Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Measuring effect size: a robust heteroscedastic approach for two or more groups

Rand R. Wilcox [a] & Tian S. Tian [b]

[a] Department of Psychology , University of Southern California , California, USA

[b] Department of Psychology , University of Houston , Houston, TX, USA

Published online: 30 Mar 2011.

To cite this article: Rand R. Wilcox & Tian S. Tian (2011) Measuring effect size: a robust heteroscedastic approach for two or more groups, Journal of Applied Statistics, 38:7, 1359-1368, DOI: 10.1080/02664763.2010.498507

To link to this article: http://dx.doi.org/10.1080/02664763.2010.498507

Journal of Applied Statistics
Vol. 38, No. 7, July 2011, 1359–1368

Taylor & Francis
Taylor & Francis Group

Measuring effect size: a robust heteroscedastic approach for two or more groups

Rand R. Wilcox[a]* and Tian S. Tian[b]

[a] *Department of Psychology, University of Southern California, California, USA;* [b] *Department of Psychology, University of Houston, Houston, TX, USA*

(*Received 5 January 2010; final version received 17 May 2010*)

Motivated by involvement in an intervention study, the paper proposes a robust, heteroscedastic generalization of what is popularly known as Cohen's d. The approach has the additional advantage of being readily extended to situations where the goal is to compare more than two groups. The method arises quite naturally from a regression perspective in conjunction with a robust version of explanatory power. Moreover, it provides a single numeric summary of how the groups compare in contrast to other strategies aimed at dealing with heteroscedasticity. Kulinskaya and Staudte [16] studied a heteroscedastic measure of effect size similar to the one proposed here, but their measure of effect size depends on the sample sizes making it difficult for applied researchers to interpret the results. The approach used here is based on a generalization of Cohen's d that obviates the issue of unequal sample sizes. Simulations and illustrations demonstrate that the new measure of effect size can make a practical difference regarding the conclusions reached.

Keywords: effect size; robust methods; outliers; heteroscedasticity explanatory power; Cohen's d

1. Introduction

When comparing two independent groups, there are, of course, several general approaches that might be used to characterize the extent to which the groups differ. One obvious approach is to use the difference between two measures of location. Using the difference between all of the quantiles can be done as described by Doksum and Sievers [6]. Some have advocated using the probability that a randomly sampled observation from the first group is smaller than a randomly sampled observation from second group (e.g. [1,3]). And of course graphical methods can be used, such as boxplots and kernel density estimators.

Yet another approach is to use the difference between measures of location relative to some measures of variation. Let μ_1 and μ_2 denote the means, and let σ_1 and σ_2 denote the standard

*Corresponding author. Email: rwilcox@usc.edu

ISSN 0266-4763 print/ISSN 1360-0532 online

DOI: 10.1080/02664763.2010.498507
http://www.informaworld.com

deviations of the two groups. The most commonly used measure, based on this last approach, is

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

where by assumption $\sigma_1 = \sigma_2 = \sigma$. (For a review of many other variations, see, for example, Algina *et al.* [2].)

The goal in this paper is to suggest a robust, heteroscedastic approach to measuring effect size that is based on measures of location and scatter. The new approach is readily extended to more than two groups and is motivated in part by at least two fundamental concerns regarding δ. First, it is not robust, where the term robust is being used in the general sense described by Huber [10], Hampel *et al.* [8] as well as Staudte and Sheather [18]. In particular, very small changes in the tails of the distributions (as measured, for example, by Kolmogorov distance) can substantially inflate the variance. This in turn can result in a large decrease in δ that can be highly misleading when effect size is viewed from a graphical perspective.

A classic example is based on the contaminated normal distribution

$$H(x) = 0.9\Phi(x) + 0.1\Phi\left(\frac{x}{10}\right),$$

where $\Phi(x)$ is the standard normal distribution. Of course, what constitutes a large difference between two groups can depend on the situation. Cohen [4, Section 2.2.3] suggests that as a general guide, $\delta = 0.2$, 0.5 and 0.8 correspond to small, medium and large effect sizes, respectively. The left panel of Figure 1 shows two normal distributions with $\delta = 1$. The right panel shows two contaminated normals with the same means, but now $\delta = 0.3$, despite the obvious similarity with the left panel. An additional concern is that when dealing with skewed distributions, the population mean μ can poorly reflect the typical response.

