HHS Public Access

Author manuscript

Multivariate Behav Res. Author manuscript; available in PMC 2017 January 01.

Published in final edited form as:

Multivariate Behav Res. 2016 ; 51(1): 23–29. doi:10.1080/00273171.2015.1099032.

A Test By Any Other Name: *P*-values, Bayes Factors and Statistical Inference

Hal S. Stern*

UC Irvine

Abstract

The exchange between Hoitjink, van Kooten and Hulsker (in press) (HKH) and Morey, Wagenmakers, and Rouder (in press) (MWR) in this issue is focused on the use of Bayes factors for statistical inference but raises a number of more general questions about Bayesian and frequentist approaches to inference. This note addresses recent negative attention directed at *p*-values, the relationship of confidence intervals and tests, and the role of Bayesian inference and Bayes factors, with an eye towards better understanding these different strategies for statistical inference. We argue that researchers and data analysts too often resort to binary decisions (e.g., whether to reject or accept the null hypothesis) in settings where this may not be required.

Keywords

Bayesian inference; confidence intervals; effect size; significance testing

This issue of *Multivariate Behavioral Research* includes an exchange between Hoitjink, van Kooten and Hulsker (in press) (HKH) and Morey, Wagenmakers, and Rouder (in press) (MWR) on the appropriate use of Bayes factors for statistical inference. Their exchange focuses on whether to use default or subjective prior distributions, the proper calibration of default prior distributions, and the proper interpretation of Bayes factors. The exchange, along with the recent controversy surrounding the decision of *Basic and Applied Social Psychology* (*BASP*) (Tramifow and Marks, 2015) to ban *p*-values, raises a number of general questions about how to perform statistical inference. This article considers both standard and Bayesian approaches to statistical inference. It begins with a discussion of common criticisms of null hypothesis significance testing and *p*-values, considers the relationship of testing and confidence intervals, addresses the role of Bayesian inference and Bayes factors, and then concludes with some practical advice about statistical inference.

BASP and the banning of p-values

For purposes of discussion we focus throughout on the setting used by HKH and MWR in which a sample of size n is observed with Y_i, $i = 1, \ldots, n$ independent and identically distributed as $N(\mu, \sigma^2)$ random variables. Their contributions use an alternative parameterization (with the standardized effect size $\delta = \mu/\sigma$ being the parameter of interest)

*Address: Hal Stern is a Professor in the Department of Statistics, University of California, Irvine, Irvine, CA 92697, sternh@uci.edu.

The author is grateful to the editor for comments that improved the presentation of material.

but this does not impact any of the discussion provided here. Attention is focused on the parameter μ which measures the scientific effect of interest (e.g., the excess (relative to chance) hit rate in the study of psi by Bem, 2011). Following HKH and MWR we initially focus on testing the null hypothesis that μ = 0. It is argued below though that the focus on significance testing (or later, on Bayes factors) rather than on estimation of effect sizes is problematic. The *t*-test of significance is commonly used to assess the null hypothesis H_o : μ = 0. To perform the *t*-test one first computes the *t*-statistic, $t=\overline{Y}/(s/\sqrt{n})$, and then inference regarding H_o is based on the *p*-value which measures the probability that a study like this would yield a t-statistic as or more extreme than the observed statistic if the null hypothsis were true. Small values of the *p*-value suggest the observed data are unlikely under the null hypothesis and thus may call that hypothesis into question. It is common to compare the *p*-value to some standard thresholds, e.g., α = .01 or .05 or .10, and then note that the observed difference is significantly different from zero (often just termed ”significant”) at the specified threshold if the *p*-value is smaller than the cutoff. We return to some negative consequences of this common practice below.

The *p*-value is a probability calculation giving the probability of an event (observing a more extreme *t*-statistic) under specific assumptions: the statistical model is correct and H_o is true. Probabilitiy calculations do not seem particularly objectionable. Why then would *BASP* ban *p*-values? Or more precisely why did the journal decide that ”… prior to publication, authors will have to remove all vestiges of the NHSTP (*p*-values, *t*-values, *F*-values, statements about “significant” differences or lack thereof, and so on).” (Tramifow and Marks, 2015, p. 1) where NHSTP in the quotation stands for null hypothesis significance testing procedure. It is true that *p*-values are often misinterpreted and abused (this is discussed further below) but that by itself does not seem like a compelling reason to ban them. The motivation for the ban is a concern with the logic that underlies significance testing and *p*-values. The question of interest in the testing framework concerns the relative likelihood of the null and alternative hypotheses given the experimental data. The difficulty is that this question can not be addressed by a calculation (the *p*-value) that assumes the null hypothesis is true. Of course, this concern is certainly not new. There have been many other critiques of this aspect of significance testing dating back at least as far as Berkson (1942), including many in psychology (Rozeboom, 1960; Cohen, 1994). Indeed, this is not the first attempt to ban *p*-values. Chapter 1 of Kline (2013) includes a history of the controversy surrounding null hypothesis testing and several attempts to reform or ban that practice.

