HHS Public Access

Author manuscript

Ann Neurol. Author manuscript; available in PMC 2018 June 01.

Published in final edited form as:

Ann Neurol. 2017 March ; 81(3): 344–347. doi:10.1002/ana.24868.

Is This Significant?

Rebecca A. Betensky, Ph.D. [Statistical Editor]

Annals of Neurology

The American Statistical Association (ASA) recently published a statement on p-values and statistical significance in response to "highly visible discussions" in recent years concerning the rigor of scientific publications.[1] These discussions have questioned the validity of statistical hypothesis testing and raised concerns about reproducibility and replicability of scientific results. One psychology journal even banned significance testing, p-values and confidence intervals.[2] The ASA had never previously issued a statement on matters of statistical practice, but felt compelled to do so in this instance to attempt to clarify "an aspect of our field that is too often misunderstood and misused" to a general audience of "researchers, practitioners and science writers." There is nothing new in the statement, but rather it is an attempt at education at a moment in which there appears to be widespread misinterpretation. In that spirit, I aim to communicate the content of the statement to the readers and contributing authors of *Annals of Neurology* and attempt to translate it into expectations of the journal for contributing authors.

Background

A scientific hypothesis is a theory; in medicine it often is about how an intervention may affect an outcome. For example, you may hypothesize that taking a statin may reduce the risk of Parkinson's disease. The hypothesis is used to make a prediction that a particular dependent variable (e.g., age at onset of PD) will change with a particular intervention (taking a statin drug). Often we assess the hypothesis within the context of a specific model for the dependent variable, such as a Cox proportional hazards model for onset of PD that is a function of statin use and potential confounders such as family history, age and certain genotypes. One estimable quantity based on this model is the hazard ratio for onset of PD for statin users as compared to non-users. Statistical hypothesis testing begins with specification of a "null hypothesis." This is most commonly a hypothesis of no difference (i.e., no effect) or of randomness (e.g., the hazard ratio for onset of PD that compares statin users to non-users is one). Importantly, the null hypothesis is the default or strawman position, which is of interest to disprove. A p-value is the probability of obtaining the observed data, or data even more extreme in its opposition to the null hypothesis, under the

Editor's Note: One of the problems that Editors and reviewers constantly face is to determine whether a finding in a research report is significant, both from a statistical and from a biological/medical standpoint (not the same thing). Over-reliance on the p-value, and an arbitrary threshold of $P<0.05$, can result in the misinterpretation of both the statistical and biological significance of the finding. If these statements seem odd to you, you are among the vast majority of my research colleagues who would benefit from reviewing the recent statement by the American Statistical Association on the meaning of p-values, and how they should be used. Having read through their document, though, I can see that this might be tough going for many neurologists, and so our Statistical Editor for Annals of Neurology, Rebecca Betensky,PhD, has volunteered to take you on a brief tour of this statement. If you use statistics at all in your work, this will be well worth your while. – Clifford B. Saper, MD, PhD, Editor-in-Chief

assumption that the null hypothesis is true. If this probability, which is calculated under the assumptions of the statistical model and the null hypothesis, is very small, indicating that the observed data are very unlikely under these assumptions, then one of three conclusions is possible: (1) the statistical model is not correct; (2) the null hypothesis is not correct; or (3) a rare event occurred to produce the observed data. Because we generally believe in our statistical model and we generally do not believe that rare events actually occur, we embrace the conclusion that the null hypothesis must not be correct and we "reject" it. A declaration of statistical significance is made when the p-value is smaller than some threshold (often set at 0.05) and is taken to indicate that the observed data provide "real" evidence in favor of rejection of the null hypothesis.

ASA principles

The ASA's statement is comprised of six concise principles, along with some explanation of each. I list them below, and include my explanations and translations.

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

   A p-value is a probability, which takes values between 0 and 1. The smaller the p-value, the less compatible the data are with the specified statistical model, meaning the model (e.g., normal distributions) in conjunction with the null hypothesis assumption on its parameters (e.g., equality of means). If the model is not in question, as is usually the case, then the smaller the p-value, the less compatible the data are with the null hypothesis assumptions on the parameters.