A related concern has to do with the usual estimate of δ: even a single outlier can result in d being relatively small when otherwise it would be large. This is a serious practical issue because modern outlier detection methods indicate that outliers are more the rule than the exception (e.g. [16,24]), a result predicted in a seminal paper by Tukey [20].

In terms of dealing with heavy-tailed distributions and outliers, Algina *et al.* [2] suggest a generalization of δ where the mean is replaced by a trimmed mean and the variance is replaced

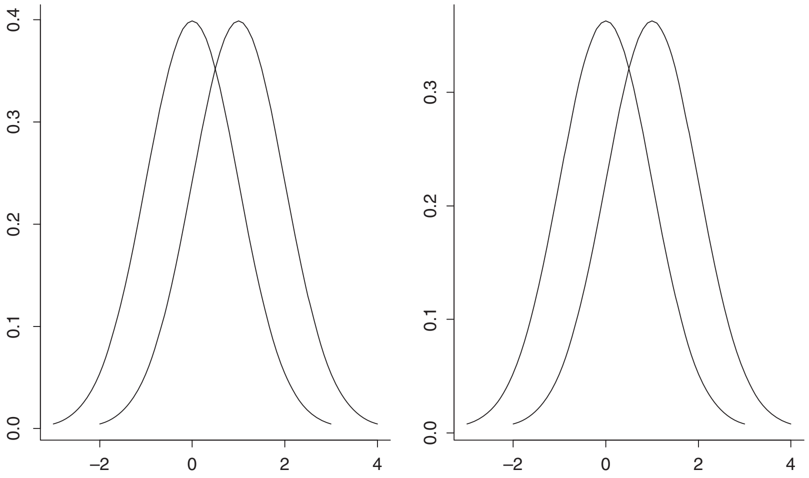

Figure 1. The measure of effect size δ is highly sensitive to the tails of a distribution, which can result in a small value for δ even when plots of the distributions indicate that there is a large difference between the groups. In the left panel, $\delta = 1$, but in the right panel $\delta = 0.3$.

by a Winsorized variance. They focus on a 20% trimmed mean and a 20% Winsorized variance that is rescaled so that it estimates the population variance under normality. There are known cases where, when reanalyzing data, robust measures of effect size indicate a large difference, while methods based on the means and standard deviations do not (e.g. [23]). That is, the choice between the mean and 20% mean can have a substantial impact on the interpretation of the data. It is not readily apparent, however, how their measure of effect size might be generalized to more than two groups.

The second general concern with δ is heteroscedasticity. A simple strategy is to use

$$\delta_1 = \frac{\mu_1 - \mu_2}{\sigma_1}$$

or

$$\delta_2 = \frac{\mu_1 - \mu_2}{\sigma_2},$$

or both, as suggested by Glass *et al.* [7]. And again a simple robust extension is to replace the means and variances with some robust analog such as trimmed means and Winsorized variances. But a possible concern is that δ_1 might indicate a large measure of effect size while δ_2 indicates the reverse. What would seem desirable is a single measure of effect size that assesses the overall difference regardless of how unequal the measures of variation might be.

Another well-known measure of effect size is ω^2 [9]. Letting Y be the random variable of interest, and assuming equal variances, Hays imagines that with probability 0.5 an observation is randomly sampled from the first group, otherwise an observation is sampled from the second group, and we predict the outcome for the jth group to be μ_j. Hays shows that

$$\sigma_Y^2 = \sigma^2 + \frac{(\mu_1 - \mu_2)^2}{4}$$

and his measure of effect size is

$$\omega^2 = \frac{(\mu_1 - \mu_2)^2}{4\sigma_Y^2}.$$

Kulinskaya and Staudte [11] suggest the measure of effect size

$$\psi = q_1 q_2 \frac{(\mu_2 - \mu_1)^2}{q_2\sigma_1^2 + q_1\sigma_2^2},$$

where $q_j = n_j/N$, n_j is the sample size associated with the jth group, and $N = n_1 + n_2$. A possible concern with ψ is that its population value depends on the q_j values. So if a study is replicated, but with different sample sizes, the effect size being estimated will differ. Kulinskaya and Staudte suggest that this is reasonable in the sense that for the same total sample sizes, the weighted effect accounts for the differing amount of information.