There are, in fact, a number of very good reasons to be concerned about the use of *p*-values. The logical difficuly described above is one important point. A significance test, and more precisely a *p*-value, can not say anything about the relative merits of two hypotheses (the null and alternative) when it is calculated assuming that one of the hypotheses is true. Another important problem is that misinterpretations and misunderstandings of the *p*-value are common. Many people continue to interpret the *p*-value as speaking to the likelihood of the null hypothesis which is impossible given its definition. A third concern is the heavy dependence of the *p*-value on the sample size. A study focused on a phenomenon characterized by a small effect size can yield low *p*-values (significant results) in large samples and on the other hand a large and potentially important effect may not be found

significant in a small sample. Another concern is that some investigators carry out a number of tests (e.g., using a range of outcome measures or a range of different statistical models) and then report a *p*-value without providing the context in which it was obtained (Simmons, Nelson, & Simonsohn, 2011).

Perhaps the biggest problem with *p*-values though is that the most common way in which they are used, a *p*-value is calculated and then compared to a threshold to determine significance, promotes a "binary" view of statistical inference that is not generally helpful. Others (e.g., Goodman, 1999) have written about the logical difficulties inherent in this common approach which combines significance testing (where the *p*-value is a measure of evidence regarding the plausibility of the null hypothesis) and the Neyman-Pearson view which chooses a significance level to control error rates as part of its hypothesis testing framework. I am especially concerned about the practical consequences of the combined approach. For many investigators results are either declared significant, in which case the researchers are likely to claim the effect is real and important, or the results are not significant, in which case researchers are likely to declare the null hypothesis must be true. It is this extreme "binary" view that in my opinion has done the most damage to science. It is problematic in many ways. A small *p*-value does not necessarily mean that an effect is practically important. It may merely reflect good fortune, particularly in an underpowered study. A large (non-significant) *p*-value implies that the data could easily have been observed under the null hypothesis. But of course the data could also have been observed under a range of alternative values of the parameter we are testing. Even top scientists in many fields miss this point and tend to dismiss findings that do not attain a desired level of significance. This observation stands behind the proposal to report a "counternull", the non-null effect size that would lead to the same *p*-value as the null hypothesis, in addition to the *p*-value (Rosenthal & Rubin, 1994). Though this proposal is not implemented often, practitioners would be wise to remember its message that both the null hypothesis and the alternative hypothesis can produce data that are "not significant".

Confidence intervals

The journal *BASP* also has banned the use of confidence intervals. This decision is not too surprising given that there is a certain equivalence of testing and confidence interval procedures. A significance test of a hypothesized value for μ will produce a *p*-value less than a given threshold (say α) when a $100(1 - \alpha)\%$ confidence interval for μ excludes the hypothesized value. Thus a series of significance tests computing *p*-values for various hypothesized values of the parameter in question can provide the same information as a confidence interval and a single confidence interval tells us which values of the parameter would be found plausible in a series of significance tests. The reason given by the *BASP* editors for extending the ban to confidence intervals is similar to the argument for the significance testing ban. The single confidence interval being computed does not provide the kind of probabilistic guarantee that the editors of *BASP* believe is required for an appropriate inference. The frequentist argument that supports the confidence interval procedure guarantees that the specified proportion of the $1–\alpha$ confidence intervals that we create will contain the true parameter value, but it is not possible to make a probabilistic claim for the one interval at hand. It is noteworthy that many other critics of signifcance testing and *p*-

