Hypothesis Testing, p Values, Confidence Intervals, Measures of Effect Size, and Bayesian Methods in Light of Modern Robust Techniques

Educational and Psychological
Measurement

2017, Vol. 77(4) 673–689

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416667983

journals.sagepub.com/home/epm

Rand R. Wilcox¹ and Sarfaraz Serang¹

Abstract

The article provides perspectives on p values, null hypothesis testing, and alternative techniques in light of modern robust statistical methods. Null hypothesis testing and p values can provide useful information provided they are interpreted in a sound manner, which includes taking into account insights and advances that have occurred during the past 50 years. There are, of course, limitations to what null hypothesis testing and p values reveal about data. But modern advances make it clear that there are serious limitations and concerns associated with conventional confidence intervals, standard Bayesian methods, and commonly used measures of effect size. Many of these concerns can be addressed using modern robust methods.

Keywords

robust methods, Bayesian methods, skewed distributions, outliers, heteroscedasticity, p values, effect size, Tukey's three decision rule

Introduction

As is well known, a perennial issue is what role null hypothesis testing (NHT) methods should play in psychological research. A comprehensive summary of the various concerns and issues is provided by Nickerson (2000) and Kline (2013). Cumming (2014), for example, argues against ever using a p value and relying instead on

¹University of Southern California, Los Angeles, CA, USA

Corresponding Author:

Rand R. Wilcox, Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061, USA.

Email: rwilcox@usc.edu

confidence intervals and measures of effect size. His criticism of p values is based in part what he claims is a common point of view: If p reveals truth, and we replicate the experiment—doing everything the same except with a new random sample—then replication p , the p value in the second experiment, should presumably reveal the same truth. As Cumming points out, this interpretation of a p value is invalid. That is, a specific misinterpretation of p values is that they reflect the likelihood of replicating a study, which clearly is not the case. Branch (2014) also criticizes p values based on this misinterpretation. A seminal paper by Cohen (1994) also criticized p values because of

its near-universal misinterpretation of p as the probability that H_0 is false, the misinterpretation that its complement is the probability of successful replication, and the mistaken assumption that if one rejects H_0 one thereby affirms the theory that led to the test. (p. 997)

Cohen goes on to encourage the increased use of exploratory studies, graphical methods, confidence intervals, and measures of effect size. As seems evident, these misinterpretations are due, at least in part, to poor training. But does this mean that p values in particular and NHT in general should be abandoned completely? If, for example, the focus is on confidence intervals and measures of effect size, do practical problems persist in our quest to understand data? Do p values provide any useful information? If the answer to this last question is yes, then abandoning p values is tantamount to suppressing information that is useful to researchers who understand what p values reveal about data.

A basic goal in this article is to suggest that p values play a useful role when viewed in the context of Tukey's three decision rule (described in the second section) in conjunction with modern insights regarding the robustness of the many methods that are now available for analyzing data. As will be evident, Tukey's three decision rule does not eliminate the possible misinterpretation of p values due to poor training, and clearly p values by themselves do not provide a deep understanding of data. The only point is that Tukey's perspective provides at least some useful information about how groups compare and the association among two or more variables. Another goal is to briefly review some basic issues regarding the robustness of confidence intervals and measures of effect size. The main point is that certain well-known techniques can be highly misleading, but there are various strategies for dealing with known concerns. The final goal is to comment on Bayesian methods from a robustness point of view.

The article is organized as follows. The second section reviews Tukey's three decision rule. The third section provides a brief review of basic robust methods with the goal of describing when a p value might be deemed useful in the context of Tukey's three decision rule. The fourth section provides perspectives on confidence intervals, some of which stem from results reviewed in the third section. The fifth section discusses measures of effect size in light of the results reviewed in the third and fourth sections. The sixth section discusses Bayesian methods from the perspective of fundamental robustness concepts that were developed about 50 years ago, the

point being that the issue of robustness, when adopting a Bayesian point of view, has not been adequately addressed.

Tukey's Three Decision Rule

Consider the common goal of testing the hypothesis that some measure of location, say θ , is equal to some specified value, θ_0 . That is, the goal is to test

$$H_0 : \theta = \theta_0. \quad (1)$$

And of course another common goal is testing

$$H_0 : \theta_1 = \theta_2, \quad (2)$$

where θ_1 and θ_2 are measures of location associated with two independent groups.

First focus on (1). A criticism of testing this hypothesis is that surely θ differs from θ_0 at some decimal place (Tukey, 1991). Jones and Tukey (2000) suggest dealing with this issue using what is generally known as Tukey's three decision rule. Let $\hat{\theta}$ be some estimate of θ . If (1) is rejected and $\hat{\theta} < \theta_0$, decide that $\theta < \theta_0$. If (1) is rejected and $\hat{\theta} > \theta_0$, decide that $\theta > \theta_0$. If (1) is not rejected, make no decision. So the goal is not to test (1), but determine whether it is reasonable to make a decision about whether θ is greater or less than θ_0 .

