and Pearson.

f) Any opinion offered about the *probability, likelihood, certainty*, or similar property for a hypothesis *cannot* be derived from statistical methods alone. In particular, significance tests and confidence intervals do not by themselves provide a logically sound basis for concluding an effect is present or absent with certainty or a given probability. This point should be borne in mind whenever one sees a conclusion framed as a statement of probability, likelihood, or certainty about a hypothesis. Information about the hypothesis beyond that contained in the analyzed data and in conventional statistical models (which give only data probabilities) must be used to reach such a conclusion; that information should be explicitly acknowledged and described by those offering the conclusion. Bayesian statistics offers methods that attempt to incorporate the needed information directly into the statistical model; they have not however achieved the popularity of P-values and confidence intervals, in part because of philosophical objections and in part because no conventions have become established for their use.

g) All statistical methods (whether frequentist or Bayesian, or for testing or estimation, or for inference or decision) make extensive assumptions about the sequence of events that led to the results presented—not only in the data generation, but in the analysis choices. Thus, to allow critical evaluation, research reports (including meta-analyses) should describe in detail the full sequence of events that led to the statistics presented, including the motivation for the study, its design, the original analysis plan, the criteria used to include and exclude subjects (or studies) and data, and a thorough description of all the analyses that were conducted.

In closing, we note that no statistical method is immune to misinterpretation and misuse, but prudent users of statistics will avoid approaches especially prone to serious abuse. In this regard, we join others in singling out the degradation of P-values into “significant” and “nonsignificant” as an especially pernicious statistical practice (Weinberg 2001).

References

Altman, D.G., and 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.

Anscombe, F.J. (1990), “The Summarizing of Clinical Experiments by Significance Levels,” *Statistics in Medicine*, 9, 703–708.

Ashworth, A. (2015), “Veto on the Use of Null Hypothesis Testing and *p* Intervals: Right or Wrong?” Taylor & Francis Editor Resources online, *http://editorresources.taylorandfrancisgroup.com/veto-on-the-use-of-null-hypothesis-testing-and-p-intervals-right-or-wrong/*, accessed Feb. 27, 2016.

Atkins, L., and Jarrett, D. (1979), “The Significance of ‘Significance Tests,”’ in *Demystifying Social Statistics*, Irvine, J., Miles, I., and Evans, J., eds. London: Pluto Press.

Bakan, D. (1966), “The Test of Significance in Psychological Research,” *Psychological Bulletin*, 66, 423–437.

Bandt, C.L., and Boen, J.R. (1972), “A Prevalent Misconception About Sample Size, Statistical Significance, and Clinical Importance,” *Journal of Periodontal Research*, 43, 181–183.

Berger, J.O., and Sellke, T.M. (1987), “Testing a Point Null Hypothesis: The Irreconcilability of P-values and Evidence,” *Journal of the American Statistical Association*, 82, 112–139.

Berkson, J. (1942), “Tests of Significance Considered as Evidence,” *Journal of the American Statistical Association*, 37, 325–335.

Bland, J.M., and Altman, D.G. (2015), “Best (but oft forgotten) Practices: Testing for Treatment Effects in Randomized Trials by Separate Analyses of Changes from Baseline in Each Group is a Misleading Approach,” *American Journal of Clinical Nutrition*, 102, 991–994.

Borenstein, M., Hedges, L.V., Higgins, J.P.T., and Rothstein, H.R. (2009), *Introduction to Meta-Analysis*, New York: Wiley.

Button, K., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafó, M.R. (2013), “Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience,” *Nature Reviews Neuroscience*, 14, 365–376.

Casella, G., and Berger, R.L. (1987a), “Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem,” *Journal of the American Statistical Association*, 82, 106–111.

Casella, G., and Berger, R.L. (1987b), “Comment,” *Stat Sci*, 2, 344–417.

Chalmers, T.C., and Lau, J. (1996), “Changes in Clinical Trials Mandated by the Advent of Meta-analysis,” *Statistics in Medicine*, 15, 1263–1268.

Chen, D.-G., and Peace, K.E. (2013), *Applied Meta-Analysis with R*, New York: Chapman & Hall/CRC.

Chia, K.S. (1997), “‘Significant-itis’An Obsession with the P-Value,” *Scandinavian Journal of Work, Environment & Health*, 23, 152–154.

Cooper, H., Hedges, L.V., and Valentine, J.C. (2009), *The Handbook of Research Synthesis and Meta-Analysis*, Thousand Oaks, CA: Sage.

Cohen, J. (1994), “The Earth is Round ($p < 0.05$),” *American Psychology*, 47, 997–1003.

Cornfield, J. (1966), “Sequential Trials, Sequential Analysis, and the Likelihood Principle,” *The American Statistician*, 25, 617–657.

Cox, D.R. (1958), *The Planning of Experiments*, New York: Wiley, p. 161.

——— (1977), “The Role of Significance Tests” (with discussion), *Scandinavian Journal of Statistics*, 4, 49–70.

——— (1982), “Statistical Significance Tests,” *British Journal of Clinical Pharmacology*, 14, 325–331.

Cox, D.R., and Hinkley, D.V. (1974), *Theoretical Statistics*, New York: Chapman and Hall.

Cumming, G. (2011), *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*, London: Routledge.

Dickersin, K. (1990), “The Existence of Publication Bias and Risk Factors for its Occurrence,” *Journal of the American Medical Association*, 263, 1385–1389.

Dwan, K., Gamble, C., Williamson, P.R., Kirkham, J.J.; Reporting Bias Group (2013), “Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias—An Updated Review,” *PLoS One*, 8:e66844.

Edwards, A.W.F. (1992), *Likelihood* (2nd ed.), Baltimore: Johns Hopkins University Press.

Edwards, W., Lindman, H., and Savage, L.J. (1963), “Bayesian Statistical Inference for Psychological Research,” *Psychological Review*, 70, 193–242.

Evans, S.J.W., Mills, P., and Dawson, J. (1988), “The End of the P-value?” *British Heart Journal*, 60,177–180.

Eyding, D., Lelgemann, M., Grouven, U., Härter, M., Kromp, M., Kaiser, T., Kerekes, M.F., Gerken, M., and Wieseler, B. (2010), “Reboxetine for Acute Treatment of Major Depression: Systematic Review and Meta-analysis of Published and Unpublished Placebo and Selective Serotonin Reuptake Inhibitor Controlled Trials," *British Medical Journal*, 341, c4737.

Fidler, F., and Loftus, G.R. (2009), "Why Figures with Error Bars Should Replace *p* Values: Some Conceptual Arguments and Empirical Demonstrations," *Journal of Psychology*, 217, 27–37.

Fisher, R. A. (1956), *Statistical Methods and Scientific Inference*, Edinburgh, UK: Oliver & Boyd.

Flanagan, O. (2015), "Journal's Ban on Null Hypothesis Significance Testing: Reactions from the Statistical Arena," *Stats Life* online, accessed 27 Feb. 2016.

Freedman, D.A., Pisani, R., and Purves, R. (2007), *Statistics* (4th ed.), New York: Norton.

Gardner, M.A., and Altman, D.G. (1986), "Confidence Intervals Rather than *P* Values: Estimation Rather than Hypothesis Testing," *British Medical Journal*, 292, 746–750.

Gelman, A. (2013), "*P*-Values and Statistical Practice," *Epidemiology*, 24, 69–72.

Gelman, A., and Loken, E. (2014), "The Statistical Crisis in Science: Data-Dependent Analysis—A 'Garden of Forking Paths'—Explains why Many Statistically Significant Comparisons Don't Hold Up," *American Scientist*, 102, 460–465. Erratum at *http://andrewgelman.com/2014/10/14/didnt-say-part-2/*, accessed Feb. 27, 2016.

Gelman, A., and Stern, H.S. (2006), "The Difference Between 'Significant' and 'Not Significant' is not Itself Statistically Significant," *The American Statistician*, 60, 328–331.

Gibbons, J.D., and Pratt, J.W. (1975), "P-Values: Interpretation and Methodology," *The American Statistician*, 29, 20–25.

Gigerenzer, G. (2004), "Mindless Statistics," *Journal of Socioeconomics*, 33, 567–606.

Gigerenzer, G., and Marewski, J.N. (2015), "Surrogate Science: The Idol of a Universal Method for Scientific Inference," *Journal of Management*, 41, 421–440.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., and Kruger, L. (1990), *The Empire of Chance: How Probability Changed Science and Everyday Life*, New York: Cambridge University Press.

Goodman, S.N. (1992), "A Comment on Replication, *p*-values and Evidence," *Statistics in Medicine*, 11, 875–879.

——— (1993), "P-values, Hypothesis Tests and Likelihood: Implications for Epidemiology of a Neglected Historical Debate," *American Journal of Epidemiology*, 137, 485–496.

——— (1994), Letter to the editor re Smith and Bates, *Epidemiology*, 5, 266–268.

——— (1999), "Towards Evidence-Based Medical Statistics, I: The *P*-value Fallacy," *Annals of Internal Medicine*, 130, 995–1004.

——— (2005), "Introduction to Bayesian Methods I: Measuring the Strength of Evidence," *Clinical Trials*, 2, 282–290.

——— (2008), "A Dirty Dozen: Twelve *P*-value Misconceptions," *Seminars in Hematology*, 45, 135–140.

Goodman, S.N., and Berlin, J. (1994), "The Use of Predicted Confidence Intervals when Planning Experiments and the Misuse of Power when Interpreting Results," *Annals of Internal Medicine*, 121, 200–206.

Goodman, S.N., and Royall, R. (1988), "Evidence and Scientific Research," *American Journal of Public Health*, 78, 1568–1574.

Greenland, S. (2009), "Dealing with Uncertainty About iInvestigator Bias: Disclosure is Informative," *Journal of Epidemiology and Community Health*, 63, 593–598.

——— (2011), "Null Misinterpretation in Statistical Testing and its Impact on Health Risk Assessment," *Preventive Medicine*, 53, 225–228.

——— (2012a), "Nonsignificance Plus High Power Does not Imply Support for the Null over the Alternative," *Annals of Epidemiology*, 22, 364–368.

——— (2012b), "Transparency and Disclosure, Neutrality and Balance: Shared Values or Just Shared Words?" *Journal of Epidemiology and Community Health*, 66, 967–970.

Greenland, S., and O'Rourke, K. (2008), "Meta-analysis," in *Modern Epidemiology* (3rd ed.), Rothman, K.J., Greenland, S., Lash, T.L., eds., Philadelphia: Lippincott-Wolters-Kluwer, pp. 682–685.

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

——— (2013a), "Living with *P*-values: Resurrecting a Bayesian Perspective on Frequentist Statistics," *Epidemiology*, 24, 62–68.

——— (2013b), "Living with Statistics in Observational Research," *Epidemiology*, 24, 73–78.

Grieve, A.P. (2015), "How to Test Hypotheses if You Must," *Pharmaceutical Statistics*, 14, 139–150.

Hanley, J.A. (1994), Letter to the Editor re Smith and Bates," *Epidemiology*, 5, 264–266.

Harlow, L.L., Mulaik, S.A., and Steiger, J.H. (1997), "What if There Were No Significance Tests?" *Psychology Press*.

Hauer, E. (2003), "The Harm Done by Tests of Significance," *Accident Analysis & Prevention*, 36, 495–500.

Hedges, L.V., and Olkin, I. (1980), "Vote-Counting Methods in Research Synthesis," *Psychological Bulletin*, 88, 359–369.

Hill, A.B. (1965), "The Environment and Disease: Association or Causation?" *Proceedings of the Royal Society of Medicine*, 58, 295–300.

Hoekstra, R., Finch, S., Kiers, H.A.L., and Johnson, A. (2006), "Probability as Certainty: Dichotomous Thinking and the Misuse of *p*-Values," *Psychological Bulletin Review*, 13, 1033–1037.

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.

Hogben, L. (1957), *Statistical Theory*, London: Allen and Unwin.

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.

Kaye, D.H. (1986), "Is Proof of Statistical Significance Relevant?" *Washington Law Review*, 61, 1333–1366.

Kaye, D.H., and Freedman, D.A. (2011), "Reference Guide on Statistics," in *Reference Manual on Scientific Evidence* (3rd ed.), Washington, DC: Federal Judicial Center, 211–302.

Kline, R.B. (2013), *Beyond Significance Testing: Statistics Reform in the Behavioral Sciences*, Washington, DC: American Psychological Association.

Lambdin, C. (2012), "Significance Tests as Sorcery: Science is Empirical—Significance Tests are Not," *Theory & Psychology*, 22, 67–90.

Land, C.E. (1980), "Estimating Cancer Risks from Low Doses of Ionizing Radiation," *Science*, 209, 1197–1203.

——— (1981), "Statistical Limitations in Relation to Sample Size," *Environmental Health Perspectives*, 42, 15–21.

Lang, J.M., Rothman, K.J., and Cann, C.I. (1998), "That Confounded *P*-Value," *Epidemiology*, 9, 7–8.

Langman, M.J.S. (1986), "Towards Estimation and Confidence Intervals," *BMJ*, 292, 716.

Lash, T.L., Fox, M.P., Maclehose, R.F., Maldonado, G., McCandless, L.C., and Greenland, S. (2014), "Good Practices for Quantitative Bias Analysis," *International Journal of Epidemiology*, 43, 1969–1985.

Lecoutre, M.-P., Poitevineau, J., and Lecoutre, B. (2003), "Even Statisticians are not Immune to Misinterpretations of Null Hypothesis Tests," *International Journal of Psychology*, 38, 37–45.

Lehmann, E.L. (1986), *Testing Statistical Hypotheses* (2nd ed.), New York, Wiley.

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, 1559–1567.

Loftus, G.R. (1996), "Psychology Will be a Much Better Science When We Change the Way We Analyze Data," *Current Directions in Psychology*, 5, 161–171.

Maheshwari, S., Sarraj, A., Kramer, J., and El-Serag, H.B. (2007), "Oral Contraception and the Risk of Hepatocellular Carcinoma," *Journal Hepatology*, 47, 506–513.

Marshall, S.W. (2006), "Commentary on Making Meaningful Inferences About Magnitudes," *Sportscience*, 9, 43–44.

Matthews, J.N.S., and Altman, D.G. (1996a), "Interaction 2: Compare Effect Sizes not P Values," *British Medical Journal*, 313, 808.

——— (1996b), "Interaction 3: How to Examine Heterogeneity," *British Medical Journal*, 313, 862.

Mayo, D.G., and Cox, D.R. (2006), "Frequentist Statistics as a Theory of Inductive Inference," in *Optimality: The Second Erich L. Lehmann Symposium*, Lecture Notes-Monograph Series, J. Rojo (ed.), Hayward, CA: Institute of Mathematical Statistics (IMS) 49, 77–97.

Mayo, D.G., and Spanos, A. (2006), "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction," *British Journal of Philosophical Science*, 57, 323–357.

Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., and Wagenmakers, E.-J. (in press), "The Fallacy of Placing Confidence in Confidence Intervals," *Psychological Bulletin Review*.

Morrison, D.E., and Henkel, R.E. (eds.) (1970), *The Significance Test Controversy* , Chicago: Aldine.

Murtaugh, P.A. (2014), "In Defense of P-Values" (with discussion), *Ecology*, 95, 611–653.

Neyman, J. (1937), "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philospohical Transactions of the Royal Society of London A*, 236, 333–380.

Neyman, J., and Pearson, E.S. (1928), "On the Use and Interpretation of Certain Test Criteria for Purposes of Statistical Inference: Part I," *Biometrika*, 20A, 175–240.

Oakes, M. (1986), *Statistical Inference: A Commentary for the Social and Behavioural Sciences*, Chichester: Wiley.

Page, M.J., McKenzie, J.E., Kirkham, J., Dwan, K., Kramer, S., Green, S., and Forbes, A. (2014), "Bias Due to Selective Inclusion and Reporting of Outcomes and Analyses in Systematic Reviews of Randomised Trials of Healthcare Interventions," *Cochrane Database System Reviews*, 10:MR000035.

Peace, K. (1988), "Some Thoughts on One-Tailed Tests," *Biometrics*, 44, 911–912.

Pearson, E.S. (1955), "Statistical Concepts in the Relation to Reality," *Journal fo the Royal Statistical Society*, Series B, 17, 204–207.

Petitti, D.B. (2000), *Meta-Analysis, Decision Analysis, and Cost-Effectiveness Analysis: Methods for Quantitative Synthesis in Medicine* (2nd ed.), New York: Oxford University Press.

Phillips, C.V. (2004), "Publication Bias In Situ," *BMC Medical Research and Methodology*, 4, 20.

Pocock, S.J., Hughes, M.D., and Lee, R.J. (1987), "Statistical Problems in the Reporting of Clinical Trials," *New England Journal of Medicine*, 317, 426–432.

Pocock, S.J., and Ware, J.H. (2009), "Translating Statistical Findings into Plain English," *The Lancet*, 373, 1926–1928.

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

——— (1987b), "Confidence Intervals Exclude Nothing," *American Journal of Public Health*, 77:, 492–493.