values actually encourage the use of confidence intervals (Wilkinson and the Task Force on Statistical Inference, 1999; Kline, 2013). I too believe that confidence intervals are valuable for a number of reasons. For one thing, in the simple setting considered here the *t* confidence interval corresponds to a Bayesian posterior interval for a diffuse (sometimes called non-informative) choice of the prior distribution of the model parameters which means that the desired probabilistic interpretation is realistic in that case. In large samples the *t*-interval is approximately the same as a Bayesian posterior interval regardless of the prior distribution which again justifies the probabilistic interpretation. Most important for me though is the simple fact that confidence intervals provide a range of values of the population parameter that are compatible with the data. The range informs us about the magnitude of sampling variability (essentially providing information about sample size that is lacking in *p*-values) and it also encourages a focus on plausible values of the effect rather than focusing on proving/disproving a single hypothesis about the population parameter. Gelman and Stern (2006) and Cummings (2011) discuss several examples where valuable information is obtained from confidence intervals in situations where testing may produce difficult to interpret results.

Bayesian inference

Many researchers, including the authors of the two pieces that motivated this article, believe that the Bayesian approach to statistical inference is the most natural way to analyze data. The Bayesian approach avoids many of the concerns that have been raised here about significance tests. We briefly describe the approach here; additional information can be obtained from any number of texts including Carlin and Louis (2008), Christensen, Johnson, Branscum, and Hanson (2010), and Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2013). The Bayesian approach to inference is characterized by the explicit use of probability distributions to draw inferences. In common with the standard frequentist approach to inference, there is a probability model (sometimes known as the data model or the sampling model) for observable quantities given underlying population parameters, often denoted $p(y|\theta)$. This is a Gaussian distribution for the Bem study. The Bayesian approach adds a prior probability distribution describing uncertainty about the underlying parameters of the sampling or data distribution, denoted by $p(\theta)$. Given these two distributions Bayes theorem provides the mathematical machinery needed to combine the information in these two distributions to obtain the posterior distribution of θ, $p(\theta|y) = p(y|\theta)p(\theta)/\int p(y|\theta)p(\theta)d\theta$. The posterior distribution is a summary of what the data and prior information tell us about which values of θ are most plausible. Bayesian methods have been a part of the statistical methodology literature for many years but have increased dramatically in popularity over the last 30 years due to advances in computational algorithms and computational devices that have made it practical to study the posterior distribution via simulation (e.g., using Markov chain Monte Carlo methods) in situations for which an analytical solution is not practical. Computation is discussed in some detail in each of the references mentioned above.

Advocates of the Bayesian approach argue that it provides a natural framework for integrating a variety of sources of information about a quantity of interest. In its most basic form it combines information from a sample of data with a priori available information

about a population parameter value. More generally though the Bayesian framework also makes it easy to incorporate information about relationships among a set of parameters. For example, we might assume that the effects of individual classrooms in an education study can be modeled as draws from a population of possible classroom effects. In the Bayesian approach this leads to hierarchical models. The population distribution in the hierarchical model is sometimes known as a random effects distribution in other approaches to inference. In addition, the Bayesian approach proves to be a natural framework for accommodating complications like unintentional missing data, order constraints on parameters, data that are censored or rounded off, etc. This is not to say that the Bayesian approach is uniquely able to handle such complications, only that in practice it seems reasonably straightforward to incorporate such issues into a comprehensive probability model. A final benefit of the Bayesian approach is that the posterior distribution enables a wide range of inferential statements to be made. One can summarize the posterior distribution by providing a summary measure (e.g., the posterior mean) for each parameter, or a posterior interval describing the range of plausible values for each parameter of interest. The posterior distribution also allows other questions of interest to be addressed (e.g., what is the probability that the parameter of interest is positive or what is the probability that one parameter is greater than a second parameter).

The previous paragraph presents some of the benefits of the Bayesian approach to inference. The primary source of controversy surrounding Bayesian methods for some researchers is a concern about how the prior distribution is specified. The discussions of HKH and MWR hit upon one of the main questions, whether to use problem-specific domain knowledge to specify a probability distribution that describes the user's subjective a priori opinion regarding the unknown parameter or whether instead to rely on some sort of default (occasionally called objective) prior distribution. There is often at least some subject matter information that can be incorporated into a prior distribution. But many users are concerned that a wide range of subjective prior distributions might produce disparate or conflicting results and are drawn to the use of default prior distributions in the hope that they can produce a form of consensus analysis. Of course, in large samples the choice of prior distribution is much less important. In practice there is much more to to say about prior distributions; interested readers can refer to the Bayesian texts listed above.

