Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 1

3-6-2019

A Robust Nonparametric Measure of Effect Size Based on an Analog of Cohen's *d*, Plus Inferences About the Median of the Typical Difference

Rand Wilcox
University of Southern California, rwilcox@usc.edu

Recommended Citation
Wilcox, R. (2018). A robust nonparametric measure of effect size based on an analog of Cohen's *d*, plus inferences about the median of the typical difference. *Journal of Modern Applied Statistical Methods, 17*(2), eP2726. doi: 10.22237/jmasm/1551905677

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2726
doi: 10.22237/jmasm/1551905677

בס"ד

ISSN 1538 − 9472

INVITED ARTICLE

A Robust Nonparametric Measure of Effect Size Based on an Analog of Cohen's *d*, Plus Inferences About the Median of the Typical Difference

Rand Wilcox
University of Southern California
Los Angeles, CA

The paper describes a nonparametric analog of Cohen's *d*, *Q*. It is established that a confidence interval for *Q* can be computed via a method for computing a confidence interval for the median of $D = X_1 - X_2$, which in turn is related to making inferences about $P(X_1 < X_2)$.

Keywords: Nonparametric methods, Cohen's *d*, Wilcoxon-Mann-Whitney method, bootstrap methods

Introduction

When comparing two independent groups, there are now a variety of methods aimed at measuring effect size (e.g., Algina, Keselman, & Penfield, 2005; Grissom & Kim, 2012; Wilcox, 2017a). For two independent random variables, say X_1 and X_2, let μ_j and σ_j denote the population mean and standard deviation, respectively, associated with the j^{th} group (j = 1, 2). Certainly, one of the better-known measures of effect size is

$$\delta = \frac{\mu_1 - \mu_2}{\sigma_p} \tag{1}$$

doi: 10.22237/jmasm/1551905677 | Accepted: October 31, 2017; Published: March 6, 2019.
Correspondence: Rand Wilcox, rwilcox@usc.edu

Rand Wilcox is a Professor in the Department of Psychology, University of Southern California. His primary interests are robust and nonparametric statistical methods.

where by assumption $\sigma_1 = \sigma_2 = \sigma_p$, say. Based on a random sample of size n_j from the j^{th} group, let $\bar{X}_j$ and s_j denote the sample mean and standard deviation, respectively. Letting

$$S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

the usual estimate of δ is

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S} \tag{2}$$

which is generally known as Cohen's *d*. There are, however, three fundamental concerns associated with Cohen's *d* that are reviewed in the next section. Some of these concerns have been addressed, but some have not.

Another well-known measure of effect size is

$$p = \mathrm{P}(X_1 < X_2) \tag{3}$$

the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second group. The Wilcoxon-Mann-Whitney (WMW) test is based in part on an estimate of *p*. In effect, an estimate of the distribution of $D = X_1 - X_2$ is used. A concern, however, is that under general conditions the WMW test performs poorly in terms of computing a confidence interval for *p*. The basic problem is that the WMW method uses an incorrect estimate of the standard error when distributions differ. Several methods have been derived for dealing with this issue, which are summarized in Wilcox (2017a).

Note that the common goal of testing

$$H_0 : p = 0.5 \tag{4}$$

is equivalent to testing

$$H_0 : \theta_D = 0 \tag{5}$$

where θ_D is the median of D. Certainly p is a useful indication of the extent two distributions differ. But θ_D is intrinsically interesting and it helps provide perspective beyond p.

One of the main goals in this paper is to suggest a measure of effect size, Q, that captures the spirit of Cohen's d and simultaneously deals with three concerns associated with d that are reviewed in the next section. The suggested approach is based in part on the estimate of the distribution of D that is used by the WMW test. An advantage of this approach is that it eliminates the issue of how to deal with heteroscedasticity, and it deals with non-normality in a sense to be described.

Let θ_j denote the population median of X_j (j = 1, 2). As is well known, under general conditions $\theta_1 - \theta_2 \neq \theta_D$. While testing H_0: $\theta_1 = \theta_2$ has been studied extensively (e.g., Wilcox, 2017a), evidently methods for computing a confidence interval for θ_D have received little to no attention. A second goal here is to describe and compare three methods for computing a confidence interval for θ_D. As will be seen, methods for computing a confidence interval for p can be used to compute a confidence interval for θ_D, which in turn can be used to compute a confidence interval for the measure of effect size Q.

The paper is organized as follows: The next section reviews practical concerns regarding Cohen's d. The third section suggests a measure of effect size, Q, aimed at dealing with these concerns and how it can be easily estimated. This is followed by a description of methods for computing a confidence interval for θ_D and Q. Simulations are used to compare these methods, which is followed by two illustrations.