——— (2001), "Low P-Values or Narrow Confidence Intervals: Which are More Durable?", *Epidemiology*, 12, 291–294.

Pratt, J.W. (1965), "Bayesian Interpretation of Standard Inference Statements," *Journal of the Royal Statistical Society*, Series B, 27, 169–203.

Rosenthal, R., and Rubin, D.B. (1994), "The Counternull Value of an Effect Size: A New Statistic," *Psychological Science*, 5, 329–334.

Rosnow, R.L., and Rosenthal, R. (1989), "Statistical Procedures and the Justification of Knowledge in Psychological Science," *American Psychologist*, 44, 1276–1284.

Rothman, K.J. (1978), "A Show of Confidence," *New England School of Medicine*, 299, 1362–1363.

——— (1986), "Significance Questing," *Annals of Internal Medicine*, 105, 445–447.

Rothman, K.J., Greenland, S., and Lash, T.L. (2008), *Modern Epidemiology* (3rd ed.), Philadelphia, PA: Lippincott-Wolters-Kluwer.

Royall, R. (1997), *Statistical Evidence,* New York: Chapman and Hall.

Rozeboom, W.M. (1960), "The Fallacy of Null-Hypothesis Significance Test," *Psychological Bulletin*, 57, 416–428.

Salsburg, D.S. (1985), "The Religion of Statistics as Practiced in Medical Journals," *The American Statistician*, 39, 220–223.

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

Schmidt, F.L. (1996), "Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers," *Psychological Methods*, 1, 115–129.

Schmidt, F.L., and Hunter, J.E. (2014), *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings* (3rd ed.), Thousand Oaks, CA: Sage.

Sellke, T.M., Bayarri, M.J., and Berger, J.O. (2001), "Calibration of p Values for Testing Precise Null Hypotheses," *The American Statistician*, 55, 62–71.

Senn, S.J. (2001), "Two Cheers for P?Values," *Journal of Epidemiology and Biostatistics*, 6, 193–204.

——— (2002a), Letter to the Editor re: Goodman 1992, *Statistics in Medicine*, 21, 2437–2444.

——— (2002b), "Power is Indeed Irrelevant in Interpreting Completed Studies," *BMJ*, 325, 1304.

Smith, A.H., and Bates, M. (1992), "Confidence Limit Analyses Should Replace Power Calculations in the Interpretation of Epidemiologic Studies," *Epidemiology*, 3, 449–452.

Sterne, J.A.C. (2009), *Meta-Analysis: An Updated Collection from the Stata Journal*, College Station, TX: Stata Press.

Sterne, J.A.C., and Davey Smith, G. (2001), "Sifting the Evidence—What's Wrong with Significance Tests?" *British Medical Journal*, 322, 226–231.

Stigler, S.M. (1986), *The History of Statistics*, Cambridge, MA: Belknap Press.

Thompson, B. (2004), "The 'significance' Crisis in Psychology and Education," *The Journal of Socio-Economics*, 33, 607–613.

Thompson, W.D. (1987), "Statistical Criteria in the Interpretation of Epidemiologic Data," *American Journal of Public Health*, 77, 191–194.

Trafimow, D., and Marks, M. (2015), Editorial, *Basic and Applied Social Psychology*, 37, 1–2.

Wagenmakers, E.-J. (2007), "A Practical Solution to the Pervasive Problem of p Values," *Psychonomic Bulletin Review*, 14, 779–804.

Walker, A.M. (1986), "Reporting the Results of Epidemiologic Studies," *American Journal of Public Health*, 76, 556–558.

Ware, J.H., Mosteller, F., and Ingelfinger, J.A. (2009), "p-Values," in *Medical Uses of Statistics* (3rd ed.), Bailar, J.C. and Hoaglin, D.C. (eds.), Hoboken, NJ: Wiley, pp. 175–194.

Weinberg, C.R. (2001), "Its Time to Rehabilitate the P-Value," *Epidemiology*, 12, 288–290.

Whitehead, A. (2002), *Meta-Analysis of Controlled Clinical Trials*, New York: Wiley.

Wood, J., Freemantle, N., King, M., and Nazareth, I. (2014), "Trap of Trends to Statistical Significance: Likelihood of Near Significant P Value Becoming More Significant with Extra Data," *BMJ*, 348, g2215, doi:10.1136/bmj.g2215.

Xu, L., Freeman, G., Cowling, B.J., and Schooling, C.M. (2013), "Testosterone Therapy and Cardiovascular Events Among Men: A Systematic Review and Meta-analysis of Placebo-Controlled Randomized Trials," *BMC Med.*, 11, 108.

Yates, F. (1951), "The Influence of *Statistical Methods for Research Workers* on the Development of the Science of Statistics," *Journal of the American Statistical Association*, 46, 19–34.

You ,B., Gan, H.K., Pond, G., and Chen, E.X. (2012), "Consistency in the Analysis and Reporting of Primary End Points in Oncology Randomized Controlled Trials from Registration to Publication: A Systematic Review," *Journal of Clinical Oncology*, 30, 210–216.

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: U Michigan Press.

Comment

Naomi S. ALTMAN

P-values and power estimates are both required in order to understand reproducibility of results. Improving the power of a study improves both the false discovery and false nondiscovery rates at any p-value threshold. An idea from the multiple testing literature, π_0, the proportion of truly null hypotheses among the tested hypotheses, can be used to reinterpret the p-value of a single test as the false discovery rate if the null is rejected. Use of bench-mark estimates of π_0 based on whether the hypothesis is a well-supported primary aim of the study, a secondary aim formulated as part of the study design, or a hypothesis proposed after examining the data can assist in interpreting the importance of observed p-values

KEY WORDS: FDR; π_0; Power; Reproducibility; Reproducible research.

Ideas from multiple testing of high dimensional data provide insights about reproducibility and false discovery rates of hypotheses supported by p-values.

Much of the statistical testing literature focuses on false rejection of the null hypothesis, also called false discovery. Individual p-values of prespecified hypotheses provide protection against a single false discovery if rejection of the null is done at some prespecified threshold such as $p < 0.05$. In many contexts, however, false nondiscovery (i.e. failure to reject the null hypothesis when it is false) is equally important, as it may lead to lack of follow-up of important hypotheses or failure to understand all the determinants of a system under study.

Both false positives and false negatives lead to irreproducibility of results. For example, sample sizes are often determined by the specification of "80% power (probability of false nondiscovery) when rejecting at $p < 0.05$." In this case, with two independent studies of the process both of which achieve the specified power and use rejection threshold $p < 0.05$, the probability of discordant results is 9.5% if the null hypothesis is true and 32% if it is false. The probability of two rejections in two trials when the null hypothesis is false is the square of the power—64%. For a fixed sample size, decreasing the significance threshold improves the probability of concordant results if the null hypothesis is true, but decreases the power to detect true discoveries. If the power is reduced to 70%, then the probability of two rejections in two trials in which the null distribution is false is only 49% and the probability of discordant results increases.

Improved study design, including larger sample size, increases power for any fixed significance threshold. What is less appreciated is that if some proportion π_0 of the hypotheses under test are truly null (and the remainder are truly not) then increased power reduces both the false discovery rate (FDR, the expected proportion of rejections for which the null is true) and the false nondiscovery rate (the expected proportion of failures to reject that are actually nonnull). If m independent hypotheses are tested, with significance threshold α and power β, we expect $\alpha\pi_0 m$ false discoveries and $\beta(1-\pi_0)m$ rejections, so that the FDR is $\alpha\pi_0/(\alpha\pi_0+\beta(1-\pi_0))$ which is a decreasing function of β. Similarly, the expected proportion of nondiscoveries that are false is $(1-\beta)(1-\pi_0)/((1-\alpha)\pi_0+(1-\beta)(1-\pi_0))$ which is also a decreasing function of β.

For high-dimensional data such as "omics" data, we can often obtain an estimate of π_0 from the observed p-values (e.g., Storey 2002; Pounds and Cheng 2004). This can be used to determine a significance threshold that produces a reasonable estimated FDR and is the basis behind methods such as Storey's q-value (Storey 2003).

For studies with much smaller numbers of test statistics, there are two reasonable ways to proceed using this paradigm. We could target a particular FDR (assuming known power) and determine whether the π_0 needed to achieve this FDR for the significance threshold is reasonable given what is known about the system under study or we could determine a benchmark value for π_0 and determine an appropriate threshold. For example, if we select FDR $= 0.05$, then at $\alpha = 0.05$ and $\beta = 0.80$, π_0 needs to be 46% or less, which implies that fewer than half of the hypotheses we expect to test are actually null. If we expect π_0 to be 90%, then to obtain FDR $= 0.05$, we need to use the threshold $\alpha < 0.0046$—however, recall that as α decreases so does β, which has not been accounted for here.

My suggestion for studies with too few hypotheses for estimation of π_0 is to establish some benchmark levels of π_0 which can be used to estimate the false discovery rate when rejecting with the observed p-value. This could be viewed as analogous to a proposal of Rosenthal to assess the "file drawer problem" in meta-analysis by computing the number of null results needed to overturn the proposed conclusion.

For example, for primary hypotheses from a study with adequate preliminary data, we might expect π_0 to be 50%—that is, equipoise . For exploratory results, found by multiple tests or fitting multiple models after the data were collected, we might expect π_0 to be much higher, say 95%. In principle, appropriate benchmarks could be determined for each discipline by a literature search, but lack of details in the literature about nonsignificant results might make this difficult. Instead, I would propose benchmarks of 50% for well-supported primary hypotheses, 75% for secondary hypotheses proposed when the study is first designed, and 95% for fortuitous findings and findings that require selection of covariates to attain statistical significance. The observed p-values can be converted into false discovery rates using the benchmark values and power appropriate to the hypothesis being tested. As explained in Storey (2003) the false

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Naomi S. Altman, Department of Statistics, The Pennsylvania State University, (Email: nsa1@psu.edu).

©2016 Naomi S. Altman

discovery rate associated with a hypothesis behaves much like the posterior probability that the null is true. Use of benchmark values of π_0 yields many of the good properties of posterior probabilities for interpreting test results without requiring formulation of a full Bayesian model for each hypothesis.

References

Pounds, S., and Cheng, C. (2004), "Improving False Discovery Rate Estimation," *Bioinformatics*, 20, 1737–1745.

Rosenthal, R. (1979), "The 'File Drawer Problem' and Tolerance for Null Results," *Psychological Bulletin*, 86, 638–641.

Storey, J.D. (2002), "A Direct Approach to False Discovery Rates," *Journal of the Royal Statistical Society*, Series B, 64, 479–498.

——— (2003), "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value," *Annals of Statistics*, 31, 2013–2035.

Comment: A Simple Alternative to P-Values

Daniel J. BENJAMIN and James O. BERGER

The ASA statement is excellent, and we completely endorse it. Unfortunately, there is a long history of such statements appearing and having only transient impact, because an alternative to p-values is not specified. There have been many proposed alternatives that are much more readily interpretable than p-values, but none has gained widespread acceptance. We believe that p-value alternatives have failed to garner general support because they have either or both of two shortcomings:

1. Being more complicated than p-values.
2. Not being acceptable to both frequentist and Bayesian schools of statistical thought.

Recently, in Bayarri et al. (2016), we have proposed an alternative that overcomes both of these hurdles.

Our proposal is a follow-up to suggestions (see Wellcome Trust Case Control Consortium 2007) that a researcher should report the *pre-experimental odds of correct rejection to incorrect rejection* of the null hypothesis. If, for example, these odds are 10 to 1, then a formal rejection of the null hypothesis is 10 times more likely to be a correct rejection than an incorrect rejection. But the pre-experimental odds depend partly on the researchers *prior odds* of the alternative hypothesis to the null hypothesis, which makes them non-frequentist.

The pre-experimental odds can be decomposed into the product of the prior odds and a frequentist component, which is determined by the experimental design and planned analysis. We can then focus attention on this frequentist component, which we call the *rejection ratio*. It is the probability of rejection when the alternative hypothesis is true, divided by the probability of rejection when the null hypothesis is true, i.e., the ratio of the power of the experiment to the Type I error of the experiment. The rejection ratio has a straightforward interpretation as quantifying the strength of evidence about the alternative hypothesis relative to the null hypothesis conveyed by the experimental result being statistically significant.

While the rejection ratio is an excellent summary of the quality of the experiment in terms of hypothesis testing, it would not work as an alternative to p-values because it suffers from the two drawbacks listed above. First, it is more complicated than a p-value because it requires power calculations. Second, although justified to unconditional frequentists, it is unsatisfactory to many (including Bayesians) because it is not data-dependent; if the rejection ratio were 10 based on a Type I error of $\alpha = 0.05$, then 10 would be reported regardless of whether the data yields a p-value of 0.05, right at the boundary of the rejection region, or a p-value of 0.000001, surely indicating much stronger evidence against the null hypothesis than $p = 0.05$. To a Bayesian, the correct data-dependent measure of the evidence in favor of the alternative hypothesis relative to the null hypothesis is the Bayes factor, given by

$$B = \frac{\text{average likelihood of the observed data under the alternative hypothesis}}{\text{likelihood of the observed data under the null hypothesis}} .$$

It would seem that we are now at a Bayesian/frequentist impasse, but this is not so - at least, it is not so for many common situations such as testing a null hypothesis of zero effect versus a two-sided alternative hypothesis of non-zero effect.[1] Indeed, for such situations, we show in Bayarri et al. (2016) that B, while data dependent, is a fully frequentist measure because its frequentist expectation under the null hypothesis precisely equals the frequentist rejection ratio. Bayesians and frequentists should thus unite in promoting B as an easily interpretable alternative to p-values, overcoming problem #2 above.

Unfortunately, B still suffers from problem #1 because the 'average likelihood' in the numerator of B needs to be computed using some assumed prior distribution for the alternative hypothesis. There is, however, a simple upper bound on B (Vovk 1993) that holds under quite general conditions (Sellke et al. 2001), namely

$$B \leq \overline{B} \equiv \frac{1}{-e\ p \log p} . \tag{1}$$

The Bayes factor bound $\overline{B}$ is the largest possible B over any (reasonable) choice of the prior distribution for the alternative hypothesis. Like B, $\overline{B}$ overcomes problem #2: it is justifiable to both Bayesians and frequentists. However, $\overline{B}$ additionally surmounts problem #1 because it is a simple function of the p-value.

Daniel J. Benjamin, University of Southern California, Los Angeles, California 90089-0001. James O. Berger, Duke University, Durham, North Carolina 27708-0187 (Email: berger@stat.duke.edu). This research was supported by the National Science Foundation, under grants DMS-1007773, DMS-1407775, and BCS-1521855, and the National Institutes of Health / National Institute on Aging under grant R01-AG042568.

[1] More generally, our recommendation here applies to any situation of *precise hypothesis testing*, by which we mean that the null hypothesis is a lower dimensional subspace of the alternative hypothesis, as in testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The major problems with p-values are muted if the hypotheses are, say, $H_0 : \theta < 0$ versus $H_1 : \theta > 0$, which are of the same dimension. As an example, suppose θ denotes the difference in mean treatment effects for cancer treatments A and B:

- Scenario 1: Treatment A = standard chemotherapy and Treatment B = standard chemotherapy + steroids. This is a scenario of precise hypothesis testing, because steroids could be essentially ineffective against cancer, so that θ could quite plausibly be essentially zero.
- Scenario 2: Treatment A = standard chemotherapy and Treatment B = a new radiation therapy. In this case there is no reason to think that θ could be zero, and it would be more appropriate to test $H_0 : \theta < 0$ versus $H_1 : \theta > 0$.

See Berger and Mortera (1999) for discussion of these issues. Moreover, in precise hypothesis testing situations that are one-sided, such as $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, $\overline{B}$ is no longer strictly an upper bound for B (although the deviations tend to be minor; Sellke (1977)).

©2016 Daniel J. Benjamin and James O. Berger

The following table shows the value of $\overline{B}$ for a wide range of p-values.

p	0.1	0.05	0.01	0.005	0.001	0.0001	0.00001
$\frac{1}{-ep\log(p)}$	1.60	2.44	8.13	13.9	52.9	400	3226

An important implication of these calculations is that results that just reach conventional levels of significance do not actually provide very strong evidence against the null hypothesis. For example, a p-value of 0.05 corresponds to a Bayes factor of at most 2.44 : 1. That is, the data imply odds in favor of the alternative hypothesis relative to the null hypothesis of at most 2.44 to 1. A p-value of 0.01 - often considered "highly significant"—corresponds to at most 8.13 to 1 odds, hardly overwhelmingly convincing odds.

Since $\overline{B}$ indicates the strongest potentially justifiable inference from the data, its use would alert researchers when seemingly strong evidence is actually not very compelling. Its use would therefore help prevent researchers from being misled into concluding too much from the statistical significance of a finding. Interestingly, although $\overline{B}$ is only an upper bound on the Bayes factor, we report evidence in Bayarri et al. (2016) that, when calculated from real data from a range of scientific fields, $\overline{B}$ is often not that far from the B implied by a scientifically reasonable alternative hypothesis.