Bayes factors

The discussion of Bayesian inference given above notes that there is considerable flexibility in how one summarizes the posterior distribution to provide desired inferences. This extends to allowing a probabilistic evaluation of the relative support for competing hypotheses if one adopts a testing framework. As described in the HKH article, the Bayes factor BF_{01} is the Bayesian approach used to compare two hypotheses H_o and H_1 (e.g., $H_o : \mu = 0$ and $H_1 : \mu \neq 0$). I must confess up front that I do not generally use Bayes factors in my applied work. I believe the posterior distribution provides the most relevant inferences for the population mean in a Bayesian analysis and generally report posterior intervals for the parameters of interest (e.g. for μ) as the summary. Despite this personal view, given that the Bayes factor is at the core of the HKH and MWR exchange I next briefly review the definition of the Bayes factor and provide my views on the questions considered by HKH and MWR.

The Bayes factor measures the ratio of the marginal likelihood of the data under the null hypothesis, denoted $p(y|H_o)$ and computed as $p(y|H_o) = \int p(y|\theta, H_o)p(\theta|H_o)d\theta$ where θ are the unknown parameters in the model H_o, and the marginal likelihood of the data under the hypothesis H_1, denoted $p(y|H_1)$ and computed in analogous fashion. Formally the Bayes factor tells a user how to modify the a priori odds in favor of H_o relative to H_1 in order to obtain the posterior odds in favor of H_o. If the BF_{01} is equal to 1 then the data provide equal support for the two hypotheses and there is no reason to change our a priori opinion about the relative likelihood of the two hypotheses, if BF_{01} is greater than 1 then the data provide support for the null hypothesis and we should increase the odds in favor of H_o, and if BF_{01} is less than 1 then the data provide support for the alternative hypothesis and we should decrease the odds in favor of H_o. Of course, to fully utilize the Bayes factor and turn it into a posterior probability that the null hypothesis is true, one must first specify a prior probability on that proposition. Without a subjective prior distribution on the two competing hypotheses, the Bayes factor becomes a measure of evidence regarding the two hypotheses but does not tell us exactly what conclusion to draw. The Bayes factor has a significant advantage over the *p*-value in that it explicitly addresses the likelihood of the observed data under each hypothesis and thus treats the two symmetrically. (Recall the *p*-value assumed that H_o was true.)

HKH (in press) address three issues that they see arising in applications of the Bayes factor in psychology, using the one-sample example described above as motivation. In that setting a requirement for forming the Bayes factor is a prior distribution for the parameter μ (or the standardized effect size μ/σ) under the alternative hypothesis that it is not zero. HKH and MWRfocus their attention on a prior distribution that is Gaussian with mean zero and for which the only unspecified parameter is the standard deviation (their τ). The first issue raised by HKH concerns the frequent use of default prior distributions rather than subjective prior distributions, the second issue concerns how to choose the default prior distribution (or more precisely the key parameter(s) of the default prior distribution) if one opts to go that route, and the third issue concerns interpretation of the Bayes factor. HKH and MWR both seem to agree on the issues associated with the choice of subjective or default prior distribution, although they disagree on how best to proceed in the specific example at hand. Thus, I do not discuss this issue further here.

For the second and third issues HKH propose to use frequency properties of the Bayesian method to assist in calibrating the default prior distribution and in selecting thresholds for interpreting the Bayes factor. Their approach is discussed further below but it is worth noting that this is not the first time frequency calculations have been proposed for use by Bayesian analysts. Rubin (1984) argued that there are frequency calculations that can be justified as being relevant to Bayesian analysts to assure that their inferences are calibrated correctly or to assist with evaluating the assumptions of their model. Similarly, Little (2006) argued for Bayesian inference but endorsed careful assessment of models using frequentist ideas. In a medical context Berry (2004) showed how Bayesian methods can be used to design clinical trials while controlling standard frequentist operating characteristics if that is desired.

To address HKH's second issue, we suppose that the default normal prior distribution is chosen and then ask how to choose the standard deviation τ. HKH propose a form of frequentist calibration; the standard deviation should be chosen so that the Bayes factor has specified frequentist properties. They give a couple of example calibration rules, one (their definition 2) being to choose the standard deviation such that the probability the Bayes factor is greater than one (favors the null) is high (say .95) if the null hypothesis is true. I share MWR's concern about this type of calibration. In automating the application of the Bayes factor in this way HKH are reproducing some of the problematic issues associated with traditional hypothesis testing.