Concerns about Cohen's *d*

Note that under normality and homoscedasticity, δ is reasonable in the sense that it provides a probabilistic sense of what a shift of δ standard deviations means. Of course, what constitutes a large effect size can depend on the situation. But to provide some perspective, consider the frequently adopted view (e.g., Cohen, 1988) that δ = 0.2, 0.5, and 0.8 correspond to small, medium, and large effect sizes, respectively. From basic principles, if the mean of a distribution is increased by 0.2 standard deviations, this corresponds to shifting the mean to the 0.58 quantile. That is, for any normal distribution, $\mu + 0.2\sigma$ corresponds to the 0.58 quantile. Similarly, δ = 0.5 and 0.8 correspond to shifting the mean to the 0.69 and 0.79 quantiles, respectively.

There are, however, fundamental concerns regarding both δ and Cohen's d. First, even a small departure from normality toward a more heavy-tailed

distribution can result in a relatively small value for δ when in fact, based on plots of the distributions, there is a relatively large difference (e.g., Algina et al., 2005; Wilcox, 2017a).

Consider, for example, the mixed normal distribution used by Algina et al. (2005). This distribution consists of sampling an observation from a standard normal distribution with probability 0.9; otherwise an observation is sampled from a normal distribution with mean zero and standard deviation ten. Figure 1 shows the distribution of this mixed normal as well as the standard normal. Although the variance of the standard normal is one, the variance of the mixed normal is 10.9.

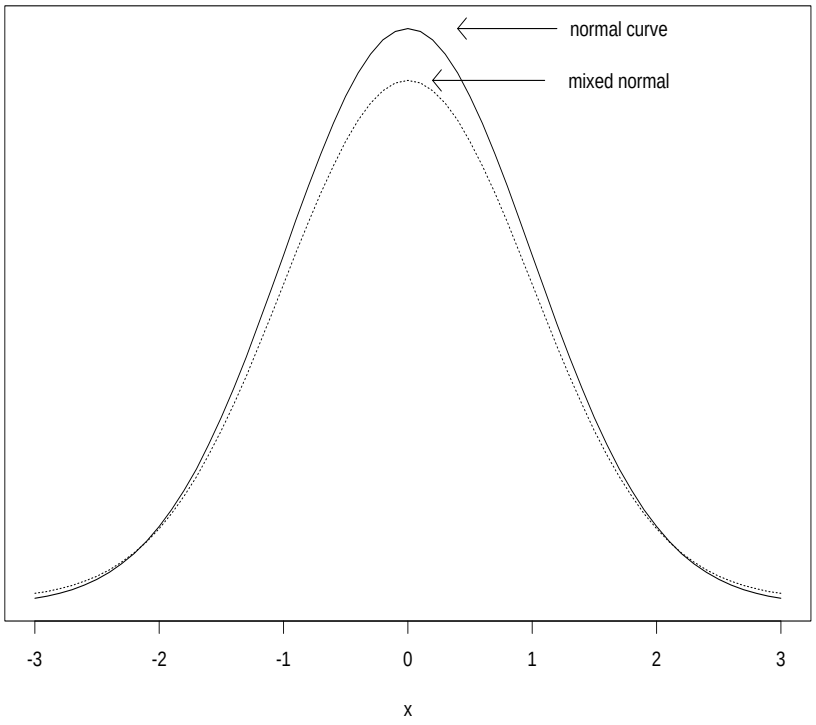

Figure 1. Shown are the standard normal and mixed normal distributions; the two curves shown here have an obvious similarity, yet the variances are 1 and 10.9

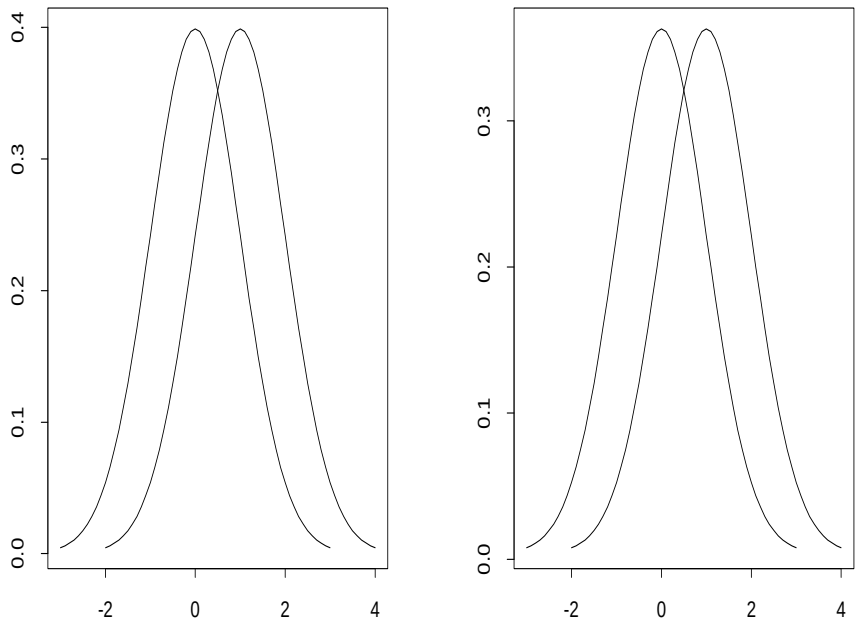

Figure 2. In the left panel, $\delta = 1$; in the right panel, $\delta = 0.3$, illustrating that a slight departure from normality can alter δ substantially