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."

   This is a common misconception. A p-value is not a probability of the truth of a null hypothesis or of random chance; it is a probability of the observed data (or data even more extreme) under the assumption of the null hypothesis.

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

   "Bright-line" rules such as $p<0.05$ for scientific conclusions can be erroneous. This is because p-values depend on many factors besides the parameters of interest. They are strongly dependent on sample size, variability of underlying measurements and the assumed statistical models. For example, different size data samples that exhibit the same effect size can lead to very different p-values, with large sample sizes yielding smaller p-values. This is true also for data that exhibit the same effect size with different standard deviations; smaller standard deviations yield smaller p-values. It is also misleading to treat an arbitrary threshold, such as 0.05, as having any scientific authority that would require dismissal of a p-value of 0.07 and embrace of a p-value of 0.04.

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

This is a critical point for authors. Data analysis is a process that often involves exploration because researchers may not know in advance of seeing the data what to expect and thus what statistical tests to conduct. Instead, they often analyze the data and then select statistical tests that reflect the salient features of the data. However, if the analytical strategy and process is not identified before analyzing the data, the reported p-values are not interpretable. This is because they are selected from a larger set of comparisons that could be made, by criteria that have been optimized to yield small p-values. As such, the p-value no longer can be interpreted as the simple probability of obtaining the observed data (or data more extreme), but rather must be interpreted in light of the fact that the probability has already been optimized. This optimization arises in common approaches such as subgroup analysis, variable selection, model selection, and threshold selection for categorization of continuous data. It is unaffectionately labeled as "cherry picking," "fishing," "p-hacking"[3], or "data dredging." The ASA statement advocates full reporting by authors of all hypotheses explored, all data collection decisions, all statistical analyses conducted and all p-values computed. This is a first and important step toward correcting the problem of covert p-value optimization, but truthfully does not offer a real solution for its full correction, because even with full transparency, selection of tests conducted and models fit still occur.

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

This is because p-values are functions of sample size and variability. It is also because the importance of an effect of a particular magnitude is a substantive matter and not a statistical matter. It is common, particularly with large datasets, to have very small differences that may reach a p-value threshold, such as 0.05. For example, how important would it be if a statin drug were found to delay median onset of PD by one month? Would you commit a very large cohort of at-risk individuals to take a drug for such a small benefit? We frequently see papers at Annals where a biomarker is statistically significantly elevated at the threshold of $p<0.05$, but the individual values for affected and control individuals overlap extensively. So, the marker may be statistically validated, but of no use in identifying individuals who are affected.

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

Again, this is because of its dependence on sample size and variability and the need for substantive interpretation. It is also because of its dependence on the assumed model; while a large p-value indicates low compatibility of the data with the assumed model and null hypothesis, there may be other models that have not been considered that could be even more compatible with the observed data. For this reason, large p-values generally do not provide evidence in favor of the null hypothesis.

Implications for authors

1. Authors should be accurate in their language surrounding p-values.
2. Authors should not describe results with p-values that are small, but exceed 0.05, as being null. That is, if a mean difference has a p-value of 0.07 the authors should not state that there was no difference, but rather they should provide the estimated mean difference, its 95% confidence interval, the p-value, and substantive interpretation. At best they can state that they could not reject the null hypothesis at a 0.05 level of significance.
3. Authors should not conclude from large p-values that the null hypothesis is true. While large p-values indicate that the data exhibit some compatibility with the null hypothesis, there is no guarantee that it is true. There may be other hypotheses that are even more compatible with the data.
4. Statistical Methods sections should fully report the analytical strategy that was pre-specified and analyses that were conducted. Results sections should include all results obtained and not just those with small p-values. Replication cohorts are particularly valuable if the statistically significant differences in the first cohort were not pre-specified. This is now standard practice in some settings, e.g., in genome-wide association studies, where very large numbers of comparisons are pre-specified (30,000 genes).
5. Authors should never report p-values in isolation, even in the Abstract; they should be accompanied by effect estimates and ideally confidence intervals.
6. Studies should be powered to detect effects that are of biological interest, i.e., if a 10% reduction in mortality over 5 years is a sufficiently large biological response, then enough subjects should be included to detect this size effect with high probability
7. Sample sizes should be sufficiently large to establish “null effects” that are clearly smaller than the level considered to be of minimum clinical importance. This requires pre-specification of the magnitude of the minimal meaningful clinical effect and calculation of the sample size that would yield a confidence interval that does not cover this effect size while also covering the null. For example, if a study is 95% powered to detect a particular effect size(e.g., life extension by one year for subjects taking a statin drug), then if the true effect size equals or exceeds that level, the null hypothesis will be rejected with 95% probability. On the other hand, if the 95% confidence interval for the true effect size covers zero, it necessarily does not cover the effect size of clinical interest, and this level can be ruled out. It is a useful contribution to the scientific literature to present definitive null results.

