Graphical Methods for Assessing Effect Size: Some Alternatives to Cohen’s d

RAND R. WILCOX
University of Southern California

ABSTRACT. Reporting effect size plays an integral role in educational and psychological research and is required by many journals. Certainly, the best-known measure of effect size is Cohen’s \( d \), which represents a substantial improvement over using \( p \) values. But Cohen’s \( d \) is known to suffer from some fundamental concerns. The author’s goal was to illustrate some graphical methods aimed at addressing those concerns. These methods can be applied in a wide range of situations, including situations in which the Wilcoxon–Mann-Whitney test is used.

Key words: density estimators, effect size, shift function, theories of human mating, Wilcoxon–Mann-Whitney test

A FUNDAMENTAL GOAL IN APPLIED RESEARCH is not just detecting a difference between two groups of participants, but rather understanding how groups differ and by how much. Cohen (1977) formulated a very general and very reasonable approach when addressing this issue. His strategy began from a graphical perspective. In particular, when looking at the distributions corresponding to the two groups, Cohen suggested that there is a large effect if the difference between the two distributions (roughly referring to a shift in location) is “visible to the naked eye” (p. 40). Cohen went on to conclude that for normal distributions having a common variance, his familiar standardized difference, \( d \), is small if it equals .2 and large if it equals .8. Today, however, a known concern about \( d \) is that it can be small when in fact, based on Cohen’s graphical perspective, the effect size is large. My goal in this article is to illustrate a graphical method aimed at correcting this problem, plus other graphical methods that might be used to add perspective. None of the graphical methods used in this article are new, but it seems fair to say
that they are not well known by most applied researchers. All of the methods are now easy to apply with existing software; therefore, the illustrations reported here may promote their use. Yet another goal is to comment on measuring effect size when using the Wilcoxon–Mann-Whitney test.

Cohen’s $d$

To begin, it helps to elaborate on why Cohen’s $d$ might erroneously suggest a small effect size when in fact the effect size is large. The classic illustration is based on what is called a contaminated normal distribution that arises as follows: Imagine two populations of participants, one having a standard normal distribution and the other having a normal distribution with mean 0 and standard deviation 10. If these two distributions are mixed, we get a contaminated normal distribution. Suppose that when sampling from the contaminated normal, there is a .9 probability that a participant comes from the first group having the standard normal distribution. Figure 1 shows a standard normal distribution and the contaminated normal just described. As is evident, there is little visible difference. But, although the standard normal has variance $\sigma^2 = 1$, the contaminat-

![FIGURE 1. Plot of the standard normal and a contaminated normal distribution. Despite the similarity, the standard normal has variance 1, but the contaminated normal distribution has variance 10.9.](image-url)
ed normal has variance $\sigma^2 = 10.9$. This illustrates the classic result that the variance is highly sensitive to the tails of a distribution, a result that became evident with the publication of the seminal article by Tukey (1960).

Now imagine two groups, both having normal distributions with $\sigma^2 = 1$; the first has mean 0 and the second has mean .8. The left panel of Figure 2 shows the distributions, $d = .8$, and, according to Cohen, this is a large effect size. Now look at the right panel of Figure 2. There is an obvious similarity with the left panel; according to Cohen’s graphical perspective, there is a large effect size, but the distributions are not normal; they are contaminated normals, and $d = .24$, suggesting what would ordinarily be interpreted as a small effect size. Therefore, Cohen’s $d$ can mask a large effect size, and, as will be illustrated later, it can exaggerate an effect size as well. The basic problem here is the sensitivity of $\sigma^2$ to the tails of the distributions. When working with actual data, this sensitivity to the tails of a distribution is related to outliers. Another issue to be discussed is the lack of robustness associated with the mean.

Put another way, the crucial issue is not whether contaminated normal distributions frequently occur in applied work. Rather, the point is that the contaminated normal illustrates a general principle: Arbitrarily small changes in the tail of a distribution can have a very large effect on the variance, which can in turn substantially alter Cohen’s $d$, resulting in large effects, from a graphical perspective, to be missed. The issue is whether outliers are common, and, in many situations within the social sciences, all indications are that the answer is yes (e.g., Wilcox, 2003; Wu, 2002).

FIGURE 2. In the left panel, Cohen’s $d = .8$. In the right panel, $d = .24$. 
An Alternative Approach via Kernel Density Estimators

How might we capture the spirit of Cohen’s graphical view of effect size without assuming normality? One possibility is to use modern methods for estimating the distributions associated with both groups and simply graph them.