When it comes to interpreting the Bayes factor after one has chosen the appropriate prior distribution, HKH are concerned about the various scales that have been proposed (see, e.g., Kass and Raftery, 1995). It is easy to understand their concern about automatic use of Bayes factors cutoffs. After all the use of .05 as a strict cutoff in significance testing was identified earlier as a key problem with significance testing. HKH propose the use of frequentist operating characteristics of the Bayes factor procedure to assist with interpretation. They generate simulated Bayes factors under the assumption that the null hypothesis is true and then again under a specific alternative hypothesis (or a series of different alternative hypotheses). Given an observed Bayes factor a researcher can use these simulations to determine the likelihood of observing such a value of the Bayes factor under the null and alternative cases. Though correct, the HKH approach seems like a great deal of work to help refine the interpretation of an observed Bayes factor, work that I do not believe that I would find useful in my applied work. The Bayes factor is a measure of evidence and comes with a natural interpretation – a Bayes factor (BF_{01}) of 1/1000 for example multiplies ones prior odds in support of the null hypothesis by 1/1000 and thus indicates very strong evidence in favor of the alternative hypothesis for all but the most committed supporters of H_o. The ambiguity of larger values less than one for the Bayes factor (i.e., 1/5, 1/10, 1/20) is informative as it indicates the data supports the alternative but that one's conclusion will depend on one's prior opinion about the relative likelihood of the two hypotheses being considered. I believe that this is consistent with what Kass and Raftery (1995) describe. I do not believe they or others suggest the use of strict cutoffs for decision making. Here HKH's desire to force the Bayes factor into a decision regarding the "correct" hypothesis is creating a binary decision where it may not be required. It is the adoption of this binary view of the world instead of choosing to focus on the magnitude of effects that leads to the HKH concerns. Summarizing Bayesian inferences by a posterior distribution rather than insisting on a binary decision would elimiinate the need for the proposed fix.

Practical advice

The HKH/MWR discussion in this issue of the journal is important. Though the details of the exchange are likely to be primarily of interest to Bayesian data analysts, the exchange in fact raises more substantial issues about statistical inference. My own view is that psychologists too often ask questions in the form of hypothesis tests when the questions might be more usefully addressed with effect sizes and interval estimates. There can be no doubt that there are scientific questions that require a decision as to whether to accept a particular hypothesis – clearly this is the case in a medical study designed to decide whether

to continue applying an existing treatment or adopt a new one. I do not however believe that hypothesis tests are always or even generally required in psychology. Understanding the size of the effect that has been observed and the role of sampling variability are the basic elements of a data analysis. In my own research I tend to present such results by summarizing the posterior distribution of a Bayesian analysis. Others may prefer to present confidence intervals based on a frequentist analysis. In either case, the binary decision regarding a specific null hypothesis, if needed, should occur only after we understand what the data are telling us about the effect under study. It is important to emphasize that this is not a new message; it has been made offen in the psychology literature over the years (e.g., Wilkinson et al, 1999; Cummings, 2011).

Earlier in this article I questioned the decision of *BASP* to ban the *p*-value and all other traces of significance testing. I continue to believe that this is a misguided and extreme reaction to the fact that people occasionally misinterpret *p*-values as being more than just a measure of evidence about H_O. In place of significance testing, the editors of *BASP* suggest greater use of descriptive statistics and effect sizes, more graphical displays of the distribution of the data, and larger sample sizes (Tramifow and Marks, 2015). These recommendations echo the suggestions of earlier groups, e.g., the Task Force on Statistical Inference (Wilkinson et al., 1999). The first two of these are outstanding suggestions and should be a key part of scientific data analyses. Indeed, I would especially emphasize the second suggestion. Figure 1 presents two data displays comparing hypothetical data for treatment and control groups in a randomized study. The left hand panel presents the data in a form that is all too recognizable to readers of many popular and highly regarded journals. For each group the height of the bar represents the mean and the vertical bar extending up from the top of the bar gives the standard deviation of the distribution (or occasionally the standard error of the mean). The right hand panel shows the mean as a horizontal line superimposed on the observed data values. It is natural to ask how we ended up with so many people choosing to display their data in the form presented at the left rather than the form presented at the right? The proposal that investigators should use bigger samples is an easy one to support but often difficult to execute because of funding. For this reason I believe it is important that any publication provide information about the amount of sampling variabiality that is present. Confidence intervals and Bayesian posterior intervals are excellent ways to communicate this information.