---

Now look at Figure 2. In the left panel are two normal distributions. Both have variances equal to one and the means are zero and one, so $\delta = 1$. In the right panel, the means are again zero and one, but the distributions are mixed normals and now $\delta = 0.3$. In terms of *d*, even a single outlier can inflate S^2, the estimate of the assumed common variance, which can result in a relatively small value for *d* even when, for the bulk of the data, there is a relatively large effect.

A related concern is the negative impact on the probabilistic interpretation of δ. Consider, for example, two mixed normals where one is shifted to have a population mean equal 0.8. So $\delta = 0.8/\sqrt{10.9} = 0.24$, which is often interpreted as being relatively small. However, from the probabilistic interpretation underlying Cohen's *d*, shifting the second mean by 0.8 corresponds to shifting it to the 0.76 quantile. That is, to the extent the probabilistic interpretation of δ is deemed

reasonable, there is a very large effect size in contrast to what is indicated by δ. In more general terms, even a small departure from normality, toward a heavy-tailed distribution, can render the probabilistic interpretation of δ misleading and even meaningless.

Skewed distributions also create concerns from the probabilistic point of view associated with δ. As an illustration, consider a lognormal distribution, which Gleason (1993) argues is relatively light-tailed. Then the population mean, $\mu = \sqrt{\exp(1)}$, corresponds to the 0.69 quantile. Now consider a second lognormal distribution that has been shifted to have mean $\mu - 0.8\sigma$, in which case $\delta = 0.8$. But $\mu - 0.8\sigma = -0.08$. So from a probabilistic point of view this effect is much bigger than what is gleaned from the interpretation of Cohen's *d* under normality because, for the lognormal distribution, $P(X < 0) = 0$.

There are at least two other concerns associated with skewed distributions. First, there is no distinction between δ and $-\delta$. When D has a symmetric distribution this is reasonable, but otherwise this is not necessarily the case from the point of view of shifting a measure of location to some quantile. Second, the mean can reflect a relatively atypical response. The strategy here is to deal with these issues by focusing on the median of D. For the special case of a lognormal distribution, of course one could simply transform to a normal distribution by taking logs. But usually simple transformations do not effectively deal with skewed distributions. As illustrated, for example, in Wilcox (2017c), transformed data can remain skewed to the point that practical concerns are not adequately addressed. Moreover, simple transformations do not deal effectively with concerns associated with heavy-tailed distributions.

Of course, there is the practical issue of whether the probabilistic interpretation of the standard deviation, under normality, is misleading based on estimates of the mean and standard deviation using data from an actual study. Illustrations that this is the case are given in Wilcox (2017a, b). To provide yet another example, consider the cortisol awakening response (CAR), which is just the difference between cortisol measured upon awakening and again about 30-45 minutes later. The CAR has been found to be associated with various measures of stress. Both enhanced and reduced CARs are associated with various psychosocial factors including depression and anxiety disorders (e.g., Bhattacharyya, Molloy, & Steptoe, 2008; Pruessner, Hellhammer, & Kirschbaum, 2003). In most studies the CAR has been found to be negative (salivary cortisol levels increase after awakening). In the Well Elderly 2 study (Clark et al., 2012), of interest was the extent cortisol increases or decreases among older adults after completion of an

intervention program generally aimed at improving their overall physical and mental health. The sample size was $n = 328$. The point here is that shifting the data by 0.8 standard deviations to the left is tantamount to shifting the sample mean to the 0.05 quantile. That is, from the probabilistic interpretation of Cohen's *d* under normality, the effect size is estimated to be larger than what is indicated by $d = 0.8$.

Yet one more concern is the homoscedasticity assumption. A simple and well-known method for dealing with heteroscedasticity ($\sigma_1 \neq \sigma_2$) is to use two measures of effect size, namely $\delta_j = (\mu_1 - \mu_2) / \sigma_j$ ($j = 1, 2$). So, a large or small effect size might be indicated depending on whether δ_1 or δ_2 is used. And of course, this does not eliminate the interpretational concerns previously described. The suggested measure of effect size eliminates the homoscedasticity assumption by focusing on the distribution of *D*.