Currently, there are no results on robust, heteroscedastic measures of effect size when dealing with more than two groups. Hays [9] describes an extension of ω^2 to more than two groups, but it assumes equal variances and it is not robust. The general approach proposed here is readily extended to more than two groups. As will become evident, the proposed approach has certain similarities to the derivation of ω^2, but it differs in a crucial manner.

It is stressed that although the focus here is on measures of effect size that offer a robust, heteroscedastic alternative to δ, this is not to suggest that the other general approaches listed above have no practical value. Our view is that several different perspectives are useful when characterizing the extent to which groups differ.

2. The proposed approach to measuring effect size

The proposed approach to measuring effect size is based on the notion of explanatory power, which was studied by Doksum and Samarov [5]. It contains as a special case measures of the strength of an association that reflect the proportion of variance accounted for by a regression line and X. One advantage of this approach is that it is readily generalized to numerous robust analogs.

To elaborate, let $\sigma^2(Y)$ be the (population) variance of Y and let $\hat{Y}$ be some predicted value of Y, given X. A measure of the strength of the association is

$$\xi^2 = \frac{\sigma^2(\hat{Y})}{\sigma^2(Y)},$$

which is called explanatory power. Within the context of least squares regression, this approach is well known. To see this, note that if $\hat{Y} = \beta_0 + \beta_1 X$ is the population least squares regression line, then

$$\xi^2 = \rho^2,$$

where ρ is Pearson's correlation. (Because $\beta_1 = \rho\sigma(Y)/\sigma(X)$, $\sigma^2(\hat{Y}) = \rho^2\sigma^2(Y)$, and the result follows.) Kulinskaya and Staudte [11] mention a weighted version of this approach, but it was not studied.

Returning to the case where the goal is to compare two independent groups based on their means, Hays imagines that with probability 0.5, an observation is randomly sampled from the first group. From a regression perspective, X represents groups, and there are two possible values for $\hat{Y}$: μ_1 if the observation is from the first group, and μ_2 if it is from the second. For the special case where the groups have equal variances, it is readily verified that from Hays' perspective, ξ^2 becomes ω^2.

Note that from a regression perspective, if we let $\bar{\mu} = (\mu_1 + \mu_2)/2$,

$$\sigma_\mu^2 = \sum_{j=1}^{J} (\mu_j - \bar{\mu})^2,$$

and explanatory power becomes

$$\xi^2 = \frac{\sigma_\mu^2}{\sigma^2(Y)}, \tag{2}$$

which has an obvious similarity to ω^2. One important difference from ω^2 is that equal variances are not assumed. To see this, note that $\sigma^2(Y)$ represents the unconditional variance of Y. It is not being assumed that $\sigma^2(Y|X = 0) = \sigma^2(Y|X = 1)$, where $X = 0$ and 1 correspond to the first and second group, respectively. The key idea is that when comparing groups, it is convenient to view $\sigma^2(Y)$ in a slightly different manner than done by Hays when developing ω^2.

Let Y_{ij} ($i = 1, \ldots, n_j$; $j = 1, 2$) be a random sample of size n_j from the jth group. In the context of a regression perspective, rather than assume that observations are sampled with probability 0.5 from the first group, as done by Hays, momentarily consider the situation where with probability 1, equal sample sizes are used. From this perspective, $\sigma^2(Y)$ can be viewed as the population variance associated with the pooled Y_{ij} variables. Then a natural estimate of $\sigma^2(Y)$ is obtained simply by pooling the Y_{ij} values and computing the usual sample variance.