As for testing (2), a criticism is that surely these two measures of location differ at some decimal place. (An exception might occur when θ is taken to be the population median and two discrete distributions are being compared that have a relatively small sample space.) If we accept this argument, then Tukey's three decision rule can be used. That is, the goal is to determine whether θ_1 is less than or greater than θ_2 and a decision is made if (2) is rejected. Even if Tukey's view is disputed, certainly in some instances there is interest in determining which group has the larger measure of location.

Note that from the point of view of Tukey's three decision rule, a p value quantifies the extent to which it is reasonable to make a decision. The closer the p value is to zero, the stronger the empirical evidence that it is reasonable to make a decision: provided the Type I error probability can be controlled reasonably well. From basic principles, a confidence interval, for example, can be used to make a decision about whether to reject the hypotheses given by (1) and (2), and it has the advantage of providing information about the extent the null hypothesis is false. But a p value provides more information about the extent it is reasonable to make a decision about the parameters of interest. In the context of Tukey's three decision rule, power corresponds to the probability of being able to make a decision about the parameter of interest.

Cohen (1994) argues that hypothesis testing does not answer a question of central interest: What is the probability that a null hypothesis is true. He goes on to note that a Bayesian approach provides a method for dealing with this issue. When testing for

exact equality, Tukey is essentially saying that, in general, we know this probability without any data: zero.

However, interpreting p values as just described is based on the assumption that the null distribution of the test statistic being used can be determined to a reasonably accurate degree. It is well known in the statistics literature that methods based on means or least squares regression are not robust in a broad sense briefly reviewed in the third section, and in particular they can be unsatisfactory in terms of providing accurate confidence intervals and relatively high power (e.g., Hampel, Ronchetti, Rousseeuw, & Stahel, 1986; Huber & Ronchetti, 2009; Maronna, Martin, & Yohai, 2006; Rousseeuw & Leroy, 1987; Staudte & Sheather, 1990; Wilcox, 2012a, 2012b, 2016). Simultaneously, the books just cited describe a vast array of techniques where the null distribution can be approximated reasonably well. For these methods, a p value can be interpreted as reflecting the strength of the empirical evidence that some unknown parameter is greater than some specified value. And when comparing two parameters, a p value reflects the strength of the empirical evidence that a decision can be made about which of the two parameters has the larger value.

Obviously, this interpretation of p values does not tell us everything we would like to know about the parameters under investigation. For example, is the difference between θ_1 and θ_2 clinically important? And if a study is replicated, a p value does not reflect the probability of being able to again make a decision about which group has the largest measure of location. Yet another limitation is that a p value does not say anything about the probability that θ_1 is greater than θ_2 . The only point is that completely abandoning p values means suppressing information that can be useful to those researchers who know how to interpret a p value in a proper way.

A Brief Review of the Basics of Robust Statistics

When describing some statistical method, what does the term robust mean? Perhaps the most obvious response, when computing a confidence interval, is that the actual probability coverage should be reasonably close to the nominal level. From the point of view of testing hypotheses, there is also the issue of achieving relatively high power. But from a modern perspective, the term robust has a much broader meaning that is relevant to basic methods for summarizing data. Roughly, population parameters, such as the mean, variance, and median, are said to be robust if arbitrarily small changes in a distribution do not substantially affect their values. From this point of view, the population mean and variance are not robust (e.g., Hampel et al., 1986; Huber & Ronchetti, 2009). This is not to suggest that they play no useful role. But it is well known that this lack of robustness can yield a highly misleading summary of data (e.g., Heritier, Cantoni, Copt, & Victoria-Feser, 2009; Maronna et al., 2006; Rousseeuw & Leroy, 1987; Staudte & Sheather, 1990; Wilcox, 2012a, 2012b, 2016). That is, even if the population mean and variance could be determined exactly, they can be highly misleading. The same is true when dealing with the slope and intercept of a regression line that is estimated via least squares.

In the context of some random sample, robustness means that a few unusual values should not dominate the statistic that is being used. The sample mean, for example, does not satisfy this goal: Even a single outlier can substantially alter the sample mean. Moreover, for heavy-tailed distributions, roughly meaning distributions where outliers are likely to occur, the standard error of the sample mean can be substantially larger than the standard error of robust estimators such as the median or a 20% trimmed mean, which has implications about the ability to replicate a result. (For illustrations, see Wilcox, 2016.) Also, even a single outlier can inflate the sample variance resulting in relatively long confidence intervals and misleading measures of effect size that are based in part on the sample variance. Using least squares regression can result in a poor fit to the bulk of data and relatively low power. Even when dealing with skewed distributions where outliers are relatively rare, inferential methods based on the mean and variance can be relatively inaccurate (e.g., Wilcox, 2012a, 2012b).

To some degree, many non-statisticians are aware of the practical concerns associated with skewed distributions, outliers, and heteroscedasticity. Some use the more effective methods for dealing with these issues, but it is evident that most do not. A practical consequence is that misleading generalizations can be made and discoveries lost.