There is a large collection of methods for accomplishing this goal. But before discussing them, it might help to put the problem in more familiar terms by momentarily assuming groups have a normal distribution. This means that a plot of the data is given by (the probability density function)

\[
f(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),
\]

where \(\mu\) and \(\sigma^2\) are the population mean and variance, respectively. An estimate of the distribution, given by \(f(x)\), is obtained simply by replacing \(\mu\) and \(\sigma^2\) with the usual sample mean and variance. But this approach is highly unsatisfactory because often distributions differ substantially from normality. What is needed are more flexible approaches to estimating \(f(x)\), and such an approach is provided by what are called kernel density estimators. Standard commercial software contains certain variations, but these methods can give absurd results when, for example, there are ceiling or floor effects (e.g., Silverman, 1986; Wilcox, 2004). So, for example, if values less than zero are impossible and values close to zero are relatively common, then certain kernel density estimators can suggest that values less than zero have a fairly high probability of occurring. There are, however, variations of kernel density estimators that virtually eliminate this problem. In particular, use an adaptive kernel density estimator (Silverman, 1986) in conjunction with an initial fit that avoids problems with a restriction in range. The initial fit used here is called an expected frequency curve, after which the adaptive kernel density estimator is applied. (Computational details can be found in Wilcox, 2004, and easy-to-use software is provided in Wilcox, 2005.)

Sex Differences in the Desire for Sexual Variety

Recently, Schmitt (2003) analyzed data dealing with the desired number of future sexual partners and concluded that men and women differ. Here, a portion of his data is used to illustrate the kernel density estimator plus other graphical methods to be described. It is stressed that nothing in this article remotely contradicts his conclusion that the groups differ. But given that the groups differ, certainly it is important to gain as deep an understanding as possible about how they differ and by how much. I argue that the differences between the men and women in Schmitt’s study are large or small depending on where we look and how we measure effect size. The main point is that by taking a global look at how the entire distributions differ, new and useful perspectives are achieved.
The focus here is on Schmitt’s (2003) North American data dealing with two issues. The first has to do with the reported number of desired sexual partners during the next month. Schmitt provides ample evidence that responses by men differ from those of women. Figure 3 shows the kernel density estimates for men (given by the solid line and based on a sample size of 1,316) and women (the dashed line, based on a sample size of 2,282). (This graph was created with the S-PLUS function g2plot in Wilcox, 2005. A version using the free software R is available.) Cohen’s $d$ is .4, suggesting a moderate effect size. The difference between the means is 1.1, and this difference, divided by the standard error for women, is 1.44, suggesting a very large difference relative to the variation among women. But for men, this ratio is only .24, suggesting a relatively small effect. Therefore, we get different perspectives depending on which variation of Cohen’s $d$ we use. But as is evident from Figure 3, there is a sense in which there is a small difference as originally conceived by Cohen: The central portions of the estimated distributions are very similar. (Both groups have a median of 1.) However, the right tails differ noticeably, roughly meaning that men tend to give higher, more extreme responses. For example, 23% of men gave a response greater than one, versus only 3% for women. That is, there is a large effect size among the more extreme responses. This will be discussed in more detail later in this article.

**FIGURE 3.** Kernel density estimates for men (the solid line) versus women (the dashed line) for the 1-month data.
It should be remarked that Schmitt (2003) generally eliminated any response greater than 100, but in Figure 3, all of the data were used. This is a minor issue for the case at hand because there is only one man who gave a response of 100, and the largest response among the women was 20.

Next, consideration is given to the second issue: How do men and women differ when asked how many sexual partners they desire over the next 30 years? Following Schmitt, (2003) any response greater than 100 is eliminated. Now Cohen’s $d$ is .58, suggesting a moderately large effect size. Figure 4 shows the estimated distributions for the two groups. Again, we see that for the central portions of the distributions, there is little difference, but now there is a bigger difference in the tails. For example, now 46% of men give a response greater than one versus 27% for women.

**Possible Reasons for Comparing Quantiles**

Before continuing, some general remarks about comparing quantiles might help. A well-known concern about comparing groups with a single measure of location, such as the mean or median, is that certain details are lost about how the groups compare. How do high-scoring participants in the first group compare with high-scoring participants in the second? How do low-scoring participants...
compare? For the data in Figures 3 and 4, there is evidence that men differ from women in terms of the higher responses, but a possible criticism is that the details of how they differ, and by how much, are not made clear. What is needed is a method that compares the quantiles.