To add perspective, momentarily consider the case of equal variances and denote the common variance by σ^2. Focusing on the Y values, when $\mu_1 \neq \mu_2$, the conceptualization just described involves a sample that is not identically distributed. Half of the Y values have a different mean.

However, it is apparent that the population mean of the pooled observations is $\bar{\mu} = (\mu_1 + \mu_2)/2$. Moreover, the value of $E(Y_{ij} - \bar{\mu})^2$ is the same for both groups. To see this, note that

$$(Y_{i1} - \bar{\mu})^2 = \left(Y_{i1} - \mu_1 + \frac{\mu_1}{2} - \frac{\mu_2}{2}\right)^2$$

so for the first group $E(Y_{i1} - \bar{\mu})^2 = \sigma^2 + (\mu_1 - \mu_2)^2/4$. The same is true for the second group, so

$$\sigma^2(Y) = \sigma^2 + \left(\frac{\mu_1 - \mu_2}{2}\right)^2.$$

from which it follows that

$$\xi^2 = \frac{2(\mu_1 - \mu_2)^2}{4\sigma^2 + (\mu_1 - \mu_2)^2}.$$

Consider again Cohen's suggestion that under normality and homoscedasticity, $\delta = 0.2$, 0.5 and 0.8 roughly correspond to small, medium and large effect sizes, respectively. If this convention seems appropriate, it is noted that this corresponds approximately to $\xi = 0.15$, 0.35 and 0.50, respectively.

There remains the issue of how to proceed when $n_1 \neq n_2$ and the variances are possibly unequal. If we simply proceed as before and estimate $\sigma^2(Y)$ with the pooled Y values, then the expected value of the resulting estimate of ξ^2 can differ substantially from the equal sample size case. That is, the population variance $\sigma^2(Y)$ is defined as the estimand of the pooled Y_{ij} values when equal samples are used with probability 1. If $\sigma^2(Y)$ is estimated by pooling the Y_{ij} values when the sample sizes are unequal, we no longer get a satisfactory estimate of $\sigma^2(Y)$. In effect, we have a situation similar to the approach used by Kulinskaya and Staudte where the measure of effect size being estimated depends on the sample sizes. The proposed strategy for dealing with this problem is to define $\sigma^2(Y)$ as was done for the equal sample size case, and then use an appropriate estimate of $\sigma^2(Y)$ when dealing with unequal sample sizes. Kulinskaya and Staudte [11, p. 101] conclude that a natural generalization of Cohen's δ to the heteroscedastic case does not appear to be possible without taking into account the relative sample sizes.

The proposed strategy is as follows. Let $m = \min(n_1, n_2)$. A simple approach is to estimate $\sigma^2(Y)$ using the first m values from each group. But an obvious concern is that this ignores information in one of the groups. To deal with this problem, assume $m = n_1 < n_2$ and consider any subset of $\{1, 2, \ldots, n_2\}$ having cardinality m. Denote these m integers by $k_1, \ldots, k_m$. Then, of course, a reasonable estimate of $\sigma^2(Y)$ is obtained by computing the usual sample variance associated with the values $Y_{11}, \ldots, Y_{n_11}, Y_{k_12}, \ldots, Y_{k_m2}$. Let $\hat{\sigma}^2(Y)$ be the sample variance of these $2m$ values. To take advantage of all the information in group 2, one could repeat this process for all unique subsets of $\{1, 2, \ldots, n_2\}$ having cardinality m and average the results. But a practical issue is that the number of such subsets can be extremely large. Consequently, the strategy here is to take K randomly sampled subsets of size m yielding K estimates of $\sigma^2(Y)$: $\hat{\sigma}_1^2(Y), \ldots, \hat{\sigma}_K^2(Y)$. Then the final estimate of $\sigma^2(Y)$ is

$$\tilde{\sigma}^2(Y) = \frac{1}{K}\sum_{k=1}^{K} \hat{\sigma}_k^2(Y).$$

And the final estimate of ξ^2 is taken to be

$$\hat{\xi}^2 = \frac{\hat{\sigma}^2(\mu)}{\tilde{\sigma}^2(Y)}, \tag{3}$$

where

$$\hat{\sigma}^2(\mu) = \left(\frac{\bar{Y}_1 - \bar{Y}_2}{2}\right)^2,$$

and $\bar{Y}_j = \sum_{i=1}^{n_j} Y_{ij}/n_j$, $j = 1, 2$.