Consider, for example, the strategy of transforming data in an attempt to deal with skewed distributions. Typically, distributions remain skewed after using the more obvious transformations (e.g., taking logs or using a Box–Cox transformation) and outliers remain (e.g., Wilcox, 2012b). There are vastly better techniques for dealing with skewed distributions (e.g., Wilcox, 2012a, 2012b, 2016), which include methods based on robust measures of location, improvements on the Wilcoxon–Mann–Whitney method, robust regression estimators, and even methods aimed at comparing quantiles other than the population median. But the many details go beyond the scope of this article.

Another common strategy is to discard valid values among the dependent variable that are flagged as an outlier, and then apply conventional methods based on means using the remaining data. This is technically unsound because it results in an incorrect estimate of the standard error. Using theoretically sound methods for dealing with outliers can make a substantial difference as illustrated, for example, in Wilcox (2012a, 2016). Even with a large sample size, invalid inferences can occur due to using a technique that violates basic principles. For example, the standard error of the sample mean is $\sigma/\sqrt{n}$, which is derived under the assumption that the observations $X_1, \dots, X_n$ are a random sample. In particular, these n observations are assumed to be independent. But when outliers are empirically identified and removed, the remaining observations are no longer independent, in which case the derivation of the standard error is no longer valid. More details about this issue and why it has practical importance can be found in Wilcox (2012a, 2012b).

Many other misconceptions are rampant, and other serious technical issues are not addressed in a theoretically sound manner. Part of the problem is that basic training

has not kept pace with advances in the statistics literature. This is beginning to change but it is evident that most non-statisticians are not adequately trained in the basic principles relevant to understanding data. These issues are fundamental when trying to understand the role of hypothesis testing methods. Consequently, some of these issues are reviewed here in the hope that it helps non-statisticians in their attempt to understand data and that it helps accelerate efforts aimed at providing improved training.

One fundamental goal has been finding mathematical methods for characterizing the extent population measures of location and scale (variation) are relatively insensitive to small changes in a distribution. As previously noted, a basic idea is that a small change in a distribution should not have an arbitrarily large impact on the associated parameters. In particular, a small shift away from a normal distribution should not substantially alter the values of the population mean and variance. There are three mathematical criteria used to measure the extent to which a parameter satisfies this goal (e.g., Staudte & Sheather, 1990; Wilcox, 2012b), but the details go beyond the scope of this article. (An outline of these three criteria is given in the appendix.) Here, it is merely noted that the population mean μ and variance σ^2 do not satisfy any of these criteria. This is not to suggest that all departures from normality are a serious concern. And it is not being suggested that the population mean and variance never have practical value. Certainly this is not the case. But the reality is that for a broad range of situations, there are fundamental concerns associated with the population mean and variance. Or from the point of view of analyzing data, there are concerns associated with the sample mean $\bar{X}$ and the sample variance s^2 regardless of how large the sample size might be.

A well-known illustration that the population variance is sensitive to slight departures from normality stems from Tukey (1960) and is based on a particular mixed normal distribution. Figure 1 shows a standard normal distribution and the mixed normal discussed by Tukey. As is evident, both distributions have a population mean $\mu = 0$. The variance of the standard normal is $\sigma^2 = 1$, but the variance of the mixed normal is $\sigma^2 = 10.9$. Roughly, this illustrates that the population variance, σ^2 , is highly sensitive to the tails of a distribution. In effect, there is the potential for relatively poor power and misleading measures of effect size when using any method based on the sample mean and the standard deviation, or when using Pearson's correlation and least squares regression. There is a vast literature on how to deal with these concerns, and there are numerous examples that they can make a substantial difference when analyzing data (e.g., Heritier et al., 2009; Staudte & Sheather, 1990; Wilcox, 2012a, 2012b, 2016).

A related concern is that even a single outlier can have a substantial impact on both the sample mean and variance. As a result, the sample mean, for example, can poorly reflect the typical response and the standard error of the sample mean can be substantially larger than the standard error associated with other location estimators that might be used. From basic principles, relatively large standard errors can result

Figure 1. Shown are a standard normal distribution (solid line) and a mixed normal. The standard normal has variance 1 and the mixed normal has variance 10.9 despite the apparent similarity of the distributions.

in relatively poor power, relatively long confidence intervals, and a relatively low likelihood of replicating a result.

There are two basic strategies for dealing with the deleterious impact of outliers. The first is to trim a specified percentage of the data and the other is to trim any values flagged as an outlier. The usual sample median belongs to the first strategy, but a concern is that it might trim too many observations. In some situations a compromise between no trimming (the sample mean) and the maximum amount of trimming (the sample median) is preferable. For example, 20% trimming often is a good choice, meaning that the smallest 20% and the largest 20% are trimmed. One argument for trimming 20% is that its standard error compares well to the standard error of the mean when sampling from a normal distribution, and as we move toward a heavy-tailed distribution, the standard error of the 20% trimmed mean can be substantially smaller. Both theory and simulations indicate that the accuracy of confidence intervals improves as the amount of trimming increases (Wilcox, 2012b).