In short, to begin the process of recovering from the misuse of p-values, we propose replacing the p-value by the Bayes factor bound $\overline{B}$. $\overline{B}$ is easily interpretable, has both Bayesian and frequentist justification, and is as simple to calculate as the p-value. Of course, we would encourage everyone to "bite the bullet" and use the more sophisticated B—again, also a fully frequentist measure—and even to go one step further by multiplying B by the prior odds to obtain the overall odds of the alternative hypothesis to the null hypothesis. But even just the initial, small step of reporting $\overline{B}$ would help researchers avoid some of the most problematic and apparently inevitable misinterpretations that arise from reliance on the p-value.

REFERENCES

Bayarri, M.J., Benjamin, D., Berger, J., and Sellke, T. (2016), "Rejection Odds and Rejection Ratios: A Proposal for Statistical Practice in Testing Hypotheses," to appear in *Journal of Mathematical Psychology*.

Berger, J., and Mortera, J. (1999), "Default Bayes Factors for Non-nested Hypothesis Testing," *Journal of the American Statistical Association*, 94, 542–554.

Sellke, T., Bayarri, M.J., and Berger, J.O. (2001), "Calibration of p Values for Testing Precise Null Hypotheses, *The American Statistician*, 55, 62–71.

Sellke, T.M. (2012), "On the Interpretation of p-Values," Tech. Rep. Department of Statistics, Purdue University.

Vovk, V.G. (1993), "A Logic of Probability, with Application to the Foundations of Statistics," *Journal of the Royal Statistical Society*, Series B, 55, 317–351.

Wellcome Trust Case Control Consortium (2007), "Genome-Wide Association Study of 14,000 Cases of Seven Common Diseases and 3,000 Shared Controls," *Nature*, 447(7145), 661–678.

It's Not the P-Values' Fault

Yoav BENJAMINI

I argue that ASA board statement about the p-values may be read as discouraging the use of p-values because they can be misused, while the other approaches offered there might be misused in much the same way. In particular, ignoring the effect of selection on statistical inferences is common yet potentially very harmful to the replicability of research results.

KEY WORDS: ASA board; Industrialized science; Selective inference.

When I was invited to participate in ASA committee, my initial response was that it would be better for the committee to draft a statement about the appropriate use of statistical tools for addressing the crisis of reproducibility and replicability (R&R) in science. Unfortunately, in response to outcries about the role of Statistics, which focused on the perceived role of the widely used p-values, the ASA board fell into the trap of formulating a statement about the p-values. The well-phrased statement demonstrates our mistake in singling out the p-value: posing the p-value as a culprit, rather than the way most statistical tools are used in the new world of industrialized science.

Admittedly, most statisticians reading this statement will agree with most of its principles (Bayesians may not agree to principle 1, frequentists will have difficulties understanding principle 6), but all principles stated are only about p-values and statistical significance. The result is a statement that will be read by our target audience as expressing very negative ASA attitude towards the p-value. As stated, the p-value "can be useful" providing "one approach to summarizing the incompatibility between a particular set of data and a proposed model for the data" (Principle 1).

On the other hand:

Principle 2: "*p-values do not measure the probability*...;

Principle 3: "scientific decisions *should not be based only on whether a p-value passes a specific threshold*" as this "...*leads to considerable distortion of the scientific process*";

Principle 4: "*P-values and related analyses should not be reported selectively*";

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

Principle 6: " ...a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis"

Nonstatistical scientists, editors, policy makers or a judges, who read these principles will conclude that the p-value is indeed a very risky statistical tool, as advertised by its opponents. Avoiding its use and discouraging its use by others is just a matter of common sense. This will be the case especially since the ASA statement offers *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.*"

Yet all of these other approaches, as well as most statistical tools, may suffer from many of the same problems as the p-values do. What level of likelihood ratio in favor of the research hypothesis will be acceptable to the journal? Should scientific discoveries be based on whether posterior odds pass a specific threshold (P3)? Does either measure the size of an effect (P5)? Isn't our best effect size estimator useless as a single measure if not supported by a statement about its uncertainty? How can we decide about the sample size needed for a clinical trial—however analyzed—if we do not set a specific bright-line decision rule? Finally, 95% confidence intervals or credence intervals (both sharing the limitations in P2) offer no protection against selection when only those that do not cover 0, are selected into the abstract (P4).

What made the p-value so useful and successful in Science throughout the 20th century, despite of the misconceptions so well described in the statement? In some sense it offers a first line of defense against being fooled by randomness, separating signal from noise, because the models it requires are simpler than any other statistical tool needs. Likelihood ratios, effect size estimates, confidence intervals, and Bayesian methods all rely on assumed models over a wider range of situations, not merely under the tested null; Bayesian tools need further modeling, in the form of priors and hierarchical structures. Most important, the model needed to calculate of the p-value can be guaranteed to hold under appropriately designed and executed randomized experiments.

The p-value is a very valuable tool, but when possible it should be complemented—not replaced—by confidence intervals and effect size estimates. The end of a 95% confidence interval that extends towards 0 indicates by how much the difference can be separated from 0 (in a statistically significant way at level 5%...). The mean difference, when supported by an assessment of uncertainty is again useful. Disappointingly, in some areas of science these methods are grossly underutilized.

Sometimes, especially when using emerging new scientific technologies, the p-value is the only way to quantify uncertainty, and can be mapped and compared across conditions (e.g.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Yoav Benjamini, Tel Aviv University, Ramat Aviv, Tel Aviv 69978, Israel (Email: ybenja@post.tau.ac.il). The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement n° [294519] PSARPS. The analysis of the papers of the Reproducibility Project was led by Yoav Zeevi, with Sofi Estashenko, Tal Galili, Tal Kozlovski, and Dan Warshavski (see replicability.tau.ac.il). Thanks to Henry Braun, Abba Krieger, and Philip Stark for comments.

©2016 *Yoav Benjamini*

functional MRI, Gene Expression, Genome Wise Association Studies). It is recognized that merely "full reporting and transparency" (Principle 4) is not enough, as selection is unavoidable in these large problems. Selection takes many forms: selection by a table, selection into the abstract, selection by highlighting in the discussion, selection into a model, or selection by a figure. Further statistical methods must be used to address the impact of selective inference, otherwise the properties each method has on the average for a single parameter (level, coverage or unbiasedness) will not hold even on the average over the selected parameters. Therefore, in those same areas, the p-value bright-line is not set at the traditional 5% level. Methods for adaptively setting it to directly control a variety of false discovery rates or other error rates are commonly used. More generally, addressing the effect of selection on inference has been a very active research area, resulting in new strategies and sophisticated tools for testing, confidence intervals, and effect-size estimates, in different setups. It deserves a separate review.

The transition in large complex problems illustrates the process occurring throughout science: the industrialization of the scientific process at the turn of the century. Experimentation is done by high throughput industrial processes and their outcomes are analyzed automatically, resulting in a large number of inferences to select from. With the availability of ever-larger databases and the ease of computations, other areas of science are undergoing similar industrialization processes, yet are slow to realize these changes. For example, the estimated number of *reported* inferences in the 100 studies included in the "reproducibility project" in Experimental Psychology (Open Science Collaboration, 2015) range from 5 to 730, with an average of 77 (± 10) per study. We currently study the actual selection process in these complex studies (rather than merely counting) but it is enough to note that only 11 studies included any partial effort to address selection. Facing such ignorance I prefer to eyeball a set of p-values to assess the effect of selection rather than view a set of confidence intervals.

In summary, when discussing the impact of statistical practices on R&R, the p-value should not be singled out nor its use discouraged: its more likely the fault of selection, not the p-values'.

Reference

Open Science Collaboration (2015), "Estimating the Reproducibility of Psychological Science," *Science*, 349, aac4716. DOI: 10.1126/science.aac4716.

P-Values Are Not What Theyre Cracked Up to Be

Donald A. Berry

The ASA is to be congratulated for its "Statement on Statistical Significance and P-values." Much has been written about p-values in the last 50 years. Many authors have been critical, with pointed warnings about misunderstanding and misinterpreting these strange but ubiquitous beasts. The cumulative impact of such criticisms on statistical practice and on empirical research has been minimal to none. Surprisingly, although statisticians can correctly define p-values and they properly struggle to not overestimate the extent of confidence one can have in a confidence interval, most statisticians do not really understand the issues in applied settings.

Recent attacks on p-values and the role of statistical significance in the "crisis of irreproducibility" has highlighted our lack of understanding. Our collective credibility in the science community is at risk. We cannot excuse ourselves by blaming non-statisticians for their failure to understand or heed what we tell them. The fault for widespread ignorance about statistical significance and for the misuses by substantive scientists of measures we promulgate is ours alone. We must communicate better even if we have to scream from the rooftops, which is exactly what the ASA is doing.

More important than the credibility of our discipline is the impact that misuse and misinterpretation of statistical significance and p-values has on science and society. Patients with serious diseases have been harmed. Researchers have chased wild geese, finding too often that statistically significant conclusions could not be reproduced. The economic impacts of faulty statistical conclusions are great.

The effects extend to the public and affect the lay persons understanding and appreciation of science. For example, anybody who leafs through newspapers has seen many studies showing statistically significant health effects of drinking coffee, with each study contradicting many earlier studies. The public learns to not believe "studies." Count me among them.

The ASA's statement will herald a statistical renaissance. Every statistician must take notice. Statisticians who think they know better than the ASA are wrong. And no teacher of an introductory statistics course can pretend to represent modern statistical philosophy and practice without discussing this statement with their students. An effective tack would be to provide them with examples from the research literature, bad examples as well as good—with my warning that it will be hard to find the latter!

Statistics texts define p-values and show how to calculate them in particular examples assuming a particular statistical model. But they fail to address the confusion and mayhem that these measures cause in practical applications. Texts accentuate the positive. They do not consider applied problems with conclusions of statistical significance when the p-value is less than 0.05, say, but have no inferential content, are scientifically meaningless, and cannot be reproduced.

There is little controversy regarding interpreting p-values as summary statistics for a particular set of data. P-values are handy measures of extremity and serve to describe a set of numbers in a way similar to that of Z-scores and confidence intervals. Errors occur when attributing scientific import to a p-value. For instance, researchers may claim that a small p-value is evidence against the null hypothesis that a treatment is ineffective. The standard Bayesian non-informative-prior data-analytic approach is similar to using p-values for inference but is potentially more dangerous because it ostensibly concludes with a posterior probability of truth.

I will expand here on Principles 1 and 4 of the ASA statement.

The statement gives this "informal" definition: "a p-value is the probability under a specified statistical model that a statistical summary of the data . . . would be equal to or more extreme than its observed value." Similarly, Principle 1 indicates that p-values "can indicate how incompatible the data are with a specified statistical model." Yes, but there are subtleties: choice of statistical model; interpretations of "the data;" deciding what is extreme.

Statisticians are trained to analyze numbers. Suppose you provide a statistician with a spreadsheet containing outcomes for a particular experimental treatment vs control and you request a p-value. To set up a model the statistician may ask about the stopping rule, about whether the two samples were independently collected, and about any covariates. After deciding whether to transform the data the statistician calculates a p-value using a test based on some assumed form of the distribution of outcomes or taking a nonparametric approach. What does the p-value mean regarding whether the treatment is better than control? Not much.

Such a p-value is a descriptive summary of a dataset but it has no inferential content. The critical issue is the interpretation of "the data" in the p-value definition. Inferences require a broader interpretation of data than one based on numbers alone. My dictionary says data are "things known or assumed as facts, making the basis of reasoning or calculation." P-values ignore many aspects of the evidence in the experiment at hand including information that is obviously known. One important piece of data is the simple fact that you gave the statistician the spreadsheet and requested a p-value. Why did you do that? Had you noticed something unusual about the outcomes? Had you requested p-values for the same data from other statisticians and didn't like their answers?

The specifics of data collection and curation and even your intentions and motivation are critical for inference. What have you not told the statistician? Have you deleted some data points

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Donald A. Berry, Department of Biostatistics, The University of Texas M.D. Anderson Cancer Center, 1400 Pressler Street, 4-5062 Pickens Academic Tower, Houston, TX 77030-1402 (Email: dberry@mdanderson.org).

©2016 Donald A. Berry

or experimental units, possibly because they seemed to be outliers? Are some entries actually the average of two or more measurements made on the same experimental unit? If so, why were there more measurements on some units than on others? Have you conducted other experiments addressing the same or related questions and decided that this was the most relevant experiment to present to the statistician? And on and on and on.

The answers to these questions may be more important for making inferences than the numbers themselves. They set the context for properly interpreting the numerical aspects of the "data." Viewed alone, p-values calculated from a set of numbers and assuming a statistical model are of limited value and frequently are meaningless.

How can one incorporate the answers to questions such as those above into a statistical analysis? Standard Bayesian data-analytic measures have the same fundamental limitation as p-values. Subjective Bayesian approaches have some hope, but exhibiting a full likelihood function for nonquantifiable data may be difficult or impossible. As a practical matter, when I worry that I dont know enough about the extra-numerical aspects of the "data" or about the possibility of incorporating this information into a quantitative measure of evidence then I resort to a "black-box warning:"

"Our study is exploratory and we make no claims for generalizability. Statistical calculations such as p-values and confidence intervals are descriptive only and have no inferential content."

When is it appropriate to use p-values for inference? An archetype is drug regulation. Drug sponsors must develop a protocol and a statistical analysis plan in advance of an experiment. These explicitly and unambiguously state the primary endpoint and how it will be analyzed. After the experiment a robot could calculate the p-value.

Principle 4 in the ASA statement is that "Proper inference requires full reporting and transparency and multiplicities." When there is a prospective study protocol and statistical analysis plan then both should be made available at the time of publication along with any deviations from the original plans. In the absence of a protocol and statistical analysis plan the credibility of conclusions is low, despite honest attempts to say what analyses had been planned, whether done or not, and what planned analyses were not done. And adjusting for associated multiplicities may be difficult in this circumstance. A pragmatic approach is to completely describe the multiplicities, keeping a log of what was done, and then giving "unadjusted" p-values, including a black-box warning similar to the one above.

The "p-value dilemma" is entwined with the bigger problems of global multiplicity and irreproducible research. Drug development is an example. Thousands of drugs are being developed worldwide. Each developer conducts clinical trials to decide whether their drug merits further development. The trials may have completely prospective protocols that are followed meticulously. Development continues into the next phase if $Z > 1.65$, say. Some drugs proceed and some do not. For those that proceed, regression to the mean sets in and most drugs fail in the next phase. A Bayesian analysis can accommodate historical information regarding other drugs to suitably regress the results of any particular clinical trial. But speaking as someone who does that, it is not easy to persuade a developer that their next trial is unlikely to be as promising as the present one! And its not just the developer who is duped. For example, even though they have lots at stake, venture capitalists over-interpret the present trial's p-value and they have trouble understanding that they cannot take the observed data at face value.

Irreproducible research is a huge problem in science and medicine. Statisticians are well positioned to teach other scientists about reproducibility of research, or lack thereof. However, most statisticians are as naïve in this regard as the scientists themselves. Newly minted statisticians tend to regard p-values as relevant scientifically and interpret statistical significance found from processing a spreadsheet of numbers as being reproducible 95% of the time. Only the cold water of experience teaches them otherwise. Again, the remedy is education. We must change the way statisticians are trained. They will, in turn, retrain the rest of the world.

In brief, p-values are not what they're cracked up to be. They serve to describe a dataset of numbers and in that sense they are useful tools. But the vast majority of small p-values do not deserve the label "statistically significant" and they do not imply any other type of scientific relevance. The ASA statement is a bold attempt to right previous misunderstandings in this regard.

Are there better approaches to inference than using p-values, in clinical research say? Absolutely. But that has not been my focus here. It is important to use any tool correctly, especially if we hope to improve it.

Comment: Is Reform Possible Without a Paradigm Shift?

John B. CARLIN

Looking back over the ASA Statement on Statistical Significance and P-values, which I think is an important and valuable contribution to a real and pressing problem, a striking feature is that so much effort appears to be needed to *counteract* the misuse and misinterpretation of what must have seemed to Fisher and other pioneer statisticians to be a simple set of ideas. Of course the originators of the concepts of the p-value and its far more invidious offshoot, the practice of reducing empirical comparisons to declarations of "statistical significance" (or otherwise), could not have anticipated how these ideas would become so embedded in the practice of nonstatistical researchers. As has been pointed out and will continue presumably to be analysed by the philosophers and historians of science, there seems to be an irresistible urge to encode scientific conclusion-making into a rule-based activity (Gigerenzer and Marewski 2015). This always seems curious to me because it appears obvious that conclusions about the empirical world can only be made tentatively (beware the black swan!). Thus inductive inference must always be couched in a language of uncertainty, in contrast with which the familiar phraseology of statistically based research ("an association was found; $P < 0.05\ldots$", "no effect was observed") just doesn't make sense.

Can the general scientific usage of statistical inference methods be reformed? The outright ban on traditional tools such as the p-value and confidence interval by the journal *Basic and Applied Social Psychology* (Trafimow and Marks, 2015) has achieved some positive outcomes, by way of much broader recognition and discussion of the underlying problems (Ashworth 2015). It is less clear whether the quality of scientific inference within the pages of the journal has improved (Lakens 2016).