2.1 *Robust generalizations*

Robust analogs of ξ^2 are apparent: replace the mean with some robust measure of location θ and the variance with some robust measure of scatter, τ^2. So the general form of the measure of effect size becomes

$$\xi^2 = \frac{\sigma_\theta^2}{\tau^2(Y)}. \tag{4}$$

Consistent with Doksum and Samarov [5], this general version of ξ^2 will be called explanatory power. And henceforth, ξ is called an explanatory measure of effect size.

2.2 *Choosing a robust measure of location*

One obvious choice for a robust measure of location, θ, is the median. But a concern with the sample median is that it represents an extreme amount of trimming: all but one or two values are trimmed. Moreover, under normality, its standard error does not compare well with the sample mean or other robust location estimators that might be used. When sampling from a very heavy-tailed distribution, roughly meaning that for a random sample of size n, a large number of outliers is likely to occur, the efficiency of the sample median can be high relative to the mean. But otherwise, alternative location estimators can have better efficiency.

One possibility is to use a compromise amount of trimming. In the context of measuring effect size, this approach is not new and was studied by Algina *et al.* [1] who used a 20% trimmed mean. For a random sample $Y_1, \ldots, Y_n$, the 20% trimmed mean is

$$\bar{Y}_t = \frac{1}{n-2g}(Y_{(g+1)} + \cdots + Y_{(n-g)}),$$

where $Y_{(1)} \le \cdots \le Y_{(n)}$ are the Y values written in ascending order and g is $0.2n$ rounded down to the nearest integer.

From an efficiency point of view, 20% trimming has been found to be a relatively good choice [15]. That is, its standard error compares well with the standard error of the mean under normality. And when sampling from distributions where outliers tend to occur, the standard error of the 20% trimmed mean can be substantially smaller.

For completeness, another well-known robust measure of location is an M-estimator with Huber's Ψ (e.g. [8,10,18,23]). Like a 20% trimmed mean, it has good efficiency under normality, but it has the added advantage of being able to handle a larger proportion of outliers. More precisely, the breakdown point of a 20% trimmed is 0.2, roughly meaning that about 20% of the values must be altered to render the 20% trimmed mean arbitrarily large or small. In contrast, Huber's M-estimator has a breakdown point of 50%, the same as the median.

2.3 *Choosing a robust measure of variation*

There are many choices for a robust measure of variation, τ^2 (e.g. [24]). Algina *et al.* [2] use a 20% Winsorized variance, which has been rescaled to estimate the usual variance under normality. This rescaling is done so that under normality, Cohen's δ has the same value when the (population) trimmed mean and Winsorized variance are replaced by the mean and variance. The Winsorized variance is a natural choice in the sense that it plays a role when dealing with the standard error of a trimmed mean. And the 20% Winsorized variance provides good protection against the deleterious effects of outliers.

The sample 20% Winsorized variance is computed as follows. Let

$$W_i = \begin{cases} Y_{(g+1)} & \text{if } Y_i \le Y_{(g+1)}, \\ Y_i & \text{if } Y_{(g+1)} < Y_i < Y_{(n-g)}, \\ Y_{(n-g)} & \text{if } Y_i \ge Y_{(n-g)}. \end{cases}$$

The 20% Winsorized variance is

$$s_w^2 = \frac{1}{n-1} \sum_{i=1}^{n} (W_i - \bar{W})^2,$$

where $\bar{W} = \sum W_i / n$ is the Winsorized mean.

