The American Statistician

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

The *p*-Value Requires Context, Not a Threshold

Rebecca A. Betensky

To cite this article: Rebecca A. Betensky (2019) The *p*-Value Requires Context, Not a Threshold, The American Statistician, 73:sup1, 115-117, DOI: 10.1080/00031305.2018.1529624

To link to this article: https://doi.org/10.1080/00031305.2018.1529624

THE AMERICAN STATISTICIAN
2019, VOL. 73, NO. S1, 115–117: Statistical Inference in the 21st Century
https://doi.org/10.1080/00031305.2018.1529624

The *p*-Value Requires Context, Not a Threshold

Rebecca A. Betensky

Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA

ABSTRACT

It is widely recognized by statisticians, though not as widely by other researchers, that the *p*-value cannot be interpreted in isolation, but rather must be considered in the context of certain features of the design and substantive application, such as sample size and meaningful effect size. I consider the setting of the normal mean and highlight the information contained in the *p*-value in conjunction with the sample size and meaningful effect size. The *p*-value and sample size jointly yield 95% confidence bounds for the effect of interest, which can be compared to the predetermined meaningful effect size to make inferences about the true effect. I provide simple examples to demonstrate that although the *p*-value is calculated under the null hypothesis, and thus seemingly may be divorced from the features of the study from which it arises, its interpretation as a measure of evidence requires its contextualization within the study. This implies that any proposal for improved use of the *p*-value as a measure of the strength of evidence cannot simply be a change to the threshold for significance.

ARTICLE HISTORY
Received March 2018
Revised September 2018

KEYWORDS
Effect size; Sample size; Statistical significance.

1. Introduction

Seventy-two prominent researchers proposed changing the default *p*-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries (Benjamin et al. 2018). They were motivated to address a leading cause of nonreproducibility of scientific studies; standards of evidence such as the usual rule for formal inference, "reject if $p < 0.05$," are too low. This is not a new concern and was raised several years ago by Ionnidis (2005), who presented data and analysis in support of his title claim that most published research findings are false. The solution proposed by Benjamin et al. (2018) is simple, but surprisingly does not implement some of the basic tenets put forth in the recent statement published by the American Statistical Association (ASA) on the topic of statistical significance and *p*-values (Wasserstein and Lazar 2016). The third principle listed in the ASA statement asserts that scientific conclusions should not be based on whether a *p*-value passes a threshold. The fifth principle acknowledges that the *p*-value, in isolation, does not measure the effect size or the importance of a result. A better solution is not to change the threshold, as suggested by Benjamin et al. (2018), but to find an alternative to exclusive reliance on threshold alone. Context matters.

In this article, I use the terms design and context to refer to characteristics of experiments such as sample size and substantively meaningful effect size, which impact the interpretation of a *p*-value and the conclusions that are drawn. I then rely on examples from the simple setting of a single normal sample with variance one to articulate and illustrate two informal principles for interpreting *p*-values. A first pair of examples (Section 2) shows how data leading to a *p*-value of 0.005 as in Benjamin et al. (2018) can lead to different inferences depending on the combination of sample size and context-specific magnitude of an interesting or important effect size. These examples rely on a functional relationship between the observed *p*-value and sample size, and the lower endpoint of a one-sided confidence interval. A second pair of examples (Section 3) shows how data leading to a large *p*-value can also lead to different inferences, also depending on the combination of sample size and context-specific magnitude of an uninteresting or inconsequential effect size. These examples similarly rely on a relationship between the observed *p*-value and sample size, and the upper endpoint of a one-sided confidence interval. Although I rely on a simple setting for examples, the informal principles for interpreting *p*-values extend in a natural way to more general settings. The article concludes (Section 4) with a summary and discussion.

2. Interpreting a Small *p*-Value

The Benjamin et al. (2018) proposal calls for reducing the *p*-value threshold from 0.05 to 0.005 as a solution to nonreproducibility in science. This section presents two examples, both with $p = 0.005$. For the first example, $p = 0.005$ is too stringent of a threshold for detecting a meaningful signal. For the second example, $p = 0.005$ is not stringent enough. What distinguishes the two examples is the context, namely, the combination of sample size, n, and size, d, of the effect judged to be meaningful. Here, I assume that d has been identified; in practice, the identification of clinically or substantively meaningful effects is complicated and may not be consistent across the various stakeholders, including patients, clinicians, regulators,

CONTACT Rebecca A. Betensky betensky@hsph.harvard.edu Department of Biostatistics, Harvard T.H. Chan School of Public Health, 655 Huntington Avenue, Boston, MA 02115

investors, and payers (Keefe et al. 2013; Rosnow and Rosenthal 2003). By definition, d is nonzero.