I believe that fundamental improvement will only be possible if and when we can agree on some broad principles about the inference task. In particular, we need to cultivate a viable language of uncertainty that is primarily focused on expressing uncertain knowledge conditional on observed data (Morey et al. 2016). To my mind the only general language that seems to have any reasonable chance of fulfilling this goal is the Bayesian use of probability. Genuine post-data conclusions seem to be possible only within a Bayesian or perhaps closely related paradigm.

However, I am well aware of the difficulties of this path. The overwhelming challenge is that as soon as we enter this paradigm we appear to require god-like knowledge of the "true" or at least "appropriate" model that should be specified, including prior distributions that will be needed to kick-start the uncertainty calculus. The difficulty of creating broadly accepted conventions for how models should be specified and checked before any conclusions based on their application to the data at hand may be trusted often seems insuperable, despite some suggested strategies (Gelman and Shalizi 2013). To many, the danger of the rules becoming even more malleable—and so even more likely to allow researchers to put arbitrary stamps of statistical authority on the conclusions they would like to draw—under this paradigm than under the traditional muddled modes of p-value-based inference outweigh its compelling inherent logic.

Although I understand this point of view I just don't see any real choice, as no one seems to be coming close to proposing a way of salvaging the traditional muddle. Indeed in a companion paper to this discussion (Greenland et al. 2016) we see a long list of misconceptions and misinterpretations, which it can be hoped scientists may start to avoid. Yet the length and complexity of the list itself suggests that the fundamental ideas that it seeks to clarify are so convoluted—and inherently unsuited to the task of uncertain inductive post-data inference—that a solution might only be possible with a more fundamental paradigm shift.

References

Ashworth, A. (2015), "Veto on the Use of Null Hypothesis Testing and p intervals: Right or Wrong?" Taylor Francis Editor Resources [online], available at *http://editorresources.taylorandfrancisgroup.com/veto-on-the-use-of-null-hypothesis-testing-and-p-intervals-right-or-wrong/*, accessed Feb. 29, 2016.

Gelman, A., and Shalizi, C. (2013), "Philosophy and the Practice of Bayesian Statistics" (with discussion), *British Journal of Mathematical and Statistical Psychology* 66, 8–80.

Gigerenzer, G., and Marewski, J.N. (2015), "Surrogate Science: The Idol of a Universal Method for Scientific Inference," *Journal of Management*, 41, 421–440.

Greenland, S., Senn, S.J., Rothman, K.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," *The American Statistician*, online supplement to "ASA Statement on Statistical Significance and P-Values," *The American Statistician*, 70.

Lakens, D. (2016), "So You Banned p-Values, How's that Working Out for You?" *The 20% Statistician: A Blog on Statistics, Methods and Open Science* [online], available at *http://daniellakens.blogspot.com.au/2016/02/so-you-banned-p-values-hows-that.html*, accessed Feb. 29, 2016.

Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D., and Wagenmakers, E.-J. (2016), "The Fallacy of Placing Confidence in Confidence Intervals," *Psychonomic Bulletin & Review*, 23, 103–123.

Trafimow, D., and Marks, M. (2015), Editorial, *Basic and Applied Social Psychology*, 37, 1–2.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. John Carlin is Director, Data Science Core and Clinical Epidemiology & Biostatistics Unit (CEBU), Murdoch Children's Research Institute & University of Melbourne Department of Paediatrics, and is Professorial Fellow, Centre for Epidemiology & Biostatistics, Melbourne School of Population & Global Health, University of Melbourne, Melbourne, Australia (Email: john.carlin@mcri.edu.au).

©2016 John B. Carlin

ASA Statement on P-Values: Two Consequences We Can Hope For

George Cobb

First off, I applaud those who put together the final version of the ASA statement, and applaud also those on the Board who approved it. By way of salute, and to highlight the challenges they faced, I pose a variant of the Birthday Problem: "How many statisticians does it take to ensure at least a 50% chance of a disagreement about p-values?" Regardless of whether your answer is 2 or 1, the question itself stands as testimony to what ASA has accomplished here in creating a consensus statement.

In addition to my salute, I offer two thoughts:

1. *What ASA has done here can set a good precedent for our organization.* Data science may be broader in scope than statistics, may be growing faster than statistics, and may be a shinier object in the eye of the media, but – to borrow a metaphor from John Kennedy: *If you purge data science of everything statistical you could bury the remains in a shoebox.*

The ASA statement on p-values is, in my opinion, an important example of how our profession can assert its relevance—to the analysis of data and to the progress of science. I would urge the Board to set aside time at least once every year to consider the potential value of similar statements. The key questions of which issues to discuss, and whether to pursue them, should be up to the Board, but two possibilities are (a) the evolving role of statistics in data science, and (b) scientific applications of Bayesian methods.

2. *What ASA has done here should spur a reshaping of the way we teach—both p-values in particular, and statistics generally*. The teaching of statistics is still in metamorphosis, no longer cocooned in its silken heritage of mathematics, but still drying its wings before taking full flight. Our introductory curriculum remains too much blinkered by its lingering obeisance to formal classical inference via hypothesis tests and confidence intervals. (Even the newer, simulation-based approaches, which I am both proud and guilty of having abetted, pay too much attention to probability-based inference.) As in the past, what we preach lags behind what we practice. The ASA statement is a useful summary of the pros and cons of p-values in practice. We should adjust our preaching to match.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. George Cobb, Mount Holyoke College South Hadley, MA (Email: gcobb@mtholyoke.edu).

©2016 George Cobb

The Problems With P-Values are not Just With P-Values

Andrew GELMAN

The ASA's statement on p-values says, "Valid scientific conclusions based on p-values and related statistics cannot be drawn without at least knowing how many and which analyses were conducted." I agree, but knowledge of how many analyses were conducted etc. is not enough. The whole point of the "garden of forking paths" (Gelman and Loken 2014) is that to compute a valid p-value you need to know what analyses *would have been done* had the data been different. Even if the researchers only did a single analysis of the data at hand, they well could've done other analyses had the data been different. Remember that "analysis" here also includes rules for data coding, data exclusion, etc.

When I was sent an earlier version of the ASA's statement, I suggested changing the sentence to, "Valid p-values cannot be drawn without knowing, not just what was done with the existing data, but what the choices in data coding, exclusion, and analysis would have been, had the data been different. This 'what would have been done under other possible datasets' is central to the definition of p-value." The concern is not just multiple comparisons, it is multiple *potential* comparisons.

Even experienced users of statistics often have the naive belief that if they did not engage in "cherry-picking . . . data dredging, significance chasing, significance questing, selective inference and p-hacking" (to use the words of the ASA's statement), and if they clearly state how many and which analyses were conducted, then they're OK. In practice, though, as Simmons, Nelson, and Simonsohn (2011) noted, researcher degrees of freedom (including data-exclusion rules; decisions of whether to average groups, compare them, or analyze them separately; choices of regression predictors and iteractions; and so on) can be and are performed after seeing the data.

A *scientific* hypothesis in a field such as psychology, economics, or medicine can correspond to any number of *statistical* hypotheses, and if the ASA is going to issue a statement warning about p-values, I think it necessary to emphasize that researcher degrees of freedom—the garden of forking paths—can and does occur even without people realizing what they are doing. A researcher will see the data and make a series of reasonable, theory-respecting choices, ending up with an apparently successful—that is, "statistically significant"—finding, without realizing that the nominal p-value obtained is meaningless.

Ultimately the problem is not with p-values but with null-hypothesis significance testing, that parody of falsificationism in which straw-man null hypothesis A is rejected and this is taken as evidence in favor of preferred alternative B (see Gelman 2014). Whenever this sort of reasoning is being done, the problems discussed above will arise. Confidence intervals, credible intervals, Bayes factors, cross-validation: you name the method, it can and will be twisted, even if inadvertently, to create the appearance of strong evidence where none exists.

What, then, can and should be done? I agree with the ASA statement's final paragraph, which emphasizes the importance of design, understanding, and context—and I would also add measurement to that list.

What went wrong? How is it that we know that design, data collection, and interpretation of results in context are so important—and yet the practice of statistics is so associated with p-values, a typically misused and misunderstood data summary that is problematic even in the rare cases where it can be mathematically interpreted?

I put much of the blame on statistical education, for two reasons.

First, in our courses and textbooks (my own included), we tend to take the "dataset" and even the statistical model as given, reducing statistics to a mathematical or computational problem of inference and encouraging students and practitioners to think of their data as given. Even when we discuss the design of surveys and experiments, we typically focus on the choice of sample size, not on the importance of valid and reliable measurements. The result is often an attitude that any measurement will do, and a blind quest for statistical significance.

Second, it seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty, an "uncertainty laundering" that begins with data and concludes with success as measured by statistical significance. Again, I do not exempt my own books from this criticism: we present neatly packaged analyses with clear conclusions. This is what is expected—demanded—of subject-matter journals. Just try publishing a result with $p = 0.20$. If researchers have been trained with the expectation that they will get statistical significance if they work hard and play by the rules, if granting agencies demand power analyses in which researchers must claim 80% certainty that they will attain statistical significance, and if that threshold is required for publication, it is no surprise that researchers will routinely satisfy this criterion, and publish, and publish, and publish, even in the absence of any real effects, or in the context of effects that are so variable as to be undetectable in the studies that are being conducted (Gelman and Carline 2014).

In summary, I agree with most of the ASA's statement on p-values but I feel that the problems are deeper, and that the solution is not to reform p-values or to replace them with some other statistical summary or threshold, but rather to move toward a greater acceptance of uncertainty and embracing of variation.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Andrew Gelman, Columbia University 1016 Social Work Building New York New York 10027 (Email: gelman@stat.columbia.edu).

©2016 Andrew Gelman

References

Gelman, A. (2014), “Confirmationist and Falsificationist Paradigms of Science,” Statistical Modeling, Causal Inference, and Social Science blog, 5 Sept. *http://andrewgelman.com/2014/09/05/confirmationist-falsificationist-paradigms-science/*.

Gelman, A., and Carlin, J. B. (2014), “Beyond Power Calculations: Assessing Type S (sign) and Type M (magnitude) Errors,” *Perspectives on Psychological Science*, 9, 641–651.

Gelman, A., and Loken, E. (2014), “The Statistical Crisis in Science,” *American Scientist*, 102, 460–465.

Simmons, J., Nelson, L., and Simonsohn, U. (2011), “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allow Presenting Anything as Significant,” *Psychological Science*, 22, 1359–1366.

The Next Questions: Who, What, When, Where, and Why?

Steven N. Goodman

There is little doubt that the ASA statement will stand the test of time better than many of the textbooks and journal articles purporting to teach the practice of statistics to scientists. And that's what we should be afraid of. As the introduction acknowledges, almost none of the statement is new. That is an understatement; some of these same principles were stated or argued about at least a century ago, with many reminders between then and now. So the question we must ask of ourselves is how could this have happened, and what can we do to change it? How could it be, almost 100 years after the formulation of the Fisherian and Neyman-Pearson approaches to statistical inference, that a body is eminent as the ASA needs to step in and remind scientists how to define an index as ubiquitous in scientific investigations as the P-value, and how to use it or other indices properly? And what will prevent us from dusting this same statement off 100 years hence, to remind the community yet again of how to do things right? If we dont start focusing on that, the most likely outcome of this effort is that the statement will literally stand the test of time, in being as needed in the next century as it is today.

I suspect that for readers outside the statistical community, the most surprising thing about the statement will be the degree of passionate commentary that it will no doubt engender. The fact that statisticians do not all accept at face value what most scientists are routinely taught as uncontroversial truisms will be a shock to many. But if we are to move science forward, we must speak to scientists. How many of these comments can be understood by the broad scientific community? If that proportion is small, we have a clue why we are in the situation we are in, and what needs to change.

In that spirit, I would like to pose those fantastically useful questions that the media is trained to ask in all their reporting: *Who, what, when, where, and why*? These questions can help force us to explore our own responsibility in not just articulating principles, but helping scientists to get things right.

Who?—Exactly who will push this agenda forward? How much can statisticians do alone, and how much must they do in partnership with other scientists? And when we say some statisticians alone, do we mean the thought leaders who write mainly about theory, or the applied statisticians in the field? Or the textbook writers? Or those who labor to teach statistics within specific disciplines? Or the department chairs? How much should we include other quantitative scientists and non-statistical scientists as allies and indeed champions?

What?—The question is two-fold. First, what are we actually recommending scientists do? We say in the statement that they should combine evidence from p-values, confidence intervals, Bayesian measures, false-discovery rates, or other measures with features of the design, conduct, and plausibility of what we are studying. Exactly how our scientists supposed to do that? Where are all the textbook examples? Where are the examples in the published literature? The statement is very good at identifying bad inferential behavior, but can we use it to figure out or dictate what is good? The fact is, bad behavior is often condoned and even encouraged within many scientific disciplines, and good behavior is neither taught nor role-modeled. What do we think would happen to a statistical colleague who tried to convince their collaborator that the $P = 0.01$ just obtained for a main finding was not sufficient to make a claim, because of the design of the experiment, the analysis, or the nature of the hypothesis? Where are the examples they can point to? If we are to make such recommendations, we need to figure out what to tell or teach people.

The second "what" is what is the profession and professionals of statistics supposed to do to maximize the adoption of these principles? Write new books? New software? Change their courses? Write in disciplinary journals? Write more blogs? Post YouTube videos? Tweet more tweets? Reward good behavior? All of the above, or A,B and D? How do we promote the complete and transparent reporting that is recommended, not just in spirit, but in reality?

When? —When do we start? For how long? How will we evaluate when enough has been accomplished?

Where?—This overlaps with the "what?" In what venues and platforms does the agenda have to be pressed forward? In funding review sections? In journal clubs? On journal pages? Which journals? At the ASA? In the halls of congress? In industry? In the OSTP (Office of Science Technology Policy), the NSF, the NIH, the FDA or the DOE? Where are the key policy leverage points, and what should those policies be? Is this amenable to a policy-level fix?

Why?—Why do this? We are all busy; who has time to reform science? Or even their co-author? Or a post-doc who asks for statistical advice the evening before the conference abstract is due? One big "why" is related to research reproducibility, which is the ring by which we can turn the nose of science. When the directors of the National Academy of Sciences, the NSF, and the NIH, as well as industry leaders all tell us that research reproducibility is high on their agenda to improve, and we can see the clear connections between the ASA principles and those concerns, we have at least our collective "why," but perhaps not the individual "why?" The ASA and other groups need to figure out how to push this issue forward in a way that statisticians see as part of their professional responsibility to further this agenda,

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Steven N. Goodman, Stanford University, Division of Epidemiology, and Meta-research Innovation Center at Stanford (METRICS) (Email: Steve.goodman@stanford.edu).

©2016 Steven N. Goodman

and to reward them for it. Perhaps some awards from the ASA for improving the practice of statistics along the lines of this policy statement could provide the role models and inspirations for others to emulate.

There are many more things that could be said about what needs to be done and who needs to do it, but this much should be clear; what follows this statement is as or more important than the statement itself. It must be someone's or some entity's designated or adopted responsibility to carry the essentials of the ASA principles forward, otherwise nothing will change. We need to formulate a vision of what success looks like, and how we will get there. If not, we can start drafting the language of the 2116 ASA statement tomorrow.

The ASA Guidelines and Null Bias in Current Teaching and Practice

Sander GREENLAND

The ASA statement is a step forward for the profession that I am happy to endorse. Nonetheless, as a compromise among many conflicting views it is bound to leave many readers dissatisfied in some respects. I should like to express my chief concern via this quote from Neyman (1977, p. 106; emphasis added) discussing a hypothetical example of a possibly carcinogenic chemical A:

> "Here we come to the subjectivity of judging importance. From the point of view of the manufacturer the error in asserting the carcinogenicity of A is (or may be) more important to avoid than the error in asserting that A is harmless. Thus, for the manufacturer of A, the 'hypothesis tested' may well be: 'A is not carcinogenic'. *On the other hand, for the prospective user of chemical A the hypothesis tested will be unambiguously: 'A is carcinogenic'.* In fact, this user is likely to hope that the probability of error in rejecting this hypothesis be reduced to a very small value!"

With this passage, Neyman makes clear that testing the "no effect" hypothesis favors only the manufacturer, precisely because it assumes asymmetrically that the cost of erroneously concluding there *is* an effect (a false positive) is higher than the cost of erroneously assuming there is *no* effect (a false negative). Note also that Neyman uses the phrase "the hypothesis tested" (as opposed to null hypothesis) making clear that the tested hypothesis could be that there *is* an effect.

Like most discussions of statistical testing I have seen, the ASA statement fails to appreciate Neyman's basic points. Against my objections, the statement maintained use of the term "null hypothesis" to designate *any* hypothesis being subject to a statistical test, whether that hypothesis is that there is no effect or that there is an effect. This is in flat contradiction of ordinary English, in which "null" means "nothing," as in "no association" (e.g., Merriam-Webster 2016, Oxford 2016a)—just as in mathematics it refers to the additive identity element—and "null hypothesis" refers to no difference (Oxford 2016b). But in the field of statistics, many statisticians use "null" to mean "the hypothesis one is attempting to nullify" or refute. Thus the term "null hypothesis" joins "significance" and "interaction" as a term whose meaning in statistical jargon deviates from common usage, resulting in profoundly confused views of inference among users and even among statistics professors and textbooks.