As another simple example of why one might want to compare quantiles, imagine that two groups have normal distributions, equal means, but unequal variances. If the second group has the larger variance, then high-scoring participants in the second group tend to score higher than high-scoring participants in the first. In other words, the upper quantiles associated with the second group are larger than the corresponding upper quantiles of the first. For low-scoring individuals, the reverse is true. Therefore, the groups differ, and if high scores are better, participants who tend to score high are better off being in the second group. But participants who tend to score low are better off being in the first group. (Lunneborg, 1986, provides additional reasons for comparing quantiles.) In terms of effect size, when we take into account the difference between quantiles, we get a more detailed and intricate understanding of how groups differ and by how much.

Returning to Schmitt’s (2003) data, consider the 1-month data. Let $q_1$ and $q_2$ be the lower quartile and upper quartiles, respectively. The estimated lower and upper quartiles for men are $\hat{q}_1 = \hat{q}_2 = 1$, and the median is $Mdn = 1$ as well. As for the women, the estimates are $\hat{q}_1 = 0$ and $\hat{q}_2 = Mdn = 1$. Because the sample medians are equal to 1 for both groups, it is evident that any reasonable method for comparing medians will fail to reject. The groups differ, as previously indicated, but the reason for rejecting lies elsewhere. Also, the most common response among both men and women is 1, which helps explain why the distributions in Figure 3 appear to be very similar near the central portion of the distributions.

As for the 30-year data, for men we have that $\hat{q}_1 = Mdn = 1$ and $\hat{q}_2 = 7$, and for women, $\hat{q}_1 = Mdn = 1$ and $\hat{q}_2 = 2$. Again, any reasonable method for comparing medians will not reject because the sample medians are equal to 1 for both groups.

**Comparisons Based on the Shift Function**

One way of getting a detailed sense of how the groups differ, without assuming normality, is via the so-called *shift function* derived by Doksum and Sievers (1976). Note that in addition to comparing the .5 quantiles (the medians), we could compare the .25 and .75 quantiles. More generally, all quantiles can be compared, an advantage of which is that we get a sense of how the tails of the distributions differ as well as the central portions. The method is based on an extension of the Kolmogorov–Smirnov test and is designed so that, under random sampling only, the probability of at least one Type I error, when comparing all quantiles, can be determined exactly. (An outline of the method is given in Wilcox, 2003, and it is performed by the S-PLUS or R function sband.
described in Chapter 8 of his book. For complete computational details, see Wilcox, 2005.)

First, consider the 1-month data. Testing at the .05 level, it is found that the quantiles between .22 and .38, as well as between .77 and .95, differ. The confidence intervals generally indicate that the difference between the quantiles is less than or equal to 1. But from about the .87 to .92 quantiles, the difference could be as high as 2. Therefore, not only do the groups differ in terms of high responses, but they differ in terms of low responses as well. For the 30-year data, the groups differ from the .54 to the .95 quantiles. As we move into the extreme right portion of the tails, the estimated difference increases. From the .54 to the .58 quantile, the difference is estimated to be 1. From the .59 to the .63 quantile, the difference is estimated to be 2. At the .75 quantile, the confidence interval indicates that the men have a response that is higher than the women to be between 8 and 12.

To add perspective, plots can help. Figure 5 plots the difference between the deciles for the 1-month data. The y-axis, marked delta, is the estimated difference between the second group and the first. That is, the y-axis reflects an estimate of $y_q - x_q$, where $y_q$ is the $q$th quantile of the second group, $q = .1, \ldots, .9$. (This plot was created with the S-PLUS function shifthd in Wilcox, 2005, with the argu-

![FIGURE 5. Plot of the difference between the deciles for the 1-month data. + = ends of the confidence intervals for the difference between the deciles. * = estimate of $y_q - x_q$.](image-url)
ment plotop set equal to T. A version of this function that runs under R is available as well.) Ends of the confidence intervals for the difference between the deciles are marked by a +. So in Figure 5, the difference between the .4 quantiles is approximately 1, and the .95 confidence interval extends from about –0.4 to –1.3. In a similar manner, significant differences are obtained at the .8 and .9 quantiles, with the largest difference occurring at the .9 quantile. A plot for the 30-year data is shown in Figure 6. Now the lower deciles are virtually identical, but as we move toward the higher quantiles, the differences increase.