It is well-known that there is a duality relating hypothesis tests and confidence intervals: we reject the null hypothesis, H_0 at level α if and only if the null value of the parameter lies outside the corresponding $1-\alpha$ level confidence interval. I rely here on a different correspondence, that between the p-value calculated for a one-sided test of $H_0 : \mu = 0$ versus the one-sided alternative hypothesis, $H_1 : \mu > 0$, and the endpoint μ_* of a one-sided confidence interval for μ of the form (μ_*, ∞). The extension to two-sided tests and confidence intervals is straightforward. Browne (2010) also elucidated the relationship between the p-value and the observed effect. He did not, however, relate the interpretation of the 95% confidence interval to a predetermined meaningful effect size as the basis for inference.

For the simple case of a random sample $X_1, X_2, \ldots, X_n$ from a normal distribution with known variance 1 and unknown mean μ, the value of the sample mean $\bar{x}$ determines both the p-value, denoted by p, and the lower endpoint of the interval, μ_*. For example, X might be the measured standardized change in systolic blood pressure from baseline to one year after some treatment and μ is the expectation of X. Of interest is the test of the hypothesis that the drug has a positive effect on blood pressure: $H_0 : \mu = 0$ versus $H_1 : \mu > 0$. The p-value is given by $p = P_{\mu=0}(Z > \sqrt{n}\bar{x})$, where Z is standard normal, $\mu_* = \bar{x} - Z_{1-\alpha}/\sqrt{n}$, and $Z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution. Inverting the equation that defines the p-value and solving for $\bar{x}$ yields that $\mu_* = (Z_{1-p} - Z_{1-\alpha})/\sqrt{n}$. Note that smaller p are associated with larger μ_*, and thus there is a threshold p^* such that p-values that are below p^* provide evidence that $\mu > d$, that is, of a non-null meaningful effect. In the systolic blood pressure example, d might be $10/\sigma$, where σ is the standard deviation of the change in blood pressure from baseline to one year. (These calculations follow from the fact that the sample mean $\bar{X}$ is normally distributed with mean μ and variance $1/n$ and so $\sqrt{n}(\bar{X} - \mu)$ is standard normal). In the setting of a two-sided test, with positive $\bar{x}$, $\mu_* = (Z_{1-p/2} - Z_{1-\alpha/2})/\sqrt{n}$ and a p-value threshold can likewise be derived.

Given the observed p-value, it is possible to calculate the lower endpoint μ_* of a one-sided $1-\alpha$% confidence interval for μ. In particular, if $p = 0.005$, the corresponding 0.995 quantile of a standard normal is $Z_{0.995} = 2.576$, and the lower endpoint of a 95% one-sided interval is $\mu_* = (2.576 - Z_{0.95})/\sqrt{n}$, where $Z_{0.95} = 1.645$. Thus, $\mu_* = (2.576 - 1.645)/\sqrt{n}$.

Now consider two examples, both with $p = 0.005$. I take as context the combination of sample size, n, and the meaningful effect size, d, defined as the smallest value of $\mu > 0$ judged to be meaningful. Note that the sample size might have been selected to attain a certain level of power to detect a particular value of μ. If it is important to fix the power at a certain level, for practical considerations such as availability of subjects and cost of the trial, this value of μ is frequently larger than d, the smallest *meaningful* value of μ. However, if it is important to design the study to detect the meaningful effect, d, it may be underpowered given the constraints of subject availability and cost.

Example 1(a): Suppose that an effect size of $d = 0.10$ is considered meaningful, and that the sample size is $n = 50$. Given that $p = 0.005$, the lower endpoint of the one-sided 95% confidence interval is equal to $\mu_* = (2.576 - 1.645)/\sqrt{50} = 0.1317$ (Table 1). Thus, with 95% confidence, $p = 0.005$ excludes values of μ that are less than or equal to 0.1317, and thus certainly those that are less than 0.10. In this context, $p = 0.005$ identifies meaningful signals, but potentially misses some signals (i.e., those between 0.10 and 0.1317). The optimal (i.e., maximum) p-value threshold corresponding to $\mu_* = 0.1$ in this context is 0.0093.