It is noted that for any amount of Winsorizing, say γ, the population Winsorized variance is readily rescaled so that it estimates the usual variance under normality. To elaborate, for any γ, $0 \le \gamma < 0.5$, the γ-Winsorized variance is

$$\sigma_w^2 = \int_{x_\gamma}^{x_{1-\gamma}} (x - \mu_w)^2 \, \mathrm{d}F(x) + \gamma[(x_\gamma - \mu_w)^2 + (x_{1-\gamma} - \mu_w)^2],$$

where x_γ is the γ quantile. So under normality, it is a simple matter to determine the population Winsorized variance under normality and rescale s_w^2 so that it estimates the variance. The rescaled estimate will be denoted by s_r^2. For 20% Winsorization, $s_r^2 = s_w^2/0.4121$.

Arguments can be made for considering other measures of dispersion. In terms of efficiency, Lax [12] concluded that two A-estimators perform relatively well, one of which corresponds to the percentage bend midvariance studied by Shoemaker and Hettmansperger [17]. The other A-estimator that performed well is called the biweight midvariance by Shoemaker and Hettmansperger. Randal [13] expanded on Lax's study and again concluded that the two A-estimators recommended by Lax perform relatively well. However, Randal's study did not include Rocke's [14] TBS (translated biweight S) estimator, and the tau measure of scale introduced by Yohai and Zamar [25]. Perhaps these alternative measures of dispersion have practical value for the problems considered here, but this issue is left to future investigations.

2.4 *The effect of heteroscedasticity on δ: some simulation results*

A practical issue is the effect of ignoring heteroscedasticity when using δ rather than ξ to measure effect size. That is, can the choice of method alter the extent an effect size is deemed to be large? Note that if the group with the larger sample size also has the larger variance, this results in a relatively small value for d. To illustrate how d compares with $\hat{\xi}$, simulations were used to estimate both effect sizes with $n_1 = 80, n_2 = 20$, where the first group has a normal distribution with mean 0.8 and standard deviation 4, and the second group has a standard normal distribution. Based on 1000 iterations, the median value of d was 0.22, which is typically considered a small effect size. (The mean value of d was nearly identical to the median.) The median value of $\hat{\xi}$ was 0.40, which would indicate a medium effect size. So even under normality, a heteroscedastic measure of effect size can make a practical difference. If instead the first group has standard deviation 1 and the second has standard deviation 4, now the median estimates are 0.42 and 0.32. That is, in contrast to the first situation, the choice between homoscedastic and heteroscedastic measure of effect size makes little difference. If instead $n_1 = n_2 = 20$, now the median d value is 0.30 and the median $\hat{\xi}$ value is 0.34. So the effect of ignoring heteroscedasticity is less of an issue with equal sample sizes, compared with the first situation considered, but it has practical consequences.

3. Measuring effect size for more than two groups

For the case of $J > 2$ groups, an obvious generalization of ξ^2 is to use

$$\sigma_\mu^2 = \frac{1}{J-1}\sum_{j=1}^{J}(\mu_j - \bar{\mu})^2$$

in the numerator of Equation (1). Again $\sigma^2(Y)$ is conceptualized as the unconditional variance of Y when equal sample sizes are used with probability 1. That is, if m observations are randomly sampled from each group, where again m is the smallest sample size, $\sigma^2(Y)$ is viewed as the population variance associated with the Jm pooled random variables. And when the sample sizes are not all equal, K random subsets of size m from the groups with the larger sample sizes can again be used to estimate $\sigma^2(Y)$. And as before, the K resulting estimates of explanatory power are averaged to yield a final estimate of $\sigma^2(Y)$. Given the speed of modern computers, estimates can be quickly obtained even when K is fairly large.