This ingrained (and not always inadvertent) null bias in standard statistical expositions leads to a profound violation of scientific neutrality in disputes, as illustrated in Neyman's example. An extreme form of this bias is seen in testing *only* the no-effect hypothesis, then claiming the evidence supports no effect because the test was "nonsignificant" ($P > 0.05$) (see Greenland 2004, 2011, 2012 and Greenland and Poole 2011 for examples of this practice among statistics and epidemiology professors when acting as expert witnesses)—even though tests of other effect sizes would show the same statistical evidence even better supports many nonzero effects.

The falsificationist ideal that (quite independently of Popper) inspired Fisher and even more so Neyman was that tests do no more than help us see what our data cannot refute or reject under an assumed data-generation model. A failure to reject may thus allow us to proceed *as if* an unrejected effect size is correct; but it does *not* and cannot tell us which effect size to proceed with among the many that would remain unrejected by the same testing procedure. Testing only the no-effect hypothesis simply assumes, without grounds, that erroneously defaulting to no effect is the least costly error, and in this sense is a methodologic bias toward the null.

Likelihoodists and Bayesians were among the earliest to recognize the problem of focusing on single hypotheses. As Edwards (1972, p. 180) wrote of significance testing,

> "What used to be called judgement is now called prejudice, and what used to be called prejudice is now called a null hypothesis. In the social sciences, particularly, it is dangerous nonsense (dressed up as 'the scientific method'), and will cause much trouble before it is widely appreciated as such."

Ironically though, the same null bias found in significance tests is tightly integrated into the confirmationist version of "objective Bayesian" testing, which assigns a spike (point prior mass) to test a point null hypothesis against a continuous composite alternative (Jeffreys 1961; Berger and Sellke 1987). That isolated spike identifies the null as the point with special added cost to falsely reject, and the mass of the spike is a surrogate for that added cost. Again Neyman's example should be borne in mind, for in such cases the spike serves only the parties invested in maintaining the null.

As Neyman's example made clear, defaulting to "no effect" as the test hypothesis (encouraged by describing tests as concerning only "null hypothesis," as in the ASA statement) usurps the vital role of the context in determining loss, and the rights of stakeholders to use their actual loss functions. Those who benefit from this default (either directly or through their clients) have gone so far as to claim assuming "no effect" until proven otherwise is an integral part of the scientific method. It is not; when analyzed carefully such claims hinge on assuming that the cost of false positives is always higher than the cost of false negatives, and are thus circular.

In some settings (such as genomic scans) false positives are indeed considered most costly by all research participants, usu-

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Sander Greenland, Department of Epidemiology and Department of Statistics, University of California, Los Angeles, CA (Email: lesdomes@ucla.edu).

ally because everyone expects few effects among those tested will be worth pursuing. But treating these settings as if scientifically universal does violence to other contexts in which the costs of false negatives may exceed the costs of false positives (such as side effects of polypharmacy), or in which the loss functions or priors vary dramatically across stakeholders (as in legal and regulatory settings).

Those who dismiss the above issues as mere semantics or legal distortions are evading a fundamental responsibility of the statistics profession to promote proper use and understanding of methods. So far, the profession has failed abjectly in this regard, especially for methods as notoriously contorted and unnatural in *correct* interpretation as statistical tests. It has long been argued that much of harm done by this miseducation and misuse could be alleviated by suppression of testing in favor of estimation (Yates 1951, p. 32–33; Rothman 1978). I agree, although we must recognize that loss functions also enter into estimation, for example via the default of 95% for confidence or credibility intervals, and in the default to unbiased instead of shrinkage estimation. Nonetheless, interval estimates at least help convey a picture of where each possible effect size falls under the same testing criterion, thus providing a more fair assessment of competing hypotheses, and making it easier for research consumers to apply their own cost considerations to reported results.

In summary, automatically defaulting to the no-effect hypothesis is no less mindless of context and costs than is defaulting to a 0.05 rejection threshold (which is widely recognized as inappropriate for many applications). Basic statistics education should thus explain the integral role of loss functions in statistical methodology, how these functions are hidden in standard methods, and how these methods can be extended to deal with settings in which loss functions vary or costs of false negatives are large.

References

Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of p-values and Evidence" (with discussion), *Journal of the American Statistical Association*, 82, 112–122.

Edwards, A.W.F. (1972), *Likelihood*, Cambridge: Cambridge Univ. Press.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford Univ. Press.

Greenland, S. (2004), "The Need for Critical Appraisal of Expert Witnesses in Epidemiology and Statistics," *Wake Forest Law Review*, 39, 291–310.

——— (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.

Merriam Webster Dictionary online (2016), accessed 24 Feb., *http://www.merriam-webster.com/dictionary/null*.

Neyman, J. (1977), "Frequentist Probability and Frequentist Statistics," *Synthese*, 36, 97–131.

Oxford English Dictionary online (2016a), accessed 24 Feb., *http://www.oxforddictionaries.com/us/definition/american_english/null*.

——— (2016b), accessed 24 Feb., *http://www.oxforddictionaries.com/us/definition/american_english/null-hypothesis*.

Rothman, K.J. (1978), "A Show of Confidence," *New England Journal of Medicine*, 299, 1362–1363.

Yates, F. (1951), "The Influence of *Statistical Methods for Research Workers* on the Development of the Science of sStatistics," *Journal of the American Statistical Association*, 46, 1934.

Fit-for-Purpose Inferential Methods: Abandoning/Changing P-values Versus Abandoning/Changing Research

John P.A. Ioannidis

P-values continue to be widely used and misused, but until now there has been a lack of consensus in the scientific community about how grave the misuse has been, how serious the consequences are, and how exactly we should proceed to remedy the situation. Many competing options exist to change the paradigm. While the very best statisticians and methodologists do not agree on the optimal future agenda, the current status quo is perpetuated, often with prominent misconceptions that we all recognize as highly problematic. At a minimum, I hope that the positions of the ASA Statement, which reflect a high (even if not perfect) level of consensus, may offer a solid launching ground for further remedial efforts.

Currently some P-values are reported in the majority of papers that perform any statistical analysis in any empirical data (Chavalarias, Wallach, Li, and Ioannidis 2016). Conversely reporting of effect sizes is less frequent and reporting of measures of uncertainty, such as confidence intervals, is far less frequent. The use/reporting of Bayesian statistics or false-discovery rate approaches remains overall exceedingly uncommon across the published literature (Chavalarias, Wallach, Li, and Ioannidis 2016). There are certainly exceptions to this overall average pattern. Bayesian statistics (Goodman1999) and false-discovery rate (Benjamini and Hochberg 1995) may be common in specific research discipline islands. Different fields of scientific inquiry are accustomed to using different inferential tools, but unfortunately this is driven mostly by tradition and by mimicking behavior, rather than careful thinking about fit-for-purpose. Different fields also differ a lot in their accepted rules of claiming success and in their silently agreed expectations, whenever they opt to use P-values. Most fields in biomedicine and social sciences are accustomed to spurious thresholds of $P < 0.05$ for making automated claims in their inferential machinery (Gigerenzer 2004). This leads to inferential blunders, especially in an era of low prior odds for a nonnull effect, highly exploratory analyses and hidden multiplicity coupled with selective reporting (Ioannidis 2005). Other fields, e.g. genomics or experimental particle physics, typically use far more stringent P-value thresholds.

The ASA Statement may offer a sound basis on contemplating how to improve the use of statistical inferences in each field and how to forgo long-established practices in favor of others that are better suited to what each scientific field aims to achieve. The best recipe is unlikely to be the same in all scientific disciplines and it is unlikely that there will be only one optimal recipe in each discipline. But some current practices immediately seem to be grossly misleading. It is fine to correct those misleading practices, regardless of what exactly they are replaced with, among several reasonable alternatives.

For example, many observational research fields such as large segments of nutritional epidemiology, electronic health record-based investigations using routinely collected data or other big data compilations, can be described as high-output machines producing copious P-value trash. With a combination of large datasets, confounding, flexibility in analytical choices (Patel, Burford, and Ioannidis 2015), and superimposed selective reporting bias, using a $P < 0.05$ value threshold to declare "success," any result can become statistically significant, but this means next to nothing. As Jeffreys put it over half a century ago: "A null hypothesis is set up and 'tested' against data: It is merely something set up like a coconut to stand until it is hit" (Jeffreys 1961) Many scientific fields are accustomed to taking an endless number of shots until they (unfortunately) hit the coconut.

Raising the bar to more stringent P-value thresholds may reduce some of that trash, but does not attack the root of the problem, and of course it may generate also some false-negatives (Ioannidis, Tarone, and McLaughlin 2011) One has to look calmly at the main principles. Does a null-hypothesis even make sense to test? Perhaps in several of the current P-value-chasing investigations nobody should really have cared about rejecting the null-hypothesis. In fact, is there any possibility that a null-hypothesis can avoid being rejected, if one can assemble a larger and larger sample size, in a setting where confounding and bias are impossible to eradicate to the point that whatever signals can be separated from the noise? Why test against the null when the null is impossible and/or meaningless? The principles of the Statement should lead to some thought before running any statistical analysis.

Sometimes, P-values should be avoided and other methods should be used instead, or simply descriptive metrics might suffice. Other times, it is doing the research that should be avoided, if the results are likely to be misleading, regardless of the inferential methods used. Some of the most prolific fields of current research (in terms of publication volume) are practically not contributing knowledge, but just expressing repeatedly how big bias can be in their domain. Then it is not an issue of abandoning P-values, it is an issue of abandoning poor research. Misleading use of P-values is so easy and automated that, especially when rewarded with publication and funding, it can become addictive. Investigators generating these torrents of P-values should be seen with sympathy as drug addicts in need of rehabilitation that will help them live a better, more meaningful scientific life in the future.

In many other fields, inferences using P-values will continue

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. John P.A. Ioannidis, C.F. Rehnborg Professor in Disease Prevention, Professor of Medicine, of Health Research and Policy, and of Statistics, Director of the Stanford Prevention Research Center, and Co-Director of the Meta-Research Innovation Center at Stanford (METRICS), Stanford University, 1265 Welch Rd, MSOB X306, Stanford, CA 94305 (Email: jioannid@stanford.edu).

©2016 John P.A. Ioannidis

to offer helpful insights, if properly used and interpreted. Other inferential methods may need to be used more frequently. In some fields, their use is overdue. For example in clinical trials and their meta-analyses, presenting effect sizes and their uncertainty should be the default and P-values can be nicely complemented, if not largely replaced, by Bayesian inferences. Using alternative inferential tools will still not solve, nevertheless, some of the problems that cause many misleading claims in this literature, in particular those related to hidden multiplicity and selective reporting biases (Dwan et al. 2013). If success is defined based on passing some magic threshold, biases may continue to exert their influence regardless of whether the threshold is defined by a P-value, Bayes factor, false-discovery rate, or anything else. Efforts to promote transparency in study design, conduct and reporting may have more to offer in this setting than blaming P-values. Studying how these efforts can be most successful is an entire field of research on its own (Ioannidis, Fanelli, Dunne, and Goodman 2015).

References

Chavalarias, D., Wallach, J., Li, A., and Ioannidis J.P. (2016), “Evolution of Reporting of P-values in the Biomedical Literature, 1990–2015,” *Journal of the Amewrican Medical Association*, in press.

Goodman, S.N. (1999), “Toward Evidence-Based Medical Statistics. 2: the Bayes Factor,” *Annals of Internal Medicine*, 130, 1005–1013.

Benjamini, Y., and Hochberg. Y. (1995), “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society*, Series B, 57, 289–300.

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

Ioannidis, J.P. (2005), “Why Most Published Research Findings are False,” *PLoS Medicine*, 2:e124.

Patel, C.J., Burford, B., and Ioannidis, J.P. (2015), “Assessment of Vibration of Effects Due to Model Specification can Demonstrate the Instability of Observational Associations,” *Journal of Clinical Epidemiology*, 68, 1046–1058.

Jeffreys, H. (1961), *Theory of Probability*, Oxford, Oxford University Press.

Ioannidis, J.P., Tarone, R., and McLaughlin, J.K. (2011), “The False-Positive to False-Negative Ratio in Epidemiologic Studies,” *Epidemiology*, 22, 450–456.

Dwan, K., Gamble, C., Williamson, P.R., Kirkham, J.J.; Reporting Bias Group (2013), “Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias—An Updated Review,” *PLoS One*, 8:e66844.

Ioannidis, J.P., Fanelli, D., Dunne, D.D., and Goodman, S.N. (2015), “Meta-research: Evaluation and Improvement of Research Methods and Practices,” *PLoS Biology*, 13:e1002264.

Comments on the "ASA Statement on Statistical Significance and P-values" and Marginally Significant P-Values

Valen E. JOHNSON

The Board of Directors of the American Statistical Association recently issued a policy statement regarding the interpretation of p-values and statistical significance. This statement provides important guidance to scientists regarding the proper use and interpretation of p-values, along with cautions to avoid their misuse. In this note, I examine the common fallacy that p-values near 0.05 provide "significant" evidence against a null hypothesis.

The ASA statement on statistical significance and p-values addresses a number of important issues regarding the interpretation of p-values and statistical hypothesis testing. In this note, I comment further on one of those issues, namely the assertion that "a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis."

To provide a context for this statement, it is useful to consider what is perhaps the most elementary of statistical hypothesis tests, that of testing whether the mean μ of a normal population is 0 when the variance is known to be σ^2, based on a random sample $(x_1, \ldots, x_n)$ of size n from that population. If the alternative hypothesis requires that $\mu > 0$ (so that a one-side test is performed), then the null hypothesis is rejected at the 5% level of significance in the uniformly most powerful test if the sample mean $\bar{x}$ exceeds $1.645\sigma/\sqrt{n}$. If $\bar{x} = 1.645\sigma/\sqrt{n}$, then the p-value of the test is 0.05.

The "weakness of evidence" provided by this p-value is revealed when one examines the likelihood ratio of the sampling density of the data under the null hypothesis to the maximum of the sampling density of the data under the alternative hypothesis. If $\phi(x|\mu, \sigma)$ denotes the normal density function with mean μ and variance σ^2 evaluated at x, then the minimum likelihood ratio equals

$$\lambda = \arg\min_{\mu>0} \prod_{i=1}^{n} \frac{\phi(x_i|0, \sigma)}{\phi(x_i|\mu, \sigma)} = 0.258. \tag{1}$$

In other words, the sampling density of the data under the null hypothesis is *at least 1/4 as large as it is under any alternative hypothesis*. If the null and alternative hypotheses are regarded as being equally likely a priori (or from a repeated sampling context, if one-half of tested null hypotheses are true), then the probability that the null hypothesis is true when $p = 0.05$ is *at least 20%.*

This fact is not new, of course, and an extended discussion of this "paradox" was provided over 50 years ago by Edwards, Lindman and Savage (1963). This paradox is not specific to z-tests or one-sided tests, and it is not caused by the specification of a point null hypothesis to conveniently represent the notion that the mean μ is close to a specified null value.

To see that the latter claim is true, it is useful to view the hypothesis testing problem from a Bayesian perspective and replace the null hypothesis that $\mu = 0$ by the assumption that μ is drawn from a prior density function $\pi_0(\mu)$ that is symmetric around 0 and is positive only when $|\mu| < 1.645\sigma/\sqrt{n}$. Then the marginal likelihood of the data is evaluated by averaging over this interval, i.e.,

$$f_0(x_1, \ldots, x_n) \equiv \int_{-1.645\sigma/\sqrt{n}}^{1.645\sigma/\sqrt{n}} \prod_{i=1}^{n} \phi(x_i|\mu, \sigma)\pi_0(\mu)d\mu. \tag{2}$$

If $\pi_1(\mu)$ is the prior density for μ assumed under the alternative hypothesis and

$$f_1(x_1, \ldots, x_n) \equiv \int_{-\infty}^{\infty} \prod_{i=1}^{n} \phi(x_i|\mu, \sigma)\pi_1(\mu)d\mu, \tag{3}$$

then the ratio λ in (1) can be replaced with the Bayes factor[1]

$$\mathrm{BF}_{01}(\bar{x}) = \frac{f_0(x_1, \ldots, x_n)}{f_1(x_1, \ldots, x_n)}. \tag{4}$$

When $p = 0.05$, it again follows that $\mathrm{BF}_{01}(\bar{x})$ will be larger than 0.258, no matter what prior density $\pi_1(\mu)$ one choses for μ, even when the point null hypothesis has been replaced by a "small interval" null hypothesis.