Example 1(b): For a contrasting example, I increase the sample size to $n = 200$ and maintain the same effect size of $d = 0.10$. The same p-value yields a lower 95% confidence limit of 0.0658, which includes values of μ less than d. Here, $p = 0.005$ is not a useful threshold relative to the meaningful effect size as it admits values of μ less than 0.10. In this example, the optimal threshold is 0.0011.

Generalizing from these examples suggests a strategy for finding the p-value threshold for concluding a meaningful effect for any given sample size:

1. Based on substantive knowledge about the applied context, select a value d for the smallest effect size considered meaningful. While ideally this is the value that is used to design the study to achieve a fixed power, practical considerations often do not permit this.

2. Take as the upper p-value threshold that value p^*, for which $\mu_* = d$. That is, reject H_0 if and only if $p < p^*$, or equivalently, the 95% confidence interval for μ, (μ_*, ∞) excludes d.

The principle: Reject the null in favor of a meaningful effect if and only if the lower 95% confidence bound exceeds the smallest effect size considered meaningful. Thus, rejecting the null means we can be 95% confident that the true effect size is at least as large as the size considered to be clinically meaningful.

As an example, consider the 1993 GUSTO-I study of streptokinase plus intravenous heparin versus rt-PA (recombinant tissue plasminogen activator) plus intravenous heparinthrombolytic drugs for acute myocardial infarction, as discussed by Lesaffre (2008). The primary endpoint was 30-day mortality. There were approximately 10,300 subjects in each of these treatment arms, and the observed percentages of 30-day mortality were 7.4% and 6.3%. The two-sided p-value testing the equality of the percentages was 0.0028, with a 95% confidence interval for the difference of (0.36%, 1.73%). The conclusion was that there was a significant reduction in 30-day mortality advantage for the rt-PA group versus the streptokinase plus intravenous heparin group. This conclusion implies that a difference as small as 0.36% is considered to be clinically meaningful (i.e., $d < 0.0036$). If this is not the case, and $d > 0.0036$, then even the small p-value of 0.0028 does not provide strong evidence of a meaningful effect.

3. Interpreting a Large *p*-Value

In the previous section, I illustrated that a small p-value relative to a fixed threshold has different meanings depending on the context. I now consider what can be learned from large p-values. Students of introductory statistics courses are taught that no conclusions can be drawn from large p-values. This maxim was reiterated in the ASA statement (Wasserstein and Lazar

2016). In this section, I illustrate that large p-values relative to a fixed threshold also have different meanings depending on the context.

Just as the lower confidence limit for the normal mean has a direct relationship with the p-value for the same one-sided test $H_0 : \mu = 0$ versus $H_1 : \mu > 0$, so does the upper limit of a one-sided confidence interval $(-\infty, \mu^*)$. In particular, simple algebra yields that $\mu^* = (Z_{1-p} + Z_{1-\alpha})/\sqrt{n}$. Note that larger p are associated with smaller μ^*, and thus there is a threshold p_* such that p-values that exceed p_* provide evidence that $\mu < d$, that is, of a nonmeaningful effect. In the setting of a two-sided test, with positive $\bar{x}$, $\mu^* = (Z_{1-p/2} + Z_{1-\alpha/2})/\sqrt{n}$ and a p-value threshold can likewise be derived.

Now consider two examples, both with $p = 0.6286$. Again, I take as context the combination of sample size, n, and the effect size, d, defined as the smallest value of $\mu > 0$ judged to be meaningful.

Example 2(a): Suppose that an effect size of $d = 0.10$ is considered meaningful, and that the sample size is $n = 50$. Given that $p = 0.6286$ and $Z_{1-p} = -0.328$, the calculations above yield that the upper endpoint of the one-sided 95% confidence interval is equal to $\mu_* = (-0.328 + 1.645)/\sqrt{50} = 0.1862$. Thus, with 95% confidence, $p = 0.6286$ excludes values of μ that are greater than or equal to 0.1862, but is uninformative about whether μ is less than $d = 0.10$ or not. In this example, a lower p-value threshold of $p_* = 0.8259$ would provide evidence of a nonmeaningful effect (i.e., $\mu < d$).

Example 2(b): For a contrasting example, I increase the sample size to $n = 200$ and maintain the same effect size of $d = 0.10$. The same p-value yields an upper 95% confidence limit of 0.0931, which excludes values of μ greater than $d = 0.10$. Here, the large p-value of 0.6286 is useful in providing evidence against a meaningful effect. In this example, a lower p-value threshold of $p_* = 0.5913$ would be sufficient to provide evidence of a nonmeaningful effect.