Measures of Effect Size Based on the Distribution of *D*

First note that if two distributions are identical, *D* has a symmetric distribution about zero. Again, let θ_D be the population median associated with *D* and let F_0 be the distribution associated with $D - \theta_D$. That is, F_0 denotes the distribution of *D* when the null hypothesis

$$H_0 : \theta_D = 0 \tag{6}$$

is true. The basic idea is to measure effect size based on the extent θ_D represents a shift in location to some relatively high or low quantile associated with F_0. More formally, the measure of effect size is taken to be

$$Q = F_0(\theta_D) \tag{7}$$

For identical distributions, $\theta_D = 0$ corresponds to the 0.5 quantile, so $Q = 0.5$. If, for example, θ_D corresponds to a shift in location to the 0.8 quantile, $Q = 0.8$. This, of course, is very similar to the probabilistic interpretation of Cohen's *d* under normality and homoscedasticity, only no parametric assumption is made about the distribution of *D* and homoscedasticity is not assumed or required.

The effect size *Q* relates to Cohen's *d* in the following manner: Note that under normality and homoscedasticity, *D* has a normal distribution with variance $2\sigma^2$. If, for example, $\delta = 0.8$, this corresponds to shifting the median of F_0 from zero to 0.8σ. Moreover, $F_0(0.8\sigma) = P\left(Z \leq 0.8/\sqrt{2}\right)$, where *Z* has a standard normal distribution.

In particular, $\delta = 0.2$, 0.5, and 0.8 correspond to $Q = 0.556$, 0.638, and 0.714, respectively. Consequently, to the extent it is deemed reasonable to view $\delta = 0.2$, 0.5, and 0.8 as being small, medium, and large effect sizes, respectively, it follows that, roughly, $Q = 0.55$, 0.65, and 0.70 would be viewed as small, medium, and large effect sizes as well.

In a similar manner, it is approximately the case that for $\delta = -0.2$, -0.5, and -0.8; $Q = 0.45$, 0.35, and 0.30, respectively. Perhaps a more convenient perspective is the value of Q relative to 0.5. That is, one might use

$$\Omega = \frac{Q - 0.5}{0.5}$$

Now, low, medium and large effect sizes under normality and homoscedasticity would roughly correspond to $|\Omega| = 0.1$, 0.3, and 0.4, respectively, still assuming that Cohen's suggestion is deemed reasonable.

Next, consider the robustness of Q in terms of the mixed normal distribution used by Algina et al. (2005). A basic issue is whether a small shift away from normality can have an inordinate influence on a measure of effect size. One well-known way of measuring the difference between two distributions, say F and G, is Kolmogorov distance, which is the least upper bound of

$$\left|\mathrm{F}(x) - \mathrm{G}(x)\right|$$

If $\mathrm{F}(x)$ has a standard normal distribution, and $\mathrm{G}(x)$ has a mixed normal, the Kolmogorov distance is approximately 0.04 (Wilcox, 2017a, section 1.1). Consequently, if F_0 is shifted from a standard normal distribution to a mixed normal, Q will be altered by at most 0.04. In contrast, this small shift away from a normal distribution lowers δ substantially as previously noted. More broadly, for any situation where a small shift in a distribution, as measured by Kolmogorov distance, has an inordinate impact on δ, using Q instead, the impact will be less severe.

Estimation of Q is straightforward. Let X_{ij} ($i = 1,\ldots, n_j$; $j = 1, 2$) be a random sample from the j^{th} group. Then an estimate of the distribution of D can be based on the $n_1 n_2$ pairwise differences

$$D_{ik} = X_{i1} - X_{k2}$$

($i = 1,\ldots, n_1$; $k = 1,\ldots, n_2$). Let $\hat{\theta}_D$ be the sample median based on the $n_1 n_2$ D_{ik} values. An estimate of Q is simply

$$\hat{Q} = \frac{1}{n_1 n_2} \sum \sum \mathrm{I}\left(D_{ik} - \hat{\theta}_D \le \hat{\theta}_D \right) \tag{8}$$

where the indicator function $\mathrm{I}\left(D_{ik} - \hat{\theta}_D \le \hat{\theta}_D \right) = 1$ if $D_{ik} - \hat{\theta}_D \le \hat{\theta}_D$; otherwise $\mathrm{I}\left(D_{ik} - \hat{\theta}_D \le \hat{\theta}_D \right) = 0$. In other words, shift the estimated distribution of *D* so that it has a median of zero, the null case, by letting

$$Y_{ik} = D_{ik} - \hat{\theta}_D$$

Then $\hat{Q}$ is given by the proportion of Y_{ik} values that are less than or equal to $\hat{\theta}_D$.

In fairness, it can be argued that in some situations *d* can be larger than *Q* in some meaningful way simply because they are based on different measures of location. For example, imagine that X_1 has a standard normal distribution and X_2 is taken to be $2Y$, where *Y* has a lognormal distribution shifted to have a median equal to zero. Based on a simulation with 4000 replications, $n_1 = 50$ and $n_2 = 10$, $\mathrm{E}(d) = -0.7$ and $\mathrm{E}(\hat{Q}) = 0.46$. So $\hat{Q}$ tends to suggest a small effect size, roughly, because θ_D is approximately equal to −0.23. In contrast, $\mathrm{E}(X_1 - X_2) = -1.3$, which helps explains why *d* tends to be relatively large in contrast to $\hat{Q}$.