Conclusions

The ASA statement conveys a few simple messages and reminders to all users of p-values. P-values are continuous measures and should not be dichotomized according to an arbitrary

threshold. Both small and large p-values need to be interpreted in conjunction with estimates of effect. Small p-values provide useful scientific knowledge if their associated estimated effects are clinically or scientifically meaningful. Large p-values indicate that there is insufficient evidence to reject the null hypothesis. In studies that are highly powered to detect effects that are scientifically meaningful, large p-values can provide evidence against those effects. If a study has 95% power to detect a meaningful effect, then a p-value that is larger than 0.05 translates into statistical evidence in opposition to that effect. However, if a study has 50% power to detect a meaningful effect, then a p-value must be larger than 0.5 to provide evidence in opposition to that effect. This highlights the importance of designing and conducting studies that are sufficiently highly powered to detect meaningful effects; when this is done, something meaningful is learned, whether positive or null.

Data-driven analyses undermine the p-value.[4] These undertakings are ubiquitous and essential in the exploration of data for new knowledge. Researchers need to understand that p-values lose their simple interpretation whenever a decision is made about a subsequent analysis that is based on results of current analyses. Correction for this is complicated and simple approaches are often not useful due to their over-correction (e.g. Bonferroni). Some authors in Political Science have suggested requirements for registration of analysis plans prior to data collection, analogous to the requirement for registration of clinical trials.[5] Others suggested a two-stage analysis process, in which the first stage of exploration informs registration of the analysis plan for the second stage (i.e., the replication cohort).[6] At very least, authors must fully report their approach in Methods sections, and all results in Results sections (or appendices).

The ASA statement reminds us all that while p-values are just probabilities, their interpretation can be non-intuitive and complicated for many researchers, which leads to misleading scientific conclusions. It is useful for all researchers and readers of the scientific literature to be reminded of the definition and proper uses of p-values, and to be wary of their many potential misinterpretations.[7] Journals, including *Annals of Neurology*, should be vigilant in requiring that strong evidence for positive or null results, as measured by effect sizes and confidence intervals, in conjunction with p-values, guide publication decisions. Journals also need to be vigilant in their management of the language around p-values, and presentations of methods and results in their articles.

References

1. Wasserstein RL, Lazar NA. The ASA's Statement on p-Values: Context, Process and Purpose. American Statistician. published online June 9, 2016.
2. Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology. 2015; 37:1–2.
3. Simonsohn U, Nelson LD, Simmons JP. P-Curve: A Key to the File-Drawer. Journal of Experimental Psychology: General. 2014; 143:534–547. [PubMed: 23855496]
4. Benjamini Y. It's not the p-values' fault. Online Supplemental Material to "The ASA's Statement on p-Values: Context, Process and Purpose,". American Statistician. published online June 9, 2016.
5. Gelman A. Preregistration of Studies and Mock Reports. Political Analysis. 2013; 21:40–41.
6. Gelman A, Loken E. The statistical crisis in science. American Scientist. 2014; 102:460–465.
7. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical Tests, P-values, Confidence Intervals, and Power: A Guide to Misinterpretations. Eur J Epidemiol. 2016; 31:337–350. [PubMed: 27209009]