The second general approach is to use a location estimator that empirically down weights or eliminates outliers. Using a good outlier detection technique is crucial, meaning that it should be able to avoid masking. (Masking means that the very presence of outliers causes them to be missed.) This eliminates any outlier detection method based on the mean and variance because the mean, and especially the

variance, can be inflated by outliers. A boxplot is one way to avoid masking. Alternative methods and their relative merits are summarized in Wilcox (2012b). At a superficial level, this approach would seem more reasonable than using a trimmed mean because a trimmed mean might eliminate values that are not outliers. However, this issue is not straightforward and there are various reasons to suspect that for many purposes, often a 20% trimmed mean is preferable. This is not to suggest that a 20% trimmed mean is always optimal: it is not. That is, neither approach dominates based on the various criteria used to compare location estimators. But the details go beyond the scope of this article. Readers interested in these details are referred to Wilcox (2012a, 2012b).

What is particularly relevant to the present article are results dealing with the goal of computing accurate confidence intervals. That is, confidence intervals that have probability coverage reasonably close to the nominal level. Theoretical and simulation studies indicate that accurate confidence intervals can be achieved for a much wider range of situations, compared to methods based on the mean or least squares regression, when using robust measures of location and robust regression methods. These advances include modern improvements on the Wilcoxon–Mann–Whitney test, methods for comparing quantiles, inferences based on regression smoothers, various multivariate techniques, and major advances related to the analysis of covariance (e.g., Wilcox, 2012a, 2012b, 2016). These results have important implications about NHT and p values, some of which are described in the next section.

Hypothesis Testing, p Values, and Confidence Intervals

Although it can be argued that a p value conveys useful information, it seems fairly evident that it does not tell us everything we would like to know about the distributions being studied. For example, it does not indicate $P(\theta > \theta_0)$ and under general conditions it does not reflect the extent θ_1 and θ_2 differ. (An important exception is described in the next section.) But in fairness, the same criticism applies to virtually all summaries of data. In effect, they reduce data down to the point that certain features of great practical importance can be missed. This is certainly true for any single measure of location as well as any measure of effect size as discussed and illustrated in the next section. Consequently, when analyzing data, multiple perspectives would seem to be essential in many situations.

Of course, plots can be invaluable, but even plots do not tell us everything we would like to know about the data. With a large sample size, plots might suggest how groups compare, and they can indicate the nature of the association between two variables that might not be evident based on standard least squares regression methods. But just how large must the sample size be in order to argue that plots suffice without any consideration of the many inferential methods that can be used? There are many ways of addressing this issue, but in the end there will be different opinions about how this should be done.

From a robustness point of view, when dealing with means, there is a concern about p values that is also relevant when computing a confidence interval: the assumption that the probability of a Type I error can be controlled reasonably well or that the actual probability coverage of a confidence interval is reasonably close to the nominal level. If sampling is from a symmetric, light-tailed distribution, confidence intervals based on Student's t perform reasonably well. As we move toward a skewed distribution, eventually it yields inaccurate confidence intervals, even when outliers are rare as illustrated, for example, in Wilcox (2012a, 2016).

For the two-sample case, the accuracy of confidence intervals for the difference between the means is a complex function of the sample sizes, the extent there is heteroscedasticity, the extent distributions differ in skewness, and the likelihood of encountering outliers. When θ is taken to be the population 20% trimmed mean or median, the Type I error probability can be controlled very well even for seemingly extreme departures from normality and relatively small sample sizes (e.g., Wilcox, 2012a, 2012b). When testing (1), for example, and θ is the population median, a distribution-free technique is available. When testing (2), a slight variation of a percentile bootstrap method performs very well using the sample median, even when there are tied values. (It has long been known that rank-based methods are unsatisfactory given the goal of testing the hypothesis that two independent groups have equal medians. See, for example, Hettmansperger, 1984.) And the method continues to perform well when using a 20% trimmed mean. So when dealing with robust measures of location, viewing hypothesis testing from the perspective of Tukey's three decision rule is reasonable for a broad range of situations, and p values would seem to provide at least some useful information.

However, when using any method based on the mean, realistic departures from normality can result in confidence intervals for which the probability coverage differs substantially from the nominal level (e.g., Westfall & Young, 1993; Wilcox, 2012a). For the one-sample Student's t test, the actual Type I error probability tends to be less than or equal to the nominal level when sampling from symmetric distributions that have heavy tails. But suppose that when computing a 0.95 (95%) confidence interval for the mean, the actual probability coverage is deemed satisfactory if it is between 0.925 and 0.975. (This is based on the so-called liberal criterion suggested by Bradley, 1978.) When sampling from a skewed, light-tailed distribution, a sample size of 200 might be needed to get good control over the Type I error probability (Westfall & Young, 1993; Wilcox, 2012a). For skewed, heavy-tailed distributions, now a sample size of 300 or more might be needed. When comparing two independent groups using means, now a crucial issue is the extent the distributions differ in terms of skewness. (For illustrations, see, e.g., Wilcox, 2012b, 2016.) But even with much larger sample sizes, not all concerns are eliminated for the simple reason that different methods are sensitive to different features of the distributions. For example, for skewed distributions, the difference between the medians might be substantially smaller or larger than the difference between the means.