Confidence Intervals for θ_D and Q

This section describes three methods for computing a confidence interval for both θ_D and *Q*. One approach is closely related to extant heteroscedastic confidence intervals for *p*. Another approach is to use a basic percentile bootstrap method.

Let $\hat{p}$ be some estimate of *p*. The first approach for computing a confidence interval for θ_D begins with any method that uses a correct estimate of the standard error of $\hat{p}$ even when distributions differ. This eliminates the WMW test because it uses an incorrect estimate of the standard when distributions differ. Based on results in Neuhäuser, Lösch, and Jöckel (2007), the focus here is on the method derived by Cliff (1996, p. 140), but arguments can be made that certain alternative methods deserve serious consideration (e.g., Ruscio & Mullen, 2012).

Let

$$\Delta = \mathrm{P}\left(X_1 < X_2 \right) - \mathrm{P}\left(X_1 > X_2 \right) \tag{9}$$

Cliff focuses on a confidence interval for Δ, which is easily extended to computing a confidence interval for p. For the i^{th} observation in the first group and the h^{th} observation in the second group, let $d_{ih} = -1$, 0, or 1 depending on whether $X_{i1} < X_{h2}$, $X_{i1} = X_{h2}$, or $X_{i1} > X_{h2}$, respectively. An estimate of Δ is

$$\hat{\Delta} = \frac{1}{n_1 n_2} \sum \sum d_{ih} \tag{10}$$

and an estimate of p is

$$\hat{p} = \frac{1}{n_1 n_2} \sum \sum \mathrm{I}\left(D_{ik}\right) \tag{11}$$

where the indicator function $\mathrm{I}(D_{ik}) = 1$ if $D_{ik} < 0$; otherwise $\mathrm{I}(D_{ik}) = 0$.

Let

$$\bar{d}_{i.} = \frac{1}{n_2} \sum_h d_{ih},$$

$$\bar{d}_{.h} = \frac{1}{n_1} \sum_i d_{ih},$$

$$s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\bar{d}_{i.} - \hat{\Delta}\right)^2,$$

$$s_2^2 = \frac{1}{n_2 - 1} \sum_{h=1}^{n_2} \left(\bar{d}_{.h} - \hat{\Delta}\right)^2,$$

$$\text{ϑ}^2 = \frac{1}{n_1 n_2 - 1} \sum \sum \left(d_{ih} - \hat{\Delta}\right)^2$$

Then

$$\hat{\sigma}^2 = \frac{\left(n_1 - 1\right) s_1^2 + \left(n_2 - 1\right) s_2^2 + \text{ϑ}^2}{n_1 n_2}$$

estimates the squared standard error of $\hat{\Delta}$. Let z be the $1 - \alpha/2$ quantile of a standard normal distribution. Cliff's $1 - \alpha$ confidence interval for Δ is

$$\frac{\hat{\Delta}-\hat{\Delta}^3 \pm z\hat{\sigma}\sqrt{\left(1-\hat{\Delta}^2\right)^2+z^2\hat{\sigma}^2}}{1-\hat{\Delta}^2+z^2\hat{\sigma}^2}$$

Let

$$C_l = \frac{\hat{\Delta}-\hat{\Delta}^3 - z\hat{\sigma}\sqrt{\left(1-\hat{\Delta}^2\right)^2+z^2\hat{\sigma}^2}}{1-\hat{\Delta}^2+z^2\hat{\sigma}^2}$$

and

$$C_u = \frac{\hat{\Delta}-\hat{\Delta}^3 + z\hat{\sigma}\sqrt{\left(1-\hat{\Delta}^2\right)^2+z^2\hat{\sigma}^2}}{1-\hat{\Delta}^2+z^2\hat{\sigma}^2}$$

Then a $1-\alpha$ confidence interval for p is (p_ℓ, p_u), where $p_\ell = (1 - C_u)/2$ and $p_u = (1 - C_\ell)/2$.