Similar comments apply to the case of p-values in two-sided z tests. In that setting, p=0.05 if $\bar{x} = \pm 1.96\sigma/\sqrt{n}$. To account for the fact that the null hypothesis is rejected for both large positive and large negative values of $\bar{x}$, it makes sense to assume that the prior density on μ is symmetrically distributed around the null value of $\mu = 0$. If one accepts this assumption, then the ratio of the sampling density of the data under the null hypothesis to the average sampling density of the data under the alternative hypothesis, obtained by averaging over any prior distribution on μ that is symmetric around 0, exceeds 0.29. If the null hypothesis is assumed be at least as likely as the alternative hypothesis a priori, then the posterior probability that the null hypothesis is true when $p = 0.05$ in a two-sided z-test is at least 0.226 (Berger and Sellke 1987).

The one-sample z-test is a special case of a null hypothesis significance test (NHST) in a one parameter exponential family model (1PEF). The Neyman-Pearson lemma guarantees the existence of uniformly most powerful tests (UMPTs) for many

Valen E. Johnson, University Distinguished Professor Texas A and M University System–Statistics MS 3143, College Station, TX 77845-3153 (Email: vjohnson@stat.tamu.edu)

[1] In general, the Bayes factor of a test can be viewed from a classical perspective as an integrated likelihood ratio, integrated with respect to the prior densities on the unknown parameters.

©2016 Valen E. Johnson

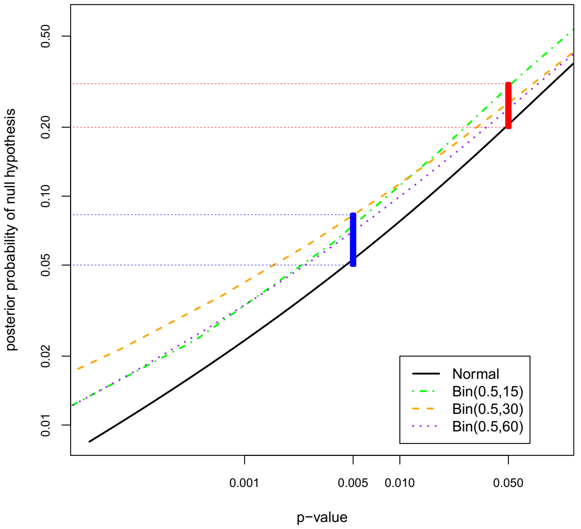


Figure 1. P-values versus posterior probabilities of null hypotheses. The curves in this plot were constructed using UMPBT alternative hypotheses and by assigning equal prior probability to the null and alternative hypotheses. Tests labeled $\text{Bin}(0.5, n)$ test a null hypothesis that a success probability is 0.5 based on a sample size of n. All tests are one-sided. Both axes are scaled logarithmically.

NHSTs in 1PEFs. As it happens, it is also possible to define uniformly most powerful Bayesian tests (UMPBTs) in the same setting by choosing the alternative hypothesis in a NHST so as to maximize the probability that the Bayes factor of the test exceeds a specified threshold (Johnson 2013a). Furthermore, the threshold of a UMPBT can be chosen so that the Bayesian test has the same type I error as the classical UMPT.

The correspondence between UMPTs and UMPBTs (matched by appropriately chosen test sizes and evidence thresholds) makes it straightforward to extend the analysis of marginally significant p-values beyond simple z-tests to more general NHSTs. Again assuming the null hypothesis is assigned prior probability of 0.5 (as it might in the case when the evidence in “a p-value near 0.05 is taken by itself”), Figure 1 displays a plot of p-values for common normal and binomial tests versus the posterior probability that the null hypothesis is true. The posterior probabilities displayed in this plot were obtained by using the UMPBT that corresponds to the size 0.05 one-sided test. Similar plots can also be constructed for two sided tests, other 1PEF tests, and (using approximate UMPBTs) t tests (Johnson, 2013b).

The red box in Figure 1 highlights the posterior probabilities of null hypotheses based on p-values of 0.05. Under the mild assumptions described above, this box shows that the posterior probability of the null hypotheses for p-values near 0.05 range between 0.20 and about 0.35. Note that when $p = 0.05$, higher posterior probabilities would be assigned to the null hypothesis for any alternative hypotheses other than the UMPBT.

The blue box in Figure 1 highlights posterior probabilities for $p = 0.005$, and shows that the corresponding posterior probabilities of null hypotheses for these z-tests and binomial tests range between approximately 1/20 and 1/12. At this level of significance, the posterior probability of the null hypotheses has fallen to the level of evidence that many scientists implicitly believe that $p = 0.05$ represents. Which begs the question, “should $p = 0.005$ be the new $p = 0.05$?” (Johnson, 2013b).

In summary, simple calculations of likelihood ratios and Bayes factors suggest that p-values near 0.05, by themselves, provide very little evidence against a null hypothesis in NHSTs. For likelihood ratios, the ratio of the data density under the null hypothesis to the data density under the alternative hypothesis exceeds 0.20 when $p = 0.05$ for common hypothesis tests. Similarly, from a Bayesian perspective using alternative hypotheses that are chosen so as to minimize the probability assigned to the null, the posterior probability of the null hypotheses typically exceeds 0.20 when $p = 0.05$ (provided that both hypotheses are assigned equal probability a priori.)

As the ASA statement asserts, “a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis.”

References

Berger, J., and Selke, T. (1987), “Testing a Point Null Hypothesis: The Irreconcilability of p Values and Evidence,” *Journal of the American Statistical Association*, 82, 112–122.

Edwards, W., Lindman, H., and Savage, L. (1963), “Bayesian Statistical Inference for Psychological Research,” *Psychological Review*, 70, 193–242.

Johnson, V.E.(2013a), “Uniformly Most Powerful Bayesian Tests,” *Annals of Statistics*, 41, 1716–1741.

——— (2013b), “Revised Standards for Statistical Evidence,” *Proceedings of the National Academy of Sciences*, 110(48), 19313–19317.

Comment

Michael LAVINE and Joseph HOROWITZ

We would like to offer some comments on the definition and interpretation of P-values. To set the stage, consider an over-simplified multiple comparisons situation in which we test 100 hypotheses H_j, for $j = 1, \dots, 100$. The jth hypothesis yields a z-score Z_j. Suppose it turns out that $|Z_{22}| = 3$ is the largest absolute value of the 100 z-scores.

A P-value is, to paraphrase the ASA statement, the probability under a specified model that a statistic would be more extreme than its observed value. Thus, to have P-values, we need models and statistics; each P-value pertains to a particular (model, statistic) pair. We focus our comments on just two of the many pairs we might consider:

Pair A ($H_{0,A}$: $\mu_{22} = 0$, statistic$_A$: Z_{22}) and

Pair B ($H_{0,B}$: $\mu_1 = \cdots = \mu_{100} = 0$, statistic$_B$: Z_{J^*}) where J^* is the maximizer of $|Z_j|$.

Each of these pairs has a P-value. In our experiment, under the obvious Normality assumption, $P_A = 2(1 - \Phi(3)) \approx .0027$, even if attention was focussed on Pair A only after the data were collected, whereas P_B cannot be calculated without further assumptions about the joint distribution of $Z_1, \dots, Z_{100}$.

With this background, we make the following observations.

1. In multiple-comparison settings one often encounters the question Q_1: *Should P-values be adjusted?*, which sounds like a technical question about statistics, to be answered by statistical theory. But because P_A ($\approx .0027$) is already a valid P-value (for A) without adjustment, Q_1 puts the emphasis in the wrong place. Often a more useful question is Q_2: *Which pair, and therefore which P-value, should we care about, A or B?*, a question about the investigation, to be answered in collaboration with the investigator in the context of background knowledge.

   One might observe that A does not accurately represent the way the data were collected. That may be true, but there is nothing in the definition of P-value to say that H_0 must reflect the experimental design. One might argue that $Z_1, \dots, Z_{100}$ ought to be modeled jointly, not separately. That may be true, but there is nothing in the definition of P-value to say that the model in the (model, statistic) pair must accurately reflect the distribution of the data. One might note that if we report P_A, some people will interpret it as a P-value for B. That may be true, but is the result of a misunderstanding and does not mean that P_A is not a valid P-value for A.

   As statisticians, we can point out the differences between A and B; we can help build models for the joint distribution of $Z_1, \dots, Z_{100}$; we can explain the different distributions of Z_{22} and Z_{J^*}; and we can help researchers think about whether they should care about A or B. But where the ASA's statement says *"[c]onducting multiple analyses of the data and reporting only those with certain p-values ... renders the reported p-values essentially uninterpretable,"* we would say instead that results should be reported so that they are useful to readers interested in A, B, or any other hypothesis that might be of interest, and so that they help readers distinguish and decide between A and B.

2. The ASA's statement says *"Cherry-picking promising findings ... leads to a spurious excess of statistically significant results."* But there are at least two points of view regarding spurious excess.

   (a) There are 100 individual hypotheses similar to $H_{0,A}$; they are $\mu_1 = 0, \dots, \mu_{100} = 0$. If all 100 hypotheses are true, then about five of them will yield P-values less than about 0.05. There is no excess of small P-values or declarations of significance.

   (b) There is a single hypothesis $H_{0,B}$. If it is true and we calculate the 100 P-values pertaining to the 100 individual hypotheses then there is a large probability that one or more of them will be less than 0.05. There is an excess of small P-values and declarations of significance.

   It seems, to us, that the purported excess of small P-values in (b) is due to treating individual P-values of type A as though they are of type B. Whether there is truly a spurious excess depends on whether we care about pairs like A or like B.

3. Whatever is the joint distribution of $Z_1, \dots, Z_{100}$, $P_B \geq P_A$. In fact, P_B is greater than or equal to each of the 100 P-values in (a) above. Assuming independence of $Z_1, \dots, Z_{100}$ gives $P_B = 1 - (0.9973)^{100} \approx .24$ which, under the usual interpretation, means that the data are compatible with $H_{0,B}$. The same data also yield $P_A \approx 0.0027$, which means that the data are not compatible with $H_{0,A}$. But because $H_{0,B} \subset H_{0,A}$ — i.e. $H_{0,B} \Rightarrow H_{0,A}$ — those two inferences about compatibility are incompatible. That's a general phenomenon of P-values pointed out by Schervish (1996): if a parameter space Θ can be partitioned into null and alternative hypotheses in two ways such that, say, $H_0 \subset H_0'$, so, necessarily, $H_a' \equiv H_0'^c \subset H_a \equiv H_0^c$, then the P-value for H_0 may be larger than the P-value for H_0', even though logic dictates that the data must be at least as compatible with H_0' as with H_0. The

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Michael Lavine, and Joseph Horowitz, University of Massachusetts, Lederle Graduate Research Tower, Box 34515, Amherst, MA 01003-9305 (Email: lavine@math.umass.edu).

©2016 Michael Lavine and Joseph Horowitz

incompatibility is inherent to P-values and cannot be resolved. Thus, P-values cannot be interpreted formally as evidence measures or, at least, the mapping between P-values and "evidence" varies according to circumstance. The ASA statement's Principle 1: *"P-values can indicate how incompatible the data are with a specified statistical model"* can be interpreted only informally, at best.

References

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

Three Inferential Questions, Two Types of P-value

Michael J. Lew

Many users of P-values don't understand why their standard practices are derided as "P-value hacking" and when and why they should abstain from "cherry picking." Such confusion is predictable as statistical methods and behaviors appropriate in some circumstances are inappropriate, dangerous, and verging on dishonest in others. The ASA statement doesn't describe such circumstances and so I will introduce some of them here. This commentary should be read more as an extension of the statement rather than a commentary on its contents, with which I agree with few reservations.

Richard Royall noted that there are three types of inferential question that can be answered with the help of statistical methods (Royall 1997).

1. What do these data say?
2. What should I believe now that I have these data?
3. What should I do or decide now that I have these data?

Those questions might seem so obvious that it is superfluous to mention them. However they are so rarely mentioned as to remain novel, and they provide a good scaffold for understanding the roles of P-values in scientific inference. Misuse of P-values often involves an implicit assumption that they provide answers to all three questions, but P-values cannot, by themselves, tell an investigator what to believe or what to decide.

P-values answer the first question by being an index to the evidential meaning of the data within a statistical model. As noted in the ASA statement, P-values are anchored to a single hypothetical value of the parameter of interest, the 'null hypothesis', within a particular statistical model, so they are not always the *best* way to answer the first question. A likelihood function gives a richer depiction of the evidence in the data about parameter values than does a P-value from the same statistical model, as the likelihood function allows comparison of the evidential support for all values of the parameter of interest. Nevertheless, P-values are a useable and defensible answer to the question of what the data say—at least when they are accompanied by adequate demonstration of the observed effect size and relevant experimental and analytical details.

An answer to "what should I believe now that I have these data?" should meld what the data say with what was known or believed beforehand. Bayesian methods formally answer that question with a prior probability distribution to represent the pre-data information or belief. The question "what should I do or decide now that I have these data?" requires consideration of what the data say in conjunction with the benefits and costs of correct and incorrect decisions or actions. In other words, a decision process requires a loss function in addition to the data. The classical Neyman and Pearson hypothesis test is probably the most widely used decision theory approach, and its loss function is built into the designed balance between false positive and false negative error rates, α and β. For example, if α is set to a smaller value than β in the pre-data study design then that loss function reflects a greater cost of false positive than false negative errors.

Researchers should be aware of the distinction between the questions answered by the exact P-value and the conventionally dichotomized hypothesis test result. To reflect what the data say, P-values have to be treated in a nondichotomous manner, as the evidence is not simply present or absent, but is graded. Converting P-values into "significant" and "not significant" can be appropriate when answering the third question with a Neyman and Pearson hypothesis test procedure, as long as a pre-study power analysis for sample size determination has been done. Unfortunately, such a power analysis is rarely performed or reported in publications of basic biomedical science (Strasak et al. 2007), and if you dichotomise a P-value by taking $\mathrm{P} \leqslant 0.05$ as "significant" without having designed the loss function, then you are using the mechanical "bright line" rule deprecated in the ASA statement. The absence of a loss function does not preclude exact P-values from serving as an answer to the first question, and a dichotomizing hypothesis test is not the only basis for a scientific conclusion (Lew 2012).

The choice of analytical procedures should be informed by the nature of the study because if you restrict your attention to answering the first question you can identify the areas where cherries are most numerous and ripe without picking them. Data from preliminary or exploratory studies intended to determine fruitful directions of enquiry can be interrogated repeatedly and intensively and results can sensibly be assessed and communicated on the basis of observed P-values, even if the study involves many comparisons, even if the comparisons are unplanned, and even if the sampling rules were ill-defined or flexible. No "correction" of those P-values for multiplicity of comparisons is necessary—or desirable—because what the data say about one hypothetical effect is not influenced by whether the analyst sees what the data say about another hypothetical effect. In contrast, if those same P-values were used with hypothesis testing procedures to provide the basis for decisions regarding hypotheses then claims of "cherry picking" and "P-hacking" would usually be correct. A pre-study power analysis is required, and all of the comparisons to be made must be included for the loss function to be correctly calibrated. Thus P-values used within a hypothesis test decision procedure often need adjustment to take the actual experimental design into account lest the statistical support for decisions or actions is weaker than claimed or implied because of a higher than reported risk of false positive outcomes. Exploratory studies should not be misrepresented as planned studies yielding answers to the third

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Michael J. Lew, University of Melbourne, Department of Pharmacology and Therapeutics, Parkville Victoria 3010, Australia. Email: michaell@unimelb.edu.au

©2016 Michael J. Lew

question.

There are two types of P-values: P-values that show what the data say, and P-values to be used in decision processes. Analytical maneuvers that should be derided as "P-hacking" and "cherry picking" in a planned study are perfectly appropriate in the setting of a preliminary study. The rights and wrongs of using P-values are context and purpose dependent.

References

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, 1559–1567.

Royall, R. M. (1997), *Statistical Evidence: A Likelihood Paradigm*, volume 71 of *Monographs on Statistics and Applied Probability*. London: Chapman & Hall.

Strasak, A, Zaman, Q. Marinell, G., and Pfeiffer, K. (2007), "The Use of Statistics in Medical Research: A Comparison of The New England Journal of Medicine and Nature Medicine," *The American Statistician*, 61, 47–55.

Comment

Roderick J. Little

The P-value statement is good, but perhaps more prominence should be given to the problems arising from the use of p-values as an *isolated* statistical measure. With very large datasets and high precision, p-values are all but useless, and the main focus is on estimates and potential sources of bias. With smaller studies where precision matters, no single measure can simultaneously answer two questions—"is the effect large?" and "is the estimated effect signal or noise?" An estimate and p-value address these questions, but I find a confidence or credibility interval much more cohesive and intuitive, giving a range of parameter values consistent with the data, and avoiding the need for a null hypothesis.

I teach a basic course in biostatistics to public health students. Confidence intervals are no problem—ideas like margin of error have even entered the vernacular. The difficulties begin with hypothesis testing. Aside from the elaborate terminology, what's the null and what's the alternative? When should the test be one-sided and when two-sided? Who cares about a point null that is never true (e.g., Nester 1996)? If a deviation in one direction is of interest, the appropriate test seems to be one-sided; but the p-value calculation is still based on the known distribution of the test statistic under the point null, ignoring all the other null values. Isn't this a sleight of hand? Then we get to power calculations, which elude most quantitatively challenged students...