Of course, some departures from normality are innocuous when dealing with means. But given some data, diagnostic tools for determining whether this is the case are relatively ineffective (e.g., Wilcox, 2012a). Currently, the safest strategy for determining whether a robust method makes a practical difference is to actually apply one.

So how should a p value be interpreted when dealing with means? When comparing two independent groups, a p value close to zero means that we can be reasonably certain that the distributions differ. From the point of view of Tukey's three decision rule, this is not a very interesting finding if we accept the idea that even with no data, surely there is some difference albeit possibly small and unimportant. The main point is that the strength of the empirical evidence that the group with the larger population mean has been correctly identified is unclear. In a broader context, p values provide useful information when using robust methods, but it might be meaningless when dealing with means or least squares regression.

Comments on Measures of Effect Size

Cohen's d arises quite naturally under the classic assumption that for two independent groups both have a normal distribution with a common variance. Cohen's d is given by

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p}, \quad (3)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the usual sample means and s_p^2 is the (mean squares within groups) estimate of the assumed common variance. As is evident, it is designed to estimate

$$\delta = \frac{\mu_1 - \mu_2}{\sigma_p}, \quad (4)$$

where μ_1 and μ_2 are the population means and σ_p^2 is the assumed common variance. But δ is not robust because it is based on the mean and variance. Moreover, d can miss a large and important difference among the bulk of the participants due to a few outliers (e.g., Algina, Keselman, & Penfield, 2005). Another concern is that it does not deal effectively with unequal variances.

Consider, for example, two normal distributions with means 0 and 0.8, and a common standard deviation equal to one. Then $\delta = 0.8$, which is often described as being relatively large. But if the distributions are mixed normals, like the mixed normal in Figure 1, $\delta = 0.24$, which is generally viewed as being relatively small despite the fact that graphically there is little difference between these two situations.

If it is assumed that normality and homoscedasticity are a reasonable approximation of reality, then Cohen's d adds little or no additional information beyond knowing the p value and the sample sizes when comparing two independent groups. The

reason is that Cohen's d can be computed based on the p value and the sample sizes (Browne, 2010). If it is decided that normality and homoscedasticity are not a reasonable approximation of reality, then both Cohen's d and p values associated with methods based on the mean can be misleading. So, at some level, the suggestion that Cohen's d is more meaningful than a p value is nonsensical.

Algina et al. (2005) suggest a robust generalization of Cohen's d that replaces the mean and variance with a trimmed mean and a Winsorized variance. To illustrate what it means to Winsorize data, imagine a sample of 10 observations. Trimming 20% means that the two smallest and two largest values are removed and the remaining data are averaged. Winsorizing 20% means that rather than trim the two smallest values, they are set equal to the smallest value not trimmed. Simultaneously, rather than trim the two largest values, they are set equal to the largest value not trimmed. The sample variance based on the Winsorized values is the Winsorized variance.

Regarding unequal variances, a simple strategy is to replace s_p with s_1 or s_2 , the Winsorized standard deviation associated with Groups 1 and 2, respectively. But this can be rather unsatisfactory because the result can suggest a large effect or a small effect depending on whether s_1 or s_2 is used. Another possibility is to use a robust analog of Pearson's correlation (Wilcox & Tian, 2011).

Another approach to measuring effect size is to estimate $P(X_1 < X_2)$, the probability that an observation randomly sampled from the first group is less than a randomly sampled from the second group. The Wilcoxon–Mann–Whitney test is based on an estimate of this probability, but it is known that under general conditions it does not provide an accurate confidence interval (e.g., Wilcox, 2012a). Substantially improved methods for computing a confidence interval for $P(X_1 < X_2)$ have been derived that also provide a p value. For comparisons of these methods, see, for example, Neuhäuser, Lösch, and Jöckel (2007) as well as Ruscio and Mullen (2012). Given the goal of determining whether this probability is greater than 0.5, the p values associated with these newer techniques reflect the extent to which it is reasonable to make a decision. A fundamental concern with the Wilcoxon–Mann–Whitney test is that when distributions differ, under general conditions it uses an incorrect estimate of the standard error. In contrast, more modern methods use a correct estimate of the standard error even when the distributions differ.