Now consider the goal of computing a confidence interval for θ_D. For notational convenience, let $\hat{p}(\mathbf{X}_1, \mathbf{X}_2)$ denote the estimate of p, where $\mathbf{X}_j$ (j = 1, 2) denotes the random sample from the j^{th} group, and let $D(\mathbf{X}_1, \mathbf{X}_2)$be the estimate of θ_D. Consider $D(\mathbf{X}_1 - \omega, \mathbf{X}_2)$ for some constant ω. Note that there is a one-to-one correspondence between $D(\mathbf{X}_1 - \omega, \mathbf{X}_2)$ and $\hat{p}(\mathbf{X}_1 - \omega, \mathbf{X}_2)$. When $\omega = 0$, $\hat{p}(\mathbf{X}_1, \mathbf{X}_2)$ corresponds to $D(\mathbf{X}_1, \mathbf{X}_2)$, the estimate of θ_D. More generally, if (p_ℓ, p_u) has probability coverage $1-\alpha$, the corresponding range of $D(\mathbf{X}_1 - \omega, \mathbf{X}_2)$ values contain θ_D with probability $1-\alpha$ as well. For some given constant q ($0 \le q \le 1$), consider the value of ω satisfying

$$\hat{p}\left(\mathbf{X}_1 - \omega, \mathbf{X}_2\right) = q \tag{12}$$

Let ω_ℓ denote the value of ω when $q = p_\ell$, the lower end of the $1-\alpha$ confidence interval for p. In a similar manner, letting ω_u denote the value of ω when $q = p_u$. Let $d_\ell = D(\mathbf{X}_1 - \omega_\ell, \mathbf{X}_2)$ and $d_u = D(\mathbf{X}_1 - \omega_u, \mathbf{X}_2)$. So, a $1-\alpha$ confidence interval for θ_D is

$$\left(d_l, d_u\right) \tag{13}$$

The value of ω, given q, is obtained simply by finding the value of ω that minimizes $|\hat{p}(\mathbf{X}_1 - \omega, \mathbf{X}_2) - q|$, which can be done via any one of several algorithms. (Here, this minimum value is obtained via the Nelder & Mead, 1965, algorithm.) So, the confidence interval given by (13) is readily computed.

Simulation results indicate that generally, (13) has reasonably accurate probability coverage. However, for an extreme shift in location, the method can result in $d_\ell = d_u$. When this occurs, it is assumed henceforth that the percentile bootstrap method, described below, is used instead.

The confidence interval for θ_D can be used to compute a confidence interval for Q. As previously noted, $\hat{Q}$ is the proportion of $Y_{ik} = D_{ik} - \hat{\theta}_D$ values that are less than or equal to $\hat{\theta}_D$. So, an approximate confidence interval for Q is (Q_ℓ, Q_u), where Q_ℓ is the proportion of Y_{ik} values less than or equal to $\hat{\omega}_l$, and Q_u is the proportion less than or equal to $\hat{\omega}_u$. This will be called method C1.

An alternative approach, stemming from (8), is to set $U_i = X_{i1} - 2\hat{\theta}_D$ ($i = 1,\ldots, n_1$), in which case a confidence interval for Q is given by applying Cliff's method based on the U_i values and X_{k2} ($k = 1,\ldots, n_2$). This will be called method C2.

Another way of computing a confidence interval for Q is via a percentile bootstrap method. Generate a bootstrap sample from the j^{th} group by randomly sampling with replacement n_j values from X_{ij} yielding X_{ij}^* ($i = 1,\ldots, n_j$). Compute Q based on this bootstrap sample and label the result Q^*. Repeat this process B times yielding Q_1^*, K, Q_B^* and let $Q_{(1)}^* \leq \mathrm{L} \leq Q_{(B)}^*$ denote the Q^* values written in descending order. Let $\ell = \alpha B/2$ and $u = B - \ell$. Then an approximate $1 - \alpha$ confidence interval for Q is $\left(Q_{(\ell+1)}^*, Q_{(u)}^*\right)$. Here $B = 500$ is used, which has been found to be satisfactory, in terms of achieving reasonably accurate confidence intervals, for a wide range of other robust methods that have been derived (Wilcox, 2017a).

Table 1. Some properties of the *g*-and-*h* distribution

g	*h*	κ_1	κ_2
0.00	0.00	0.00	3.00
0.00	0.20	0.00	21.46
0.20	0.00	0.61	3.68
0.20	0.20	2.81	155.98

A percentile bootstrap method can be used to compute a confidence interval for θ_D as well. Simply proceed as just described. The only difference is that estimates of θ_D are used rather than estimates of Q.

Simulation Results

Simulations were used to check the small sample properties of the confidence intervals for Q. Preliminary results indicated situations where method C2 did not perform well. So, for brevity, these results are not reported.

Two sample sizes are considered: 10 and 40. Data were generated from four distributions: normal, symmetric and heavy-tailed, skewed and light-tailed, and skewed and heavy-tailed. More precisely, data were generated from *g*-and-*h* distributions (Hoaglin, 1985), which arise as follows. Let Z be a random variable having a standard normal distribution. Then

$$V = \frac{\exp(gZ)-1}{g}\exp\left(hZ^2/2\right) \quad \text{if } g > 0$$
$$V = Z\exp\left(hZ^2/2\right) \quad \text{if } g = 0$$

has a *g*-and-*h* distribution, where g and h are parameters that determine the first four moments. The four distributions used here are the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0$), an asymmetric distribution with relatively light tails ($h = 0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 1 summarizes the skewness (κ_1) and kurtosis (κ_2) of these distributions.