I have to teach hypothesis testing since it is so prevalent in biomedical research, but life would be much easier if we could just focus on estimates with their associated uncertainty. The basic artifice of hypothesis testing as a concept is perhaps the root cause of the problem, and I doubt that it will be solved by judicious and subtle statements like this one from the ASA Board.

Johnson's (2013) work points to an excessively high level of significance (5%) as a factor contributing to the failure to replicate science. On a lighter note, this inspired the following limerick, which I offer as my lame contribution to the debate:

In statistics, one rule did we cherish:
P point oh five we publish, else perish!
Said Val Johnson, "that's out of date,
Our studies don't replicate
P point oh oh five, *then* null is rubbish!"

References

Johnson, V. E. (2013), "Revised Standards for Statistical Evidence," *Proceedings of the National Academy of Sciences*, 110, 48, 19313–19317.

Nester, M. R. (1996), "An Applied Statisticians Creed," *Applied Statistics*, 45, 4, 401–410.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Roderick J. Little, University of Michigan (Email: rlittle@umich.edu).

©2016 Roderick J. Little

Don't Throw Out the Error Control Baby With the Bad Statistics Bathwater: A Commentary

Deborah G. MAYO

The American Statistical Association is to be credited with opening up a discussion into p-values; now an examination of the foundations of other key statistical concepts is needed.

Statistical significance tests are a small part of a rich set of "techniques for systematically appraising and bounding the probabilities (under respective hypotheses) of seriously misleading interpretations of data" (Birnbaum 1970, p. 1033). These may be called *error statistical methods* (or *sampling theory*). The error statistical methodology supplies what Birnbaum called the "one rock in a shifting scene" (ibid.) in statistical thinking and practice. Misinterpretations and abuses of tests, warned against by the very founders of the tools, shouldn't be the basis for supplanting them with methods unable or less able to assess, control, and alert us to erroneous interpretations of data.

P-value. The significance test arises to test the conformity of the particular data under analysis with H_0 in some respect:

> To do this we find a function $t = t(y)$ of the data, to be called the test statistic, such that
>
> - the larger the value of t the more inconsistent are the data with H_0;
> - the corresponding random variable $T = t(Y)$ has a (numerically) known probability distribution when H_0 is true.
>
> ...[We define the] p-value corresponding to any t as $p = p(t) = P(T \geq t; H_0)$ (Mayo and Cox 2006, p. 81).

Clearly, if even larger differences than t occur fairly frequently under H_0 (p-value is not small), there's scarcely evidence of incompatibility. But even a small p-value doesn't suffice to infer a genuine effect, let alone a scientific conclusion—as the ASA document correctly warns (Principle 3). R. A. Fisher was clear that we need not isolated significant results:

> ...but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result. (Fisher 1947, p. 14)

If such statistically significant effects are produced reliably, as Fisher required, they indicate a genuine effect. This is the essence of statistical falsification in science. The logic differs from inductive updating probabilities of a hypothesis, or a comparison of how much more probable H_1 makes the data than does H_0, as in likelihood ratios. Given the need to use an eclectic toolbox in statistics, it's important to avoid expecting an agreement on numbers from methods evaluating different things. Hence, it's incorrect to claim a p-value is "invalid" for not matching a posterior probability based on one or another prior distribution (whether subjective, empirical, or one of the many conventional measures).

Effect sizes. Acknowledging Principle 5, tests should be accompanied by interpretive tools that avoid the fallacies of rejection and nonrejection. These correctives can be articulated in either Fisherian or Neyman-Pearson terms (Mayo and Cox 2006; Mayo and Spanos 2006). For an example of the former, looking at the p-value distribution under various discrepancies from H_0: $\mu = \mu_0$ allows inferring those that are well or poorly indicated. If you very probably would have observed a more impressive (smaller) p-value than you did, if $\mu > \mu_1$ (where $\mu_1 = \mu_0 + \gamma$), then the data are good evidence that $\mu \leq \mu_1$. This is akin to confidence intervals (which are dual to tests) but we get around their shortcomings: We do not fix a single confidence level, and the evidential warrant for different points in any interval are distinguished. The same reasoning allows ruling out discrepancies when p-values aren't small. This is more meaningful than power analysis, or taking nonsignificant results as uninformative. Most importantly, we obtain an evidential use of error probabilities: to assess how well or *severely tested* claims are. Allegations that frequentist measures, including p-values, must be misinterpreted to be evidentially relevant are scotched.

Biasing selection effects. We often hear it's too easy to obtain small p-values, yet replication attempts find it difficult to get small p-values with preregistered results. This shows the problem isn't p-values but failing to adjust them for cherry picking, multiple testing, post-data subgroups and other *biasing selection effects*. The ASA correctly warns that "[c]onducting multiple analyses of the data and reporting only those with certain p-values" leads to spurious p-values (Principle 4). The *actual* probability of erroneously finding significance with this gambit is not low, but high, so a *reported* small p-value is invalid. However, the same flexibility can occur with likelihood ratios, Bayes factors, and Bayesian updating, with one big difference: The direct grounds to criticize inferences as flouting error statistical control is lost (unless they are supplemented with principles that are not now standard). The reason is that they condition on the actual data; whereas error probabilities take into account other outcomes that could have occurred but did not.

The introduction of prior probabilities—which may also be data dependent—offers further leeway in determining if there has even been replication failure. Notice the problem with biasing selection effects isn't about long-run error rates, it's being unable to say that the *case at hand* has done a good job of avoiding misinterpretations.

Model validation. Many of the "other approaches" rely on statistical models that require "diagnostic checks and tests of fit

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Deborah G. Mayo, Department of Philosophy, Virginia Tech, (Email: mayod@vt.edu).

©2016 Deborah G. Mayo

which, I will argue, require frequentist theory significance tests for their formal justification" (Box 1983, p. 57), leading Box to advocate ecumenism. Echoes of Box may be found among holders of different statistical philosophies. "What we are advocating, then, is what Cox and Hinkley (1974) call 'pure significance testing', in which certain of the model's implications are compared directly to the data…" (Gelman and Shalizi 2013, p. 20).

We should oust recipe-like uses of p-values that have been long lampooned, but without understanding their valuable (if limited) roles, there's a danger of blithely substituting "alternative measures of evidence" that throw out the error control baby with the bad statistics bathwater.

References

Birnbaum, A. (1970), "Statistical Methods in Scientific Inference (letter to the Editor)," *Nature* 225(5237), 1033.

Box, G. (1983), "An Apology for Ecumenism in Statistics," in *Scientific Inference, Data Analysis, and Robustness*, eds. G. E. P. Box, T. Leonard, and D. F. J. Wu, New York: Academic Press, 51–84.

Cox, D., and Hinkley, D. (1974), *Theoretical Statistics*, London: Chapman and Hall.

Fisher, R. A. (1947), *The Design of Experiments* (4th ed.), Edinburgh: Oliver and Boyd.

Gelman, A., and Shalizi, C. (2013), "Philosophy and the Practice of Bayesian Statistics" and "Rejoinder," *British Journal of Mathematical and Statistical Psychology*, 66(1), 8–38; 76–80.

Mayo, D. G., and Cox, D. R. (2006), "Frequentists Statistics as a Theory of Inductive Inference," in *Optimality: The Second Erich L. Lehmann Symposium*, ed. J. Rojo, Lecture Notes-Monograph Series, Institute of Mathematical Statistics (IMS), 49, 77–97.

Mayo, D. G., and Spanos, A. (2006), "Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction," *British Journal for the Philosophy of Science*, 57(2), 323–357.

ASA Statement on P-values: Some Implications for Education

Anne Michèle MILLAR

Over the last 20 years there has been a phenomenal improvement in our approach to teaching statistics. Current texts emphasize the understanding of concepts and the use of technology, illustrated with real world data. However the vast majority of our introductory statistics text books teach null hypothesis significance testing as the definitive method (and the only method) for statistical inference, and many still recommend the default of a 5% significance level to assess the test results. This needs to change.

Introductory statistics courses are usually service courses for students in other disciplines. Only a minority of these students will implement statistical methods in their careers, but they do need to understand the statistical content of their reading. This content will include fixed level testing, so it is important to continue to teach an understanding of the approach, without advocating it, while clearly addressing the disadvantages. It is hard for students to grasp the concept of a p-value. While some texts have done an excellent job of listing the many misconceptions along with correct interpretations when first introducing the concept, it is important that we continue to reinforce the correct concept throughout the introductory course (and second and third courses!). We also need to make our students aware that p-values are not the “only way.” At a minimum we can include confidence intervals whenever we perform a test, to assess practical significance in addition to statistical significance. Ideally we would include an introduction to Bayesian methods.

Authors, you are in the perfect position to ensure our texts are in keeping with current statistical practice. An additional chapter to introduce some of the ideas for Bayesian methods would be appreciated, and a book introducing Bayesian methods, suitable for introductory students with little quantitative background, would be a wonderful resource.

How can we implement change as teachers? First, within our own teaching. I have made some major changes in my lecture presentations and student activities as a result of my involvement with the ASA statement. Second, we can act as advocates. Start with your colleagues, encourage them to read and discuss the statement. Contact the publishers and authors of the texts you currently use, or texts you are considering for adoption, and request updates. New editions usually appear at regular intervals, and my experience has been that the authors not only appreciate feedback but actively encourage it—they want to satisfy the needs of their audience. Be aware of the exercises and the solutions provided when inspecting your text. Often the content and the worked examples have been updated and new exercises added, but the old exercises and their solutions remain unchanged. The instructor (and student) solution manuals are not necessarily written by the original authors, and the solutions left from earlier editions may no longer conform to current statistical practice. This an issue not only for students, but for instructors teaching statistics for the first time who rely on such manuals as teaching resources. We need to keep publishers aware of our concerns.

I would like to thank the ASA for leading the discussion around p-values and developing the statement, and for inviting me to the October 2015 meeting. I have learned a lot, and am truly grateful

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Anne Michèle Millar, Associate Professor, Department of Mathematics, Mount Saint Vincent University, 166 Bedford Highway, Halifax, N.S. B3M 2J6, CANADA (Email: Michele.Millar@msvu.ca).

©2016 Anne Michèle Millar

Disengaging from Statistical Significance

Kenneth J. Rothman

In the marketplace of scientific results, the preferred currency by which results have been valued has been statistical significance, expressed either as a dichotomous label or by the underlying p-value, which may be given as a number or an inequality. Like other modern currencies, the value of this one is not inherent but derived from widely held assumptions and expectations. Indeed, reliance on statistical significance is as misplaced as faith in some dubious paper monies. At the risk of stretching the analogy, I suggest that a version of Gresham's Law has been operating, allowing statistical significance to force out of circulation better ways to analyze data, and leaving us with results that are, all too often, astonishingly misleading.

As the ASA statement (ASA 2016) indicates, a fundamental flaw of relying on statistical significance for inference is the need to dichotomize all results into those that are significant or not significant. This practice degrades vast efforts to collect and analyze quantitative data into a mere label. Furthermore, if ever there were a false dichotomy, it is the dichotomy between significant and not. The label is assigned by an arbitrary rule and inevitably has less information than the p-value from which it derives. Moreover, the p-value itself is a handicapped approach to interpretation because it doesn't measure effect size. Instead, it blends together information on estimated effect size and the precision of that estimate (Lang, Rothman, and Cann 1998). Although p-values and confidence intervals are closely related, a confidence interval, in contrast to a p-value, expresses separately both effect size and precision (Poole 2001). This advantage of confidence intervals illustrates that it usually takes two numbers to measure two distinct characteristics. Unfortunately, all too often we have seen the reported confidence interval used merely to determine if the null value lies within it or not, debasing the confidence interval into a label, a surrogate significance test (Cumming 2012).

The correspondence between results that are statistically significant and those that are truly important is far too low to be useful. Consequently, scientists have embraced and even avidly pursued meaningless differences solely because they are statistically significant, and have ignored important effects because they failed to pass the screen of statistical significance. These are pernicious problems, and not just in the metaphorical sense. It is a safe bet that people have suffered or died because scientists (and editors, regulators, journalists and others) have used significance tests to interpret results, and have consequently failed to identify the most beneficial courses of action (Hauer 2004; Schmidt and Rothman 2014).

How do we fix this problem? The reliance on statistical significance testing is ingrained in the social system of many sciences, and therefore reflexive on the part of many members of those social systems, making it difficult to counter. Nonetheless, we can and should advise today's students of statistics that they should avoid statistical significance testing, and embrace estimation instead. Those who have tried offering this advice know it can be challenging. Students all too often fear that their success will be measured by publications and grants that are evaluated by reviewers who esteem statistical significance. Despite such inertia, in epidemiology there has been an encouraging trend toward reporting confidence intervals to supplement or even supplant statistical significance and p-values, toward using confidence intervals to measure effect size and to gauge precision rather than to test null hypotheses, and toward avoiding the fallacy of considering every statistically non-significant result as if it were evidence for a null relation.

Real change will take the concerted effort of experts to enlighten working scientists, journalists, editors and the public at large that statistical significance has been a harmful concept, and that the estimates of meaningful effect measures is a much more fruitful research aim than the testing of null hypotheses. This statement of the ASA does not go nearly far enough toward that end, but it is a welcome start and a hopeful sign.

References

American Statistical Association (ASA) (2016), "ASA Statement on Statistical Significance and P-values," *The American Statistician*, 70, DOI: 10.1080/00031305.2016.1154108.

Cumming, G. (2012), *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis,* New York: Routledge.

Hauer, E. (2004), "The Harm Done by Tests of Significance," *Accident Analysis and Prevention*, 36, 495–500.

Lang, J., Rothman, K. J., and Cann, C. I. (1998), "That Confounded P-Value," *Epidemiology*, 9, 7–8.

Poole, C. (2001), "Low P-values or Narrow Confidence Intervals: Which are More Durable?" *Epidemiology*, 12, 291–294.

Schmidt, M., and Rothman, K. J. (2014), "Mistaken Inference Caused by Reliance on and Misinterpretation of a Significance Test," *International Journal of Cardiology*, 177, 1089–1090.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Kenneth J. Rothman, RTI International–Health Solutions, 3040 East Cornwallis Road, Research Triangle Park, NC 27709-2194 (Email: krothman@rti.org).

©2016 Kenneth J. Rothman

Are P-Values the Problem?

Stephen Senn

I welcome the ASA report. In my discussion I shall use *direct interpretation* to refer to P-values as probabilities of (functions of) statistics given hypotheses and inverse interpretation to probabilities of hypotheses (or parameters) given statistics.

The ASA report is suitably measured and cautious when discussing P-values, reflecting, in my opinion, a view, expressed by several of those consulted, that many of the problems of which P-values are accused are problems of inference generally, not problems of P-values per se. Among these problems of inference are the purely scientific such as

1. Inference is difficult
2. The inferential content of (say) an experiment cannot easily be summarized in one statistic (recognized in the final sentence of the ASA statement)

but also problems of human psychology and society such as

3. An impatience with the necessary nuances of expression that good statistical reporting requires
4. The (usual) prejudice of scientific journals in favor of "positive" results (Senn 2013)
5. The common habit of transforming shades of gray into either black or white
6. The desire of individual scientists for recognition and reward.

My view is that P-values are statistics that play a definite but limited role in statistical inference (Senn 2001), that inferences will be all the better for recognising their limitations but will be worse if we attempt to replace them rather than supplement them.

In fact, I claim that much current criticism of P-values is misplaced and reflects a false history of how they came about. A common story is as follows

1. For over 125 years scientists were happily calculating posterior probabilities using (what are now called) Bayesian approaches
2. Starting in the 1920s RA Fisher(Fisher 1925) persuaded them to calculate P-values instead
3. Because Fisherian P-values overstate "significance" they became very popular with scientists
4. We need to return scientists to the path of Bayesian virtue.

But this history is false. Tail area probabilities were being calculated well before Fisher but were given an inverse interpretation. Student's famous paper (Student 1908) gives an example. They were also occasionally given the modern direct interpretation provided in the informal definition in the ASA statement. See, for example, Karl Pearson's chi-square paper (Pearson 1900). Fisher pointed out that inverse interpretation was highly dependent on the assumed prior distribution. The subsequent neo-Bayesian revolution proved him right. He stressed the direct interpretation as being safer. Of course he introduced a whole host of statistical techniques, including many tests, but his most influential technical innovation as regards P-values per se was to suggest a doubling. This is not uncontroversial, but, since it increases the P-value, cannot be represented as giving significance more easily than what went before(Senn 2015b).

The origin of the modern claim that P-values give significance too easily is that the direct interpretation is compared to an indirect interpretation using a *different* Bayesian system: not the one developed by Laplace (Laplace 1951) but that which Harold Jeffreys (Jeffreys 1961) developed between the two wars (Senn 2015a). The response of Jeffreys to the challenge Broad made in showing that the Laplacian formulation could not provide an appreciable *posterior* probability of a scientific law being true (Broad 1918) was to place a lump of *prior* probability on its being true. (An early alternative development by (Haldane 1932) has been noted by (Etz and Wagenmakers 2015)).