The articles in this issue focus attention on the critical role of statistical inference (*p*-values, Bayes factors, confidence intervals, posterior distributions) in scientific analyses and the importance of avoiding errors in the application of such tools. Before ending this article however it is worth repeating a critical point recently made by Leek and Peng (2015). Scientific studies involve much more than the statistical analysis stage. There are numerous other points in the research process as well. We should also be scrutinizing experimental design, data collection, data editing, preliminary data analyses, and the choice of statistical models (see, e.g., Funder, Levine, Mackie, Morf, Sansome, & West, 2014; Wilkinson et al., 1999). At each step in the research process errors can be made that negatively impact the scientific enterprise. Researchers should be careful to avoid such errors and to use the best available methods for understanding the magnitude of the effects under study.

References

Bem DJ. Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. Journal of Personality and Social Psychology. 2011; 100:407–425. [PubMed: 21280961]

Berkson J. Tests of significance considered as evidence. Journal of the American Statistical Association. 1942; 37:325–335.

Berry DA. Bayesian statistics and the efficiency and ethics of clinical trials. Statistical Science. 2004; 19:175–187.

Carlin, BP.; Louis, TA. Bayesian Methods for Data Analysis. 3rd. Boca Raton, FL: Chapman and Hall / CRC; 2008.

Christensen, R.; Johnson, W.; Branscum, A.; Hanson, TE. Bayesian Ideas and Data Analysis. Boca Raton, FL: Chapman and Hall / CRC; 2010.

Cohen J. The earth is round ($p < .05$). American Psychologist. 1994; 49:997–1003.

Cumming, G. Understanding the New Statistics: Effect Sizes, Confidence Intervals and Meta-Analysis. New York, NY: Routledge; 2011.

Funder DC, Levine JM, Mackie DM, Morf CC, Sansone C, Vazier S, West SG. Improving the dependability of resaerch in personality and social psychology: Recommendations for research and educational practice. Personality and Social Psychology Review. 2014; 18:3–12. [PubMed: 24214149]

Gelman, A.; Carlin, JB.; Stern, HS.; Dunson, DB.; Vehtari, A.; Rubin, DB. Bayesian Data Analysis. 3rd. Boca Raton, FL: Chapman and Hall / CRC; 2013.

Gelman A, Stern H. The difference between “significant” and “not significant” is not itself statistically significant. The American Statistician. 2006; 60:328–331.

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Annals of Internal Medicine. 1999; 130:995–1004. [PubMed: 10383371]

Hoitjink H, van Kooten P, Hulsker K. Why Bayesian psychologists should change the way they use the Bayes factor. Multivariate Behavioral Research. (in press).

Hoitjink H, van Kooten P, Hulsker K. Bayes factors have frequency properties, this should not be ignored: a rejoinder to Morey, Wagenmakers, and Rouder. Multivariate Behavioral Research. (in press).

Kass RE, Raftery AE. Bayes factors. Journal of the American Statistical Association. 1995; 90:773–795.

Kline, RB. Beyond Significance Testing: Statistics Reform in the Behavioral Sciences. 2nd. Washington, DC: American Psychological Association; 2013.

Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. Nature. 2015; 520:612. [PubMed: 25925460]

Little RJ. Calibrated Bayes: A Bayes/frequentist roadmap. The American Statistician. 2006; 60:213–223.

Morey RD, Wagenmakers E-J, Rouder JN. Calibrated Bayes factors should not be used: a reply to Hoitjink, van Kooten, and Hulsker. Multivariate Behavioral Research. (in press).

Rosenthal R, Rubin DB. The counternull value of an effect size: A new statistic. Psychological Science. 1994; 5:329–334.

Rozeboom WW. The fallacy of the null hypothesis significance test. Psychological Bulletin. 1960; 57:416–428. [PubMed: 13744252]

Rubin DB. Bayesianly justfiable and relevant frequency calculations for the applied statistician. Annals of Statistics. 1984; 12:1151–1172.

Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011; 22:1359–1366. [PubMed: 22006061]

Tramifow D, Marks M. Editorial. Basic and Applied Social Psychology. 2015; 37:1–2.

Wilkinson L. the Task Force on Statistical Inference. Statistical methods in psychology: Guidelines and explanation. American Psychologist. 1999; 54:594–604.