Table 2 reports estimates of the actual value of α, when computing a $1 - \alpha = 0.95$ confidence interval, based on 4000 replications when $\theta_1 - \theta_2 = \xi$. The results in Table 2 are based on two choices for ξ: 0 and 0.8. Although the importance of a Type I error depends on the situation, Bradley (1978) suggests that as a general guide, when computing a 0.95 confidence interval, the actual value of α should be between 0.025 and 0.075. Both methods satisfy this criterion for all of the situations considered. So, the percentile bootstrap does not dominate method C1, but for very small sample sizes, it might be argued that the percentile bootstrap method is preferable. When both sample sizes are equal to 40, there appears to be little separating the two methods.

Table 2. Estimates of the actual value of α when computing a 0.95 confidence interval for Q

g	h	ξ	Method	n = (10, 10)	n = (10, 40)	n = (40, 40)
0.0	0.0	0.0	PB	0.053	0.061	0.058
0.0	0.0	0.0	C1	0.058	0.064	0.052
0.0	0.0	0.8	PB	0.028	0.049	0.040
0.0	0.0	0.8	C1	0.061	0.066	0.051
0.0	0.2	0.0	PB	0.056	0.067	0.050
0.0	0.2	0.0	C1	0.044	0.047	0.042
0.0	0.2	0.8	PB	0.030	0.048	0.044
0.0	0.2	0.8	C1	0.045	0.057	0.042
0.2	0.0	0.0	PB	0.056	0.067	0.050
0.2	0.0	0.0	C1	0.062	0.065	0.047
0.2	0.0	0.8	PB	0.031	0.050	0.043
0.2	0.0	0.8	C1	0.061	0.073	0.051
0.2	0.2	0.0	PB	0.056	0.068	0.050
0.2	0.2	0.0	C1	0.040	0.049	0.040
0.2	0.2	0.8	PB	0.030	0.052	0.044
0.2	0.2	0.8	C1	0.042	0.063	0.035

Table 3. Estimates of the actual value of α when computing a 0.95 confidence interval for θ_D

g	h	ξ	Method	n = (10, 10)	n = (10, 40)	n = (40, 40)
0.0	0.0	0.0	CM	0.065	0.064	0.054
0.0	0.0	0.0	PM	0.060	0.068	0.056
0.0	0.0	0.8	CM	0.061	0.070	0.054
0.0	0.0	0.8	PM	0.060	0.071	0.059
0.0	0.2	0.0	CM	0.062	0.067	0.049
0.0	0.2	0.0	PM	0.061	0.074	0.052
0.0	0.2	0.8	CM	0.069	0.069	0.038
0.0	0.2	0.8	PM	0.064	0.075	0.057
0.2	0.0	0.0	CM	0.064	0.072	0.058
0.2	0.0	0.0	PM	0.058	0.072	0.060
0.2	0.0	0.8	CM	0.059	0.070	0.050
0.2	0.0	0.8	PM	0.060	0.073	0.062
0.2	0.2	0.0	CM	0.071	0.073	0.043
0.2	0.2	0.0	PM	0.064	0.072	0.047
0.2	0.2	0.8	CM	0.068	0.065	0.041
0.2	0.2	0.8	PM	0.062	0.069	0.060

Table 3 reports the results when computing a 0.95 confidence interval for θ_D, where CM indicates the confidence interval based on (13) and PM is the percentile bootstrap method. Note that in general, there is little separating method CM from PM. For $n_1 = n_2 = 40$, CM provides a slight advantage.

A few simulations were run comparing the power of methods PB and C1. All indications are that there is little separating the two methods. Consider, for example, $\xi = 1.2$ and $n_1 = n_2 = 10$. Under normality, power, using PB, was estimated to be 0.736 versus 0.742 using C1. For $(g, h) = (0.2, 0.0)$, $(0.0, 0.2)$, and $(0.2, 0.2)$ the power for method PB was estimated to be 0.726, 0.588, and 0.602, respectively, versus 0.736, 0.592, and 0.602 using C1. As for testing H_0: $\theta_D = 0$, there is little difference between CM and PM. The length of the confidence intervals using a bootstrap method versus a non-bootstrap method can differ, but neither approach always has the shorter length. Similar results were obtained with $\xi = 0.8$.

Two Illustrations

The first illustration is based on data reported by Dana (1990) where the goal was to investigate issues related to self-awareness and self-evaluation. One segment of his study measured the time subjects could keep a portion of an apparatus in contact with a specified target. Cohen's d is −0.23 and the estimate of Q is 0.35. So, based on a commonly-used perspective, Q suggests a medium effect size in contrast to Cohen's d. The 0.95 confidence interval for Q based on the percentile bootstrap method is (0.136, 0.540), and it is (0.188, 0.548) using C1.