An important issue is that different measures of effect size can paint a decidedly different picture regarding the extent groups differ. Consider, for example, a study by Pedersen, Miller, Putcha-Bhagavatula, and Yang (2002) that dealt with the sexual attitude of young adults. The general goal was to assess whether men seek more short-term mates than women. A portion of the study asked males ($n=105$) and females ($n=156$) how many sexual partners they desire over the next 30 years. Comparing the mean responses, the p value is 0.279 and Cohen's $d=0.168$, which is typically interpreted as being small. Trimming just 1% and using the method derived by Yuen (1974) now the p value is 0.04 and the Algina et al. (2005) robust measure of effect size is $d_t=0.31$. Increasing the amount of trimming to 10%, now the p value is 0.003 and $d_t=0.46$, which is typically interpreted as being a medium

effect size. The distributions are skewed with outliers, which helps explain these results. Switching to the robust measure of effect size derived by Wilcox and Tian (2011) gives similar results. So an initial impression might be that outliers are masking a medium effect size among the bulk of the participants. However, the median response for both males and females is one. So any measure of effect size based on the median is zero, the point being that several measures might be needed to get a good sense of how groups compare.

Concerns About Bayesian Methods

The primary goal in this section is to comment on Bayesian methods from the perspective of modern robust techniques. Some results on computing accurate posterior intervals are provided. But a more fundamental issue has to do with the notion of robustness from a broader perspective as discussed in the third section. In particular, there is the issue of characterizing the impact of arbitrarily small changes in a distribution. As previously noted, and as outlined in the appendix, the mathematical techniques for dealing with this issue are well developed from a frequentist point of view. The goal here is to comment on where we stand when a Bayesian perspective is adopted.

Before continuing, some general comments about Bayesian methods might help. The first approach to making inferences about some unknown parameter that characterizes some distribution was developed by Bayes in connection with a binomial probability function. A goal was to find some interval that is likely to include the true probability of success, p . Bayes addressed this problem in two papers published posthumously in the years 1764 and 1765. The strategy was to assume that p has a uniform distribution. Letting X denote the number of successes in n trials, Bayes showed that the (posterior) distribution of p given X is a beta distribution, which today is easily evaluated using extant software. In the year 1774, Laplace developed the same method, apparently independently of Bayes (Hald, 1998). Laplace made major contributions that included extensions to making inferences about the population mean of a normal distribution. From the point of view of inferential methods, a Bayesian approach dominated for about 50 years simply because it was the only method available. Then, in the year 1811, Laplace developed the frequentist approach to statistical inference, and in 1814, he developed a new approach to computing a confidence interval. Rather than view parameters as random variables, as done by Bayes, Laplace viewed parameters as fixed, unknown constants. Inferences are made based solely on the available data in the context of a sampling distribution, using standard techniques that are routinely covered in a typical introductory statistics course. (Of course, Laplace did not have the benefit of Student's t distribution.) In effect, Laplace created a controversy that is still with us today: When and how should a Bayesian method be used?

Bayesian methods aimed at making inferences about a population parameter became controversial soon after Laplace's frequentist approach was developed.

Gauss, for example, was highly critical of the Bayesian approach and preferred the frequentist approach (Hald, 1998). Today the controversy remains. For example, Gelman (2008), who is an advocate of the Bayesian approach, states, “Bayesian inference is one of the more controversial approaches to statistics” (p. 445). His paper is devoted to summarizing why. Clearly some psychologists look favorably upon Bayesian methods (e.g., Cumming, 2014; Morey, Rouder, Verhagen, & Wagenmakers, 2014). But some statisticians argue that it should not be used (e.g., Norton, 2011).

It is not our goal to discuss the controversies surrounding Bayesian methods other than to acknowledge that controversies exist. It seems safe to predict that agreement about when and how a Bayesian method should be used will not occur any time soon. But when discussing Bayesian methods, it seems important to at least acknowledge the controversies surrounding Bayesian techniques for the benefit of non-statisticians who might not be aware that controversies exist. The point here is that from a robustness point of view, there are serious concerns beyond the issues discussed by Gelman (2008). It is not being suggested, however, that a Bayesian perspective never has any practical value. For example, from a philosophy of science perspective, Earp and Trafimow (2015) use a Bayesian analysis to illustrate a point regarding the role of replicating a study given the goal of supporting or refuting some original result. The concerns about Bayesian methods discussed here are not relevant to their analysis.

Switching to a Bayesian method does not alter the fact that the population mean and variance are not robust. The classic Bayesian method assumes that the population mean has a normal distribution. It is further assumed that given μ , the observed data have a normal distribution. Efforts have been made to deal with nonnormality (e.g., Minsker, Srivastava, Lin, & Dunson, 2014), but more research is needed to determine whether a practical method can be derived. One strategy is to model the prior with some heavy-tailed distribution (e.g., Fúquene, Cook, & Pericchi, 2009). But this does not necessarily deal with situations where the conditional distribution of X , given μ , is nonnormal. Holmes (2014) summarizes the inherent concerns about the robustness of Bayesian methods in the context of the mathematical foundations of robust methods as described in Huber and Ronchetti (2009). In practical terms, from a Bayesian point of view, how does one avoid excessively long posterior intervals when dealing with heavy-tailed distributions? Another serious issue is how to get reasonably accurate posterior intervals when dealing with skewed distributions, particularly when outliers are likely to occur. Another basic concern is that Bayesian methods generally impose some parametric family of distributions. When attention is focused on the mean, the concern is that the basic requirements for a robust parameter, outlined in the appendix, will not be met. (For more mathematical details, see the definition of infinitesimal robustness and the derivation of an influence function in Wilcox 2012b, section 2.1.2, or Staudte and Sheather, 1990.)