Doksum and Sievers (1976) suggested plotting the quantiles of the first group versus the difference between the quantiles. Often, this provides a good overall sense of how the groups compare, but here it primarily reflects the difference between the extreme quantiles. Figure 7 shows this plot for the 1-month data (which was created with the S-PLUS function sband in Wilcox, 2003). The point that should be stressed is that the bulk of this plot reflects the extreme tails of the distributions. The x-axis indicates the response given by the men, but the proportion of men giving a response ≥ 5 was only .052. The y-axis is the estimated difference between the quantiles (delta = y_q – x_q, where x_q is the qth quantile for men). The dashed lines indicate the (simultaneous) confidence band for the differences between all quantiles. Note that the upper dashed line extends along the

![Image](image_url)

**FIGURE 6.** Plot of the difference between the deciles for the 30-year data. + = ends of the confidence intervals for the difference between the deciles. • = estimate of y_q – x_q.
The shift function for the 1-month data. The $x$-axis indicates the responses given by men. The $y$-axis is the estimated difference between the corresponding quantile, which is indicated by the solid line. If, for example, 5 is the .95 quantile among men, the plot indicates the difference between 5 and the .95 quantile for women.
the control group gain about 12 grams. As we move along the $x$-axis, meaning that rats gain more weight, the effect decreases. Output from the shift function indicates that differences occur between the .09 and .39 quantiles. (Again, the dashed lines are confidence bands. When the upper dashed line drops below zero, or the lower dashed line is above zero, this indicates a significant difference. The + marks the median for the first group and the ¶s indicate the lower and upper quartiles.)

**Comments on the Wilcoxon–Mann-Whitney Test**

Schmitt (2003) reported results when using the Wilcoxon–Mann-Whitney test for comparing groups and concluded that groups differ. When this test rejects, it is certainly reasonable to conclude that distributions differ, as done by Schmitt, but some additional information can be gained using modern extensions of this test.

Before continuing, note that the Wilcoxon–Mann-Whitney test is sometimes described as a method for comparing the medians of two groups, but it is known that it is unsatisfactory for this purpose (e.g., Hettmansperger, 1984), the point being that a failure to find differences between the medians does not contradict a significant result when applying the Wilcoxon–Mann-Whitney test, as reported by Schmitt (2003).
The Wilcoxon–Mann-Whitney test is based on an estimate of a measure of effect size that reflects the probability that a randomly sampled observation from the first group is larger than a randomly sampled observation from the second. More formally, for two independent variables, say $X$ and $Y$, let

$$p_1 = P(X > Y),$$
$$p_2 = P(X = Y),$$

and

$$p_3 = (X < Y).$$

If tied values never occur, then the Wilcoxon–Mann-Whitney test is based on a direct estimate of $p_1$. (The usual $U$ statistic, divided by the product of the sample sizes, estimates $p_1$.) If ties occur, it is based on a direct estimate of $Q = p_1 + .5p_2$. The Wilcoxon–Mann-Whitney test is a satisfactory approach to testing the hypothesis that distributions are identical, but it is known that in terms of making inferences about the measure of effect size $p_1$ or $Q$, the Wilcoxon–Mann-Whitney test is unsatisfactory (e.g., Cliff, 1996; Wilcox, 2003). The reason is that when distributions are identical, a correct estimate of the standard error of the estimate of $p_1$ (or $Q$) is used, but otherwise this is not the case. The result is that, under general conditions, confidence intervals for $p_1$ (or $Q$) violate basic principles and can yield inaccurate results even with large sample sizes. Methods that correct this problem have been devised (e.g., Brunner, Domhof, & Langer, 2002; Brunner & Munzel, 2000; Cliff).