Generalizing from these examples suggests a strategy for finding the p-value threshold for concluding a nonmeaningful effect for any given sample size:

1. Based on substantive knowledge about the applied context, select a value d for the smallest effect size considered meaningful. While ideally this is the value that is used to design the study to achieve a fixed power, practical considerations often do not permit this.
2. Take as a lower p-value threshold that value, p_*, for which $\mu^* = d$. That is, accept H_0, that is, conclude no meaningful effect, if and only if $p > p_*$, or equivalently, the 95% confidence interval for μ, $(-\infty, \mu^*)$ excludes d.

The principle: Accept the null with respect to a prespecified d if and only if the upper 95% confidence bound falls below the smallest effect size considered meaningful. Thus, accepting the null means we can be 95% confident that the true effect size is no larger than the minimal size considered to be clinically meaningful.

As an example in the different context of a two-sided test of a relative risk, the RE-LY trial of atrial fibrillation compared dabigratran to warfarin with respect to risk of stroke or systemic embolism (Connolly et al. 2009). A relative risk of 1.46 was identified as the clinically meaningful threshold for noninferiority of dabigatran relative to warfarin; that is, if the upper two-sided 95% confidence limit (i.e., the upper one-sided 97.5% limit) for the relative risk fell below 1.46, noninferiority could be declared. The upper one-sided 97.5% limit was used to account for the two dabigatran dose groups that were tested versus warfarin and because superiority was tested, as well. The relative risk for the 6015 subjects in the 110 mg dabigatran group versus the 6022 subjects in the warfarin group was 0.91, with a 95% confidence interval of (0.74,1.11) and a p-value of 0.34. Because the upper limit of 1.11 is below 1.46, this dose group of dabigatran could be concluded to be noninferior to warfarin. In this setting, the large p-value of 0.34 (and associated confidence interval) is large enough to declare noninferiority of dabigatran.

4. Summary

In conjunction with the design and context of the study, such as sample size and the minimum meaningful effect size, which are inputs to the calculation of confidence limits for measures of effect, the p-value may indeed be informative about the effect of interest and/or about the null. However, absolute thresholds for the p-value do not render it meaningful with regard to a positive or null effect; the thresholds depend on n and d. This understanding expands on the ASA statement (Wasserstein and Lazar 2016), which enumerates truisms about the p-value, but does not provide guidance regarding best uses of the p-value, and provides nuance to the simple stringent threshold suggested by Benjamin et al. (2018). In summary, I have elucidated the importance of contextualizing the p-value within the salient features of the study when formal hypothesis testing is undertaken. When this is done, it can be a useful measure of evidence for the truth.

Funding

This work was supported in part by NIH Grant UL1TR001102.

References

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., and Cesarini, D. (2018), "Redefine Statistical Significance," *Nature Human Behaviour*, 2, 6. [115,117]

Browne, R. H. (2010), "The t-Test p Value and Its Relationship to the Effect Size and $P(X > Y)$," *The American Statistician*, 64, 30–33. [116]

Connolly, S. J., Ezekowitz, M. D., Yusuf, S., Eikelboom, J., Oldgren, J., Parekh, A., Pogue, J., Reilly, P. A., Themeles, E., Varrone, J., and Wang, S. (2009) "Dabigatran Versus Warfarin in Patients With Atrial Fibrillation," *New England Journal of Medicine*, 361, 1139–1151. [Erratum, N Engl J Med 2010;363:1877.] [117]

Ionnidis, J. P. A. (2005), "Why Most Published Research Findings Are False," *PLoS Medicine*, 2, e124. [115]

Keefe, R. S. E., Kraemer, H. C., Epstein, R. S., Frank, E., Haynes, G., Laughren, T. P., and Leon, A. C. (2013), "Defining a Clinically Meaningful Effect for the Design and Interpretation of Randomized Controlled Trials," *Innovations in Clinical Neuroscience*, 10, 4S–19S. [116]

Lesaffre, E. (2008), "Superiority, Equivalence, and Non-inferiority Trials," *Bulletin of the NYU Hospital for Joint Diseases*, 66, 150–154. [116]

Rosnow, R. L., and Rosenthal, R. (2003), "Effect Sizes for Experimenting Psychologists," *Canadian Journal of Experimental Psychology*, 57, 221–237. [116]

Wasserstein, R. L., and Lazar, N. A. (2016), "The ASA's Statement on p-Values: Context, Process and Purpose," *The American Statistician*, 70, 129–133. [115,117]