To illustrate a concern with skewed distributions, a few simulation results are reported when generating data from a g -and- h distribution (Hoaglin, 1985). If Z has a standard normal distribution, then by definition

$$V = \begin{cases} \frac{\exp(gZ)-1}{g} \exp(hZ^2/2), & \text{if } g > 0 \\ Z \exp(hZ^2/2), & \text{if } g = 0 \end{cases}$$

has a g -and- h distribution where g and h are parameters that determine the first four moments. A normal distribution corresponds to $g = h = 0$. With $g = 1$ and $h = 0$, the g -and- h distribution corresponds to a lognormal distribution that has been shifted to have a population median equal to zero. (The skewness is 6.2, and the kurtosis is 114.) This distribution is relatively light-tailed in the sense that the proportion of points likely to be declared an outlier, based on a boxplot, is relatively small. (The expected proportion of points declared an outlier is approximately 7%.) Computing a Bayesian 0.95 posterior interval, assuming normality and a noninformative prior, the actual probability coverage, based on sample sizes 20, 40, and 100 is .855, .880, and .901, respectively. (These estimates are based on simulations with 10,000 replications.) Moving toward a more heavy-tailed distribution where again $g = 1$ but $h = .2$, in which case the expected proportion of points declared an outlier is approximately 8%, the estimates are now .765, .801, and .841. For $g = h = .2$, in which case the expected proportion of points declared an outlier is approximately 5%, the estimates are .932, .934, and .935. So the accuracy of any posterior interval depends on a combination of at least three factors: skewness, the likelihood of encountering outliers, and the sample size.

Consider again the sexual attitude study. Focusing on the males, a Bayesian 0.95 posterior interval for the mean, which assumes normality with an uninformative prior, is $(-45.5, 141.5)$. Using Student's t , the 0.95 confidence interval is $(-48.3, 178.2)$. There is one extreme outlier: 6,000. If it can be argued that this response is invalid and therefore can be discarded, now a 0.95 confidence interval, based on Student's t distribution, is $(3.7, 12.0)$, and the Bayesian 0.95 posterior interval is $(3.8, 11.8)$. Using all of the data, the 0.95 confidence intervals for the population 20% trimmed mean and median are $(1.6, 3.7)$ and $(1, 3)$, respectively. So confidence intervals and Bayesian posterior intervals can differ substantially in terms of their length as well as the value around which they are centered.

Concluding Remarks

In summary, it is suggested that NHT and p values are among a vast array of tools that help provide a deep and accurate understanding of data. Situations are encountered where p values can be misleading when using the more obvious hypothesis testing techniques. But there is a broad range of situations for which p values provide useful information in the context of Tukey's three decision rule. Of course, more than p values is needed to understand data, but caution is needed because the more obvious methods for computing confidence intervals and measuring effect size are not robust. The same concerns apply when using extant Bayesian techniques. There is a vast literature regarding robust frequentist (non-Bayesian) methods, but at the moment truly robust Bayesian methods have not been derived.

There is no single method that is always optimal. But this does not mean that all methods should be abandoned. A seemingly natural strategy is to use diagnostic tools to determine which method is best. But because the relative merits of competing methods depend on several factors that interact in complicated ways, this approach can be unsatisfactory. Currently, the only reliable way of determining whether two different techniques yield different perceptions of data is to use both and compare the results.

Of course, testing multiple hypotheses raises the issue of controlling the probability of one or more Type I errors. There are well-known techniques for dealing with this issue (e.g., Hochberg & Tamhane, 1987; Wilcox, 2012b). A possible concern is that when controlling the probability of one or more Type I errors, as the number of tests being performed increases, power will decrease. For this reason, exploratory methods seem crucial. Confirmatory studies could then be used with a limited number of hypothesis testing techniques so as to control the probability of one or more Type I errors and simultaneously achieve relatively high power. A possible advantage of this two-stage procedure is that it helps address the issue of replication.

Appendix

This appendix briefly outlines three basic methods used to judge the robustness of a parameter. For a more detailed description of these criteria plus related issues, see, for example, Hampel et al. (1986), Huber and Ronchetti (2009), Staudte and Sheather (1990), and Wilcox (2012b).