The second illustration is based on a study dealing with mild traumatic brain injury (Almeida-Suhett et al., 2014). Briefly, 5-6-week-old male Sprague-Dawley rats received a mild controlled cortical impact (CCI) injury. The dependent variable used here is the stereologically estimated total number of GAD-67-positive cells in the basolateral amygdala (BLA). One group was measured seven days after surgery and was compared to the sham-treated control group that received a craniotomy but no CCI injury. The results based on the contralateral side of the BLA are reanalyzed here.

Using Cliff's method to compare the control group to the Day 7 group, the estimate of $P(X_1 < X_2)$ is $p = 0.24$. The estimate of θ_D is 1220.35. The 0.95 confidence interval for θ_D, using method CM, is (800.38, 1794.90). Using the percentile bootstrap method, the 0.95 confidence interval is (849.88, 1794.90). Using method C1, the 0.95 confidence interval for Q is (0.803, 0.991). Using method PB, it is (0.761, 1.000).

Concluding Remarks

It is not being suggested that in some sense Q dominates all other measures of effect size. Certainly p, for example, given by (3), is a useful and important measure of

effect size as argued by, among others, Cliff (1996), Ruscio (2008), and Newcombe (2006). The suggestion is that Q is just one of several measures of effect size that can help provide a more nuanced understanding of data. To the extent the probabilistic interpretation of Cohen's d is deemed useful, Q provides a generalization that helps deal with both non-normality and heteroscedasticity. But measures of location, such as θ_D, also provide a potentially useful characterization of the extent groups differ.

For both illustrations, method C1 yielded shorter confidence intervals than method PB. It is not being suggested, however, that this is always the case. The simulations indicate that the reverse can happen and that neither method dominates in terms of achieving the shortest confidence interval.

Finally, R functions for applying the bootstrap methods described in this paper are being added to the R package WRS. The percentile bootstrap method for computing a confidence interval for Q is performed by the R function shiftPBci, and the bootstrap method for θ_D is performed by the R function wmwpb. The R function QS2ci performs method C1 and loc2dif.ci performs method CM.

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*(3), 317-328. doi: 10.1037/1082-989x.10.3.317

Almeida-Suhett, C. P., Prager, E. M., Pidoplichko, V., Figueiredo, T. H., Marini, A. M., Li, Z., … Braga, M. (2014). Reduced GABAergic inhibition in the basolateral amygdala and the development of anxiety-like behaviors after mild traumatic brain injury. *PLoS ONE, 9*(7), e102627. doi: 10.1371/journal.pone.0102627

Bhattacharyya, M. R., Molloy, G. J., & Steptoe, A. (2008). Depression is associated with flatter cortisol rhythms in patients with coronary artery disease. *Journal of Psychosomatic Research, 65*(2), 107-113. doi: 10.1016/j.jpsychores.2008.03.012

Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh, M., … Azen, S. P. (2012). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology & Community Health, 66*(9), 782-790. doi: 10.1136/jech.2009.099754

Cliff, N. (1996). Ordinal methods for behavioral data analysis. Mahwah, NJ: Erlbaum. doi: 10.4324/9781315806730

Dana, E. (1990). *Salience of the self and salience of standards: Attempts to match self to standard* (Unpublished doctoral dissertation). University of Southern California, Los Angeles, CA.

Gleason, J. R. (1993). Understanding elongation: The scale contaminated normal family. *Journal of the American Statistical Association, 88*(421), 327-337. doi: 10.1080/01621459.1993.10594325

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research*. New York, NY: Routledge. doi: 10.4324/9781410612915

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal, 7*(4), 308-313. doi: 10.1093/comjnl/7.4.308

Neuhäuser, M., Lösch, C., & Jöckel, K.-H. (2007). The Chen-Luo test in case of heteroscedasticity. *Computational Statistics & Data Analysis, 51*(10), 5055-5060. doi: 10.1016/j.csda.2006.04.025

Newcombe, R. G. (2006) Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine, 25*(4), 543-557. doi: 10.1002/sim.2323

Pruessner, J. C., Hellhammer, D. H., & Kirschbaum, C. (1999). Burnout, perceived stress, and cortisol responses to awakening. *Psychosomatic Medicine, 61*(2), 197-204. doi: 10.1097/00006842-199903000-00012

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*(1), 19-30. doi: 10.1037/1082-989x.13.1.19

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*(2), 201-223. doi: 10.1080/00273171.2012.658329

Wilcox, R. R. (2017a). *Introduction to robust estimation and hypothesis testing* (4th ed.). San Diego, CA: Academic Press.

Wilcox, R. (2017b). *Modern statistics for the social and behavioral sciences: A practical introduction* (2nd ed.). New York, NY: Chapman & Hall/CRC. doi: 10.1201/9781315154480

Wilcox, R. R. (2017c). *Understanding and applying basic statistical methods using R*. New York, NY: Wiley.