Cliff (1996) argued that $\delta = p_1 - p_3$ should be used as a measure of effect size. It is easy to see that $\delta = 1 - 2Q$. Using Cliff’s method for computing a confidence interval for $\delta$ (which was applied with the S-PLUS function cid in Wilcox, 2003), the estimate of $\delta$ for the 1-month data is $\hat{\delta} = .32$ and the .95 confidence interval is (.28, .35). This seems to suggest that there is a moderate difference between the groups. (Under normality, $\hat{\delta} = .32$ corresponds, approximately, to $d = .6$.) As for the 30-year data, $\hat{\delta} = .24$, with a .95 confidence interval equal to (.21, .28). The estimates of $p_1$, $p_2$, and $p_3$ are .46, .40, and .14, respectively. Therefore, the probability that a randomly sampled man and woman will give the same response is estimated to be .4, and the probability that a man gives a higher response is estimated to be .46. Applying the Brunner–Munzel method to the 1-month data, $\hat{Q} = .34$, with a .95 confidence interval equal to (.325, .358), and for the 30-year data, $\hat{Q} = .38$, with a .95 confidence interval equal to (.358, .396).

There are graphical methods, beyond those already covered, that help add perspective and assess effect size when using a Wilcoxon–Mann-Whitney test or any of its modern extensions. To explain, again consider any two independent variables, $X$ and $Y$, let $\theta_1$ be the population median of $X$, let $\theta_2$ be the population median of $Y$, and let $\theta_d$ be the population median associated with $d = X - Y$. In words, imagine that we randomly sample an individual from the first group,
yielding the value $X$; we do the same for the second group, yielding $Y$, and we note the difference, $X - Y$. If we repeat this process millions of times (and in theory infinitely many times), the median of the resulting differences is $\theta_d$.

Although the difference between the means is equal to the mean of $d$, there are general conditions where $\theta_d \neq \theta_1 - \theta_2$. As previously noted, the Wilcoxon–Mann-Whitney test and its modern extensions do not test $H_0$: $\theta_1 = \theta_2$, the hypothesis that the groups have equal medians. But if there are no tied values, they do test $H_0$: $\theta_d = 0$, the hypothesis that the difference scores have a median of zero. The main point here is that a plot reflecting the distribution of $d$ might help researchers judge effect size. That is, if the samples are $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, set $D_{ij} = X_i - Y_j$ and plot the $D_{ij}$ values. The plot provides perspective on how far zero is from the median of the $D_{ij}$ values, and a graphical depiction of the dispersion among the $D_{ij}$ values might help as well. In particular, if the groups do not differ in any manner, a plot of $D_{ij}$ should be approximately symmetric about zero.

Figure 9 shows a boxplot of the $D_{ij}$ values for the 1-month data. The median of the $D_{ij}$ values is 0, suggesting a small effect size, but the asymmetric nature of the plot indicates a large effect size in the sense that for a randomly sampled man versus woman, relatively large differences can occur.
More Comments Regarding Cohen’s $d$

Another feature of Cohen’s $d$ that should be kept in mind is its reliance on the sample mean that is not robust in the general sense described by Hampel, Ronchetti, Rousseeuw, and Stahel (1986), Huber (1981), Staudte and Sheather (1990), and Wilcox (2005). In essence, even a single outlier can render the sample mean a poor reflection of what is typical, which in turn might color any perceptions of effect size based on $d$. For the 1-month data, the sample mean for the men is 1.73, which is approximately equal to the .77 quantile. For women, the mean is .63, which corresponds, approximately, to the .42 quantile. This illustrates how Cohen’s $d$ might be relatively large even when the central portions of a distribution are fairly similar. For the 30-year data, the means are 6.5 and 2.04, which correspond, approximately, to the .75 and .79 quantiles. That is, in this case, Cohen’s $d$ is based in part on comparing, approximately, the upper quartiles. This is not to suggest that means be abandoned, but only that additional summaries of data can be important and useful.

Concluding Remarks

Certainly, the most common approach to comparing groups is to use the difference between some measure of location, such as the mean or median. This difference adds perspective, and I do not suggest that this approach be abandoned. Rather, the goal was to illustrate that various methods developed in recent years can help provide perspective on effect size and how groups differ. Very easy-to-use software is available for applying all of the illustrated methods, and additional functions are available that were not described here (Wilcox, 2003, 2005). For example, if S-PLUS or the free software R is being used, and the data for a single group are stored in the variable $x$, the command `akerd(x)` will produce a plot of the data based on the adaptive kernel density estimator. Distributions for two groups can be plotted with the function `g2plot`. When applying one of the modern extensions of the Wilcoxon–Mann-Whitney test, with the data for the second group stored in the variable $y$, the command `cid(x,y,plotit = T)` will produce an (adaptive kernel density) estimate of the distribution of the difference scores that helps assess effect size. The command `cid(x,y,plotit = T,pop = 3)` will create a boxplot instead.

REFERENCES