The distinction is mainly between testing *point and dividing* hypotheses. (Cox (1977) used *plausible* for the former.) In the frequentist framework it makes very little difference which you use; in the Bayesian it is crucial (Casella and Berger 1987). However, many (see, e.g., Colquhoun 2014) have implicitly assumed that in re-calibrating direct inferences as inverse ones, no recalibration of significance levels is necessary. In my view this is as false as to argue, in moving to dosing patients by weight limit rather than age limit, that an age group that was 10 years and older should now become 10kg and heavier (Senn 2001).

P-values should be retained for a limited role as part of the machinery of error-statistical approaches. Even within that system they need to be supplemented by other devices (Mayo 1996). This system is valuable precisely because it is independent of the Bayesian one and trying to make it more Bayesian in behavior misses the point.

I am not arguing against Bayesian inference. I am arguing that it is valuable when those employing it pay very careful attention to appropriate specification of prior distributions making explicit the (prior) evidence on which these rest. This turns out to be much more difficult than many suppose (Senn 2011, 2007). In my opinion, it is advisable to go beyond default Laplacian or Jeffreys (point) prior distributions and certainly there is no point in modifying P-values to make them more "Bayesian."

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Stephen Senn, Luxembourg Institute of Health Ringgold Standard Institution – Competence Center for Methodology and Statistics, 1a rue Thomas Edison, Strassen 1445, Luxembourg, (Email: stephen.senn@lih.lu). This work was supported by the European Union's 7th Framework Programme for research, technological development and demonstration under grant agreement no. 602552. "IDEAL"

©2016 Stephen Senn

In short, the problem is less with P-values per se but with making an idol of them. Substituting another false god will not help (Gigerenzer and Marewski 2015).

I welcome the ASA statement as a sensible and measured contribution to improving the use of statistical inferential methods.

References

Broad, C. (1918), “On the Relation Between Induction and Probability,” *Mind* no. 27, 389–404.

Casella, G., and Berger, R.L. (1987), “Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem,” *Journal of the American Statistical Association*, 82, 106–111.

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

Cox, D.R. (1977), “The Role of Significance Tests,” *Scandinavian Journal of Statistics*, 4, 49–70.

Etz, A., and Wagenmakers, E.-J. (2015), “Origin of the Bayes Factor,” arXiv preprint arXiv:1511.08180.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

Gigerenzer, G., and Marewski, J.N. (2015), “Surrogate Science The Idol of a Universal Method for Scientific Inference,” *Journal of Management*, 41, 421–440.

Haldane, J.B.S. (1932), “A Note on Inverse Probability,” Paper read at Mathematical Proceedings of the Cambridge Philosophical Society.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford: Clarendon Press.

Laplace, P.S. (1951), *A Philosophical Essay on Probabilities* (English translation), Toronto: Dover.

Mayo, D. (1996), *Error and the Growth of Experimental Knowledge*, Chicago: University of Chicago Press.

Pearson, K. (1900), “On the Criterion that a Given Sytem of Deviation from the Probable in a Correlated System of Variables is Such that it can Reasonably Supposed to have Arisen from Random Sampling,” *Philosophical Magazine*, 50, 157–175.

Senn, S.J. (2001), “Two Cheers for P-values,” *Journal of Epidemiology and Biostatistics*, 6, 193–204.

——— (2007), “Trying to be Precise About Vagueness,” *Statistics in Medicine*, 26, 1417–1430.

——— (2011), “You May Believe You are a Bayesian but You are Probably Wrong,” *Rationality, Markets and Morals*, 2, 48–66.

——— (2013), *Authors are also Reviewers: Problems in Assigning Cause for Missing Negative Studies* . Available from *http://f1000research.com/articles/2-17/v1*.

——— (2016), *Double Jeopardy?: Judge Jeffreys Upholds the Law* 2015a [cited 13 February 2016]. Available from *http://errorstatistics.com/2015/05/09/stephen-senn-double-jeopardy-judge-jeffreys-upholds-the-law-guest-post/*.

——— (2016), *The Pathetic P-Value* 2015b [cited 13 February 2016 2016]. Available from *http://errorstatistics.com/2015/03/16/stephen-senn-the-pathetic-p-value-guest-post/*.

Student (1908), “The Probable Error of a Mean,” *Biometrika*, 6, 1–25.

Comment

Dalene Stangl

This author hopes that this discussion of p-values stirs the statistical education community to do a thorough self-evaluation of its course content and teaching. The normative misuse of p-values as a simplistic one-size-fits-all decision rule has always been problematic, and it is our community's responsibility to teach why using the p-value in this way is problematic. It is also our responsibility to teach more suitable alternatives (e.g., Bayesian decision-theoretic methods) despite the difficulty of doing so. To date, we in the statistics education community have been resistant to this consideration, instead focusing on finding conceptually easier ways of producing p-values (e.g., resampling and randomization methods), hence missing the bigger problem and perpetuating misuse.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Dalene Stangl, Duke University, Box 90251, Durham, NC 27708-0251 (Email: dalene@stat.duke.edu).

©2016 Dalene Stangl

The Value of p-Values

P.B. Stark

I agree with the spirit of the ASA p-value statement, but I disagree with some of the content, for instance:

- The informal definition of a p-value at the beginning of the document is vague and unhelpful.[1]
- The statement draws a distinction between "the null hypothesis" and "the underlying assumptions" under which the p-value is calculated. But the null hypothesis *is* the complete set of assumptions under which the p-value is calculated.
- The "other approaches" section ignores the fact that the assumptions of some of those methods are identical to those of p-values. Indeed, some of the methods use p-values as input (e.g., the False Discovery Rate).
- The statement ignores the fact that hypothesis tests apply in many situations in which there is no parameter or notion of an "effect," and hence nothing to estimate or to calculate an uncertainty for.
- The statement ignores the crucial distinction between frequentist and Bayesian inference.[2]

I offer the following plainer-language alternative:

Science progresses in part by ruling out potential explanations of data. p-values help assess whether a given explanation is adequate. The explanation being assessed is often called "the null hypothesis."[3]

If the p-value is small, either the explanation is wrong, or the explanation is right but something unlikely happened—something that had a probability equal to the p-value.[4] Small p-values are stronger evidence that the explanation is wrong: the data cast doubt on that explanation.

If the p-value is large, the explanation accounts for the data adequately—although the explanation might still be wrong.[5] Large p-values are not evidence that the explanation is right: lack of evidence that an explanation is wrong is not evidence that the explanation is right. If the data are few or low quality, they might not provide much evidence, period.

There is no bright line for whether an explanation is adequate: scientific context matters.

A p-value is computed by *assuming* that the explanation is right. The p-value is *not* the probability that the explanation is right.[6]

p-values do not measure the size or importance of an effect, but they help distinguish real effects from artifacts. In this way, they complement estimates of effect size and confidence intervals.

Moreover, p-values can be used in some contexts in which the notion of "effect size" does not make sense. Hence, p-values may be useful in situations in which estimates of effect size and confidence intervals are not.

Like all tools, p-values can be misused. One common misuse is to hunt for explanations that have small p-values, and report only those, without taking into account or reporting the hunting. Such "p-hacking," "significance hunting," selective reporting, and failing to account for the fact that more than one explanation was examined ("multiplicity") can make the reported p-values misleading.

Another misuse involves testing "straw man" explanations that have no hope of explaining the data: null hypotheses that have little connection to how the data were collected or generated. If the explanation is unrealistic, a small p-value is not surprising. Nor is it illuminating.

Many fields and many journals consider a result to be scientifically established if and only if a p-value is below some threshold, such as 0.05. This is poor science and poor statistics, and creates incentives for researchers to "game" their analyses by p-hacking, selective reporting, ignoring multiplicity, and using inappropriate or contrived null hypotheses.

Such misuses can result in scientific "discoveries" that turn out to be false or that cannot be replicated. This has contributed to the current "crisis of reproducibility" in science.

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. P.B. Stark, University of California, Berkeley–Statistics, Berkeley, CA 94720-3860 (Email: pbstark@berkeley.edu).

[1] See footnote 4 below. The reference to "extreme" values of "a statistical summary" limits the scope to tests based on a test statistic. It is an inaccurate and confusing substitute for a simpler statement about monotonicity (i.e., nesting) of rejection regions.

[2] The document has other problems, among them: It characterizes a p-value of 0.05 as "weak" evidence against the null hypothesis, but strength of evidence depends crucially on context. It categorically recommends using multiple numerical and graphical summaries of data, but there are situations in which these would be gratuitous distractions—if not an invitation to p-hacking!

[3] The use of the term "null hypothesis" is not entirely consistent, but in general, the null hypothesis asserts that the probability distribution $\mathbb{P}$ of the data X is in some specified set $\mathcal{P}$ of probability distributions on a measurable space $\mathcal{X}$. A "point null hypothesis" or "simple null hypothesis" completely specifies the probability distribution of the data, i.e., $\mathcal{P}$ is a singleton set. In the context of testing whether some parameter θ is equal to θ_0, some authors write $H_0 : \theta = \theta_0$ as the null hypothesis. This is (perhaps not deliberate) shorthand for the hypothesis $X \sim \mathbb{P}_{\theta_0}$, where $\{\mathbb{P}_\theta\}_{\theta\in\Theta}$ is a pre-specified family of probability distributions on $\mathcal{X}$ that depends on a parameter θ known a priori to be in the set Θ, which contains θ_0.

[4] The simplest general definition of a p-value of a point null hypothesis I know of is as follows. Suppose the null hypothesis is that $\mathbb{P}$ is the probability distribution of the data X, which takes values in the measurable space $\mathcal{X}$. Let $\{R_\alpha\}_{\alpha\in[0,1]}$ be a collection of $\mathbb{P}$-measurable subsets of $\mathcal{X}$ such that (1) $\mathbb{P}(R_\alpha) \le \alpha$ and (2) If $\alpha' < \alpha$ then $R_{\alpha'} \subset R_\alpha$. Then the p-value of H_0 for data $X = x$ is $\inf_{\alpha\in[0,1]}\{\alpha : x \in R_\alpha\}$.

[5] Here, "adequately" is with respect to the chosen test.

[6] This is a common misinterpretation. Other misinterpretations are that 1 minus the p-value is the probability that the *alternative hypothesis* (a different explanation of the data) is true, and that the p-value is the probability of observing the data "by chance."

©2016 P.B. Stark

The Significance of the ASA Statement on Statistical Significance and P-Values

Stephen T. ZILIAK

Little p-value
What are you trying to say
Of significance?

Far more in science and society than it positively should. That is one way to express the point of the American Statistical Association Statement on Statistical Significance and P-Values (ASA 2016). The cost of incorrect interpretation of Student's t-statistics, Fisher's p-values, and other tests of statistical significance has been shown to be unsustainably large (for a large scale survey see Ziliak and McCloskey (2008)). The ASA Statement on Statistical Significance and P-values is going to help.

Content aside, it is worth noting that the Statement emerged from a humane, Socratic, and Tocquevillean model of democracy and dialectic. (I'm an economist so I know the difference.) The Statement has seen more discerning eyes than a model on a catwalk. Statistician eyes, scientist eyes, journalist and business and Bayesian eyes, over and over. Still it stands, I think most will agree. Hats off to Ron Wasserstein and the ASA Board, who made openness and transparency, widespread democracy and dialectic the chief virtues of the drafting process. Over the course of the past year the Statement on Statistical Significance has evolved with the benefit of constant counsel from leading statisticians and scientists worldwide. Thus the Statement can be treated as a repeated and unbiased sample of best practice thinking about statistical significance (plus or minus a small error).

Despite occasional disagreements—some of them fundamental to the philosophy of science—the drafting Committee did what pundits and skeptics alike thought impossible. Together we agreed that the current culture of statistical significance testing, interpretation, and reporting has to go, and that adherence to a minimum of six principles can help to pave the way forward for science and society. Adherence to principles (2) through (6) will be productive, most of us believe, of a steady and rising stream of large net benefits to more than science. In economic policy. In health and drugs and medicine. And in every realm of life, from agronomy to zoology, including law, that is touched by the test of statistical significance.

Some, for example financial investors and publishing scientists, could begin immediately to reap the benefits of change implied by this Statement. There is a hunger for change among journal editors and referees; among grantors and journalists, lawyers, and decision-makers. Virtually no one is happy with mushy p's though they, as I and others have shown, are treated like the main dish of science. In abbreviated form I believe the most important principles for the much-anticipated paradigm shift are:

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.*"

The null hypothesis stipulating no numerical or practical difference between object A and object B may be true even when the p-value is low, below 0.05, for example. And an alternative hypothesis and effect size $A > B$ may be for legal or medical or commercial purposes important even when the p-value takes on higher values. The p-value is not an error-probability. And it doesn't measure the probability of a hypothesis given the evidence. Not even when searching for the Higgs boson or Einstein's ripples, through a lot of evidence. The ripples and boson probably exist. But the p-value or other test of statistical significance does not prove the role of random chance.

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

Thus bright-line rules of acceptance and rejection, such as the false equations of $p > 0.05 =$ "insignificant, accept the null" and $p < 0.05 =$ "significant, reject the null" should be in most cases banished. (Many Presidents of the American Statistical Association have long argued such, from Kruskal and Zellner to Morganstein and Utts.) Recently the Supreme Court of the United States unanimously agreed in Matrixx v. Siracusano, 9-0, that statistical significance is neither necessary nor sufficient (Bialik 2011). Statistical significance does not mean scientific or business or policy "importance." And lack of statistical significance according to an arbitrary line and custom does not mean lack of importance. As Savage (1954, p. 116) noted long ago, "The cost of an observation in utility may be negative as well as zero or positive; witness the cook that tastes the broth"

4. "*Proper inference requires full reporting and transparency.*"

For example, if the published regression models are the result of dropping and adding variables until the t, p, R-squared and other values reach a certain level of significance, the published report should be transparent and say so. Drawing on the work of Committee members and others, Ziliak (2016) describes a number of easy-to-adopt changes to improve inferences with the style of the scientific research paper.

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

Student's t-statistic is a signal-to-noise ratio, and the p-value

Online discussion of the ASA Statement on Statistical Significance and P-Values, *The American Statistician*, 70. Stephen T. Ziliak, Roosevelt University College of Arts and Sciences–Economics, 430 S. Michigan Ave, Chicago, IL 60605-1313 (Email: sziliak@roosevelt.edu).

©2016 Stephen T. Ziliak

is the probability of observing a t-statistic equal to or larger than the one you see in the data, assuming to be true the null hypothesis and other data and modeling assumptions. The point here is that in p, t, or other form, the signal-to-noise ratio is not telling us what we want mainly to know: the answer to the expected size-matters/how much question. William Sealy Gosset aka "Student" (1876–1937) himself would agree, and would feel that his test of significance has been much abused. In a letter of 1905 to Karl Pearson the Guinness brewer and pioneer of small sample theory and applications said:

> When I first reported on the subject [of "The Application of the 'Law of Error' to the Work of the Brewery" (1904)], I thought that perhaps there might be some degree of probability which is conventionally treated as sufficient in such work as ours and I advised that some outside authority [in mathematics, such as Karl Pearson] should be consulted as to what certainty is required to aim at in large scale work. However it would appear that in such work as ours the degree of certainty to be aimed at must depend on the *pecuniary advantage to be gained by following the result of the experiment, compared with the increased cost of the new method, if any, and the cost of each experiment* (quoted in Ziliak [2008]).

Suppose diet pills Oomph and Precision are priced the same and bring the same side-effects. Oomph promises to remove 20 pounds on average but it is uncertain in actual effect, at plus or minus 10 pounds on either side of 20. Pill Precision promises a 5-pound weight-loss but its variance is found in well-designed studies to be much lower, at plus or minus 0.5 pounds. What pill is best? The signal-to-noise ratio of pill Oomph is $2(20/\pm 10)$ while for pill Precision the signal rises five times higher, to an impressive $10(5/\pm 0.5)$. Pill Precision, though more "significant," has, so to speak, no oomph. Precision works at best less effectively than Oomph at its worst. When choosing between two diet pills—between two tax policies, two blood-thinning medicines, or two paths for climate change—the signal-to-noise ratio—Student's t—is not the point. The point, what matters most times, is the expected size and meaning of uncertain effects across the whole distribution. What pill or path you favor should depend on the expected size and net value—the expected loss function—of acting as if the favored pill or hypothesis is true. Begin by not favoring the test of statistical significance.

References

American Statistical Association (2016), "Statement on Statistical Significance and P-Values," *The American Statistician*, 70.

Bialik, C. (2011), "Making a Stat Less Significant," *The Wall Street Journal*, April 11th, The Numbers Guy column. *http://www.wsj.com/articles/SB10001424052748703712504576235683249040812*.

Savage, L. (1954), *The Foundations of Statistics*, New York: Dover.

Ziliak, S. (2008), "Guinnessometrics: The Economic Foundation of 'Student's' *t*," *Journal of Economic Perspectives*, 22, 199–216.

——— (2016), "Statistical Significance and Scientific Misconduct: Improving the Style of the Published Research Paper," *Review of Social Economy*, 74 (1), 83–97. *http://dx.doi.org/10.1080/00346764.2016.1150730*.

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