Consider any function $f(x)$, not necessarily a probability density function. Suppose it is desired to impose a restriction on this function so that it does not change drastically with small changes in x . One way of doing this is to insist that $f(x)$ be continuous. Next, note that measures of location can be viewed in the context of functionals. That is, a measure of location is based on a function that maps a (probability density) function into the real line. The point is that the notion of continuity has been extended to functionals. Parameters that correspond to a continuous functional are said to have qualitative robustness. The population mean μ does not have this property in contrast to a γ trimmed mean, $\gamma > 0$.

Again consider any function $f(x)$, not necessarily a probability density function. Another way of ensuring that it does not change drastically with small changes in x is to insist that it has a bounded derivative. Viewing this requirement from the point of view of a functional, parameters that satisfy this criterion are said to have a bounded influence function and have infinitesimal robustness. Again, μ does not satisfy this criterion in contrast to a γ trimmed mean, $\gamma > 0$.

Finally, a parameter does not have quantitative robustness if an arbitrarily small change in a distribution can alter its value in an arbitrarily large manner. For a formal proof that μ does not satisfy this criterion in contrast to a γ trimmed mean, $\gamma > 0$, see, for example, Staudte and Sheather (1990) or Wilcox (2012b).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10, 317-328.
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Branch, M. N. (2014). Malignant Side Effects of null-hypothesis significance testing. *Theory & Psychology*, 24, 256-277.
- Browne, R. H. (2010). The *t*-test, *p* value and its relationship to the effect size and $P(X > Y)$. *American Statistician*, 64, 30-33 (correction, 64, 195).
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997-1003.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7-29.
- Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*. Advance online publication. doi:10.3389/fpsyg.2015.00621
- Fuquene, J. A., Cook, J. D., & Pericchi, L. R. (2009). A case for robust Bayesian priors with applications to clinical trials. *Bayesian Analysis*, 4, 817-846.
- Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, 3, 445-450.
- Hald, A. (1998). *A history of mathematical statistics from 1750 to 1930*. New York, NY: Wiley.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York, NY: Wiley.
- Heritier, S., Cantoni, E., Copt, S., & Victoria-Feser, M.-P. (2009). *Robust methods in biostatistics*. New York, NY: Wiley.
- Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. New York, NY: Wiley.
- Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Exploring data tables trends and shapes* (pp. 461-514). New York, NY: Wiley.
- Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: Wiley.
- Holmes, C. (2014). *Robust (Bayesian) inference*. Retrieved from http://mlss2014.hiit.fi/mlss_files/Holmes-MLSS.pdf
- Huber, P. J., & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). New York, NY: Wiley.
- Jones, L. V., & Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods*, 5, 411-414.
- Kline, R. B. (2013). *Beyond significance testing* (2nd ed.). Washington, DC: American Psychological Association.

- Maronna, R. A., Martin, D. R., & Yohai, V. J. (2006). *Robust statistics: Theory and methods*. New York, NY: Wiley.
- Minsker, S., Srivastava, S., Lin, L., & Dunson, D. (2014). *Robust and scalable Bayes via a median of subset posterior measures* (Technical report). Retrieved from arxiv.org/pdf/1403.2660
- Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, 25, 1289-1290.
- Neuhäuser, M., Lösch, C., & Jöckel, K.-H. (2007). The Chen–Luo test in case of heteroscedasticity. *Computational Statistics & Data Analysis*, 51, 5055-5060.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5, 241-301.
- Norton, J. D. (2011). Challenges to Bayesian confirmation theory. In D. M. Gabbay, P. Thagard, & J. Wood (Eds.), *Handbook of the philosophy of science* (Vol. 7, pp. 391-440). New York, NY: Elsevier.
- Pedersen, W. C., Miller, L. C., Putcha-Bhagavatula, A. D., & Yang, Y. (2002). Evolved sex differences in sexual strategies: The long and the short of it. *Psychological Science*, 13, 157-161.
- Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression & outlier detection*. New York, NY: Wiley.
- Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, 47, 201-223.
- Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York, NY: Wiley.
- Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow, & H. Mann (Eds.), *Contributions to probability and statistics* (pp. 448-485). Stanford, CA: Stanford University Press.
- Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science*, 6, 100-116.
- Westfall, P. H., & Young, S. S. (1993). *Resampling based multiple testing*. New York, NY: Wiley.
- Wilcox, R. R. (2012a). *Modern statistics for the social and behavioral sciences: A practical introduction*. New York, NY: Chapman & Hall/CRC Press.
- Wilcox, R. R. (2012b). *Introduction to robust estimation and hypothesis testing* (3rd ed.). New York, NY: Elsevier.
- Wilcox, R. R. (2016). *Understanding and applying basic statistical methods using R*. New York, NY: Wiley.
- Wilcox, R. R., & Tian, T. (2011). Measuring effect size: A robust heteroscedastic approach for two or more groups. *Journal of Applied Statistics*, 38, 319-331.
- Yuen, K. K. (1974). The two sample trimmed *t* for unequal population variances. *Biometrika*, 61, 165-170.