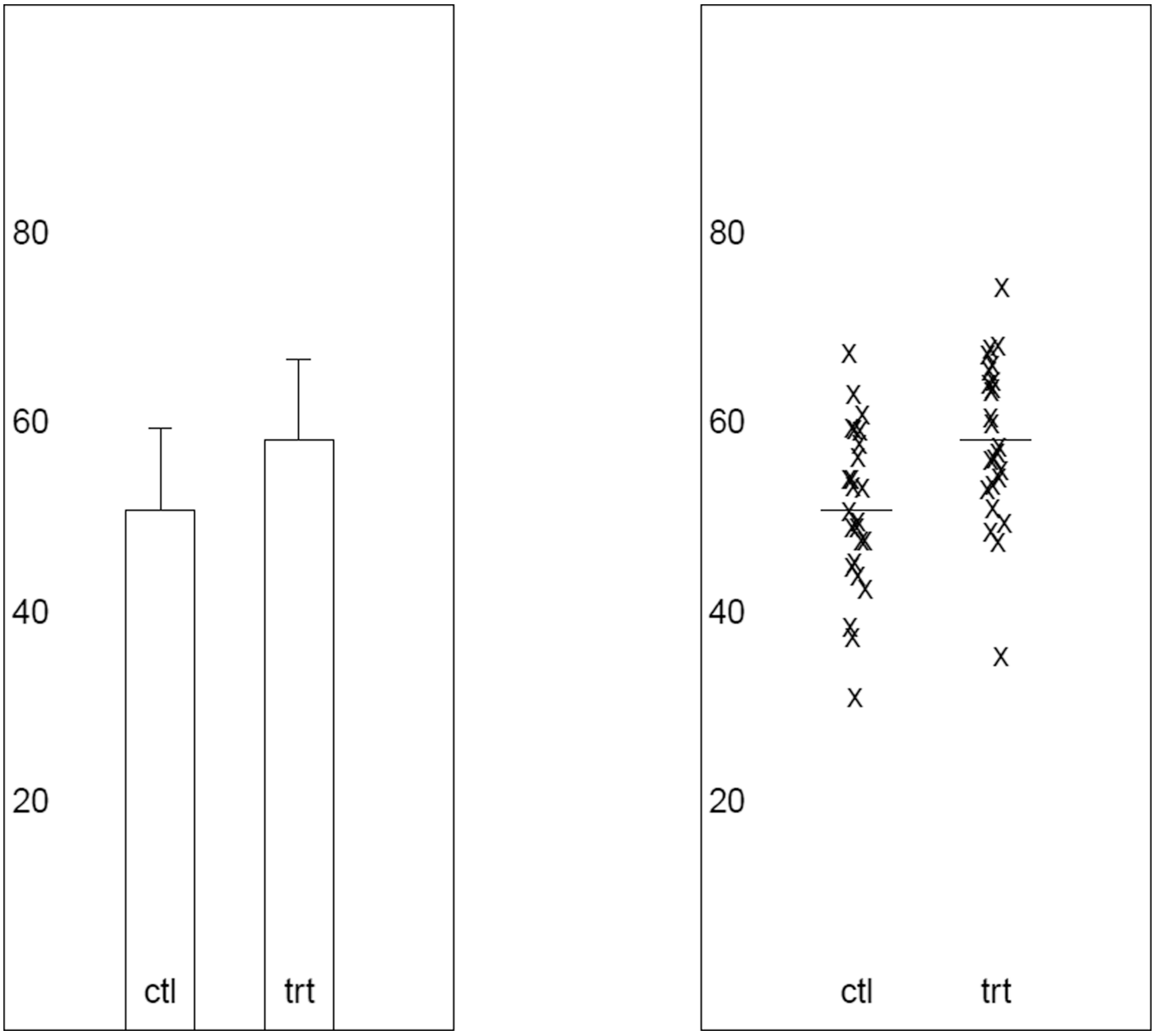

Figure 1.

Data display for a hypothetical treatment (trt) and control (ctl) group comparison. The figure on the left shows a vertical bar for each group with the height of the bar equal to the sample mean and with a vertical bar extending one standard deviation above the mean. The figure on the right shows data points in each group with horizontal line indicating position of the sample mean. Points have been jittered horizontally to make it easier to see multiple similar observations.