4. Some illustrations

The motivation for this paper stems from the first author's involvement in an intervention study aimed at improving the overall well being of older adults. One portion of the study consisted of two groups: an ethnic match between the participant and therapist and a non-match. The total sample size was 173 and the intervention lasted six months. Three outcome variables of interest dealt generally with physical measures based on version 2 of the 36-item Short-Form Health Survey SF-36v2 [21,22]. The three measures reflect bodily pain, physical function and a physical composite score. Researchers in this field routinely use Cohen's d to assess the success of intervention and for the situation at hand the values for Cohen's d were 0.59, 0.21 and 0.42, which were interpreted as small to medium effect sizes. However, boxplots indicated that there are outliers suggesting that for the typical participant, perhaps the efficacy of the intervention was being underestimated. Using the robust explanatory measure of effect size proposed here, with a 20% trimmed mean and Winsorized variance, yielded 0.5, 0.25 and 0.39, suggesting a medium to large effect size, the only point being that the choice of method can impact the perceived difference between two groups.

In another portion of this study, eight groups were compared based on a measure of overall mental health. The total sample size was 360. The eight groups corresponded to levels of activities outside the home. The researchers wanted an overall assessment of the extent the eight groups differ, but it was unknown how to proceed in a robust manner that allows heteroscedasticity. The explanatory measure of effect size was 0.5 suggesting a relatively large effect, again using a 20% trimmed mean and Winsorized variance.

As another illustration, Thompson and Randall-Maciver [19] report four measurements for male Egyptian skulls from five different time periods: 4000 BC, 3300 BC, 1850 BC, 200 BC and 150 AD. There are 30 skulls from each time period and four measurements: maximal breadth, basibregmatic height, basialveolar length and nasal height. For illustrative purposes, we focus on the maximal breadth measure. If the proposed measure of effect is used with the mean and variance, the overall effect size among all five time periods is 0.35, which some would interpret as a medium effect size. But using a 20% trimmed mean and Winsorized variance, the effect size is 0.52, which is relatively large and again illustrates that using a robust, heteroscedastic method can make a practical difference in the assessed effect size.

5. Concluding remarks

In summary, an alternative to Cohen's d was proposed that has three practical characteristics: it allows heteroscedasticity, a robust version is easily implemented, and it is readily extended to more than two groups. An appeal of the proposed method is that has a simple interpretation. Even when using means, a method that allows heteroscedasticity can make a practical difference. And as was illustrated, robust versions can make a substantial difference regarding the assessment of the extent two or more groups differ.

Consider multivariate data associated with two independent groups. Wilcox [24, Section 6.9] describes a generalization of the Wilcoxon–Mann–Whitney test that is based on projecting the points onto the line passing through the measures of location associated with the two groups. It is noted that the measure of effect size proposed here is readily extended to this situation. For the skull data, if we compare the first two time periods using all four skull measures, the estimated effect size is 0.67, a relatively high value. Comparing the first and last time periods, the estimate is 0.82.

Finally, an R package at http://r-forge.r-project.org/projects/wrs/ contains the functions yuenv2 and t1wayv2, which estimate the proposed measure of effect size. Alternatively, go to www-rcf.usc.edu/~rwilcox/.

References

[1] L. Acion, J.J. Peterson, S. Temple, and S. Arndt, *Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects*, Statist. Med. 25 (2006), pp. 591–602.

[2] J. Algina, H.J. Keselman, and R.D. Penfield, *An alternative to Cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case*, Psychol. Meth. 10 (2005), pp. 317–328.

[3] N. Cliff, *Ordinal Methods for Behavioral Data Analysis*, Erlbaum, Mahwah, NJ, 1996.

[4] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, New York, 1977.

[5] K.A. Doksum and A. Samarov, *Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression*, Ann. Statist. 23 (1995), pp. 1443–1473.

[6] K.A. Doksum and G.L. Sievers, *Plotting with confidence: Graphical comparisons of two populations*, Biometrika 63 (1976), pp. 421–434.

[7] G.V. Glass, B. McGraw, and M.L. Smith, *Meta Analysis in Social Research*, Sage, Newbury Park.

[8] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, and W.A. Stahel, *Robust Statistics*, Wiley, New York, 1986.

[9] W.L. Hays, *Satistics*, Holt, Rinehart & Winston, Fort Worth, 1988.

[10] P.J. Huber, *Robust Statistics*, Wiley, New York, 1981.

[11] E. Kulinskaya and R.G. Staudte, *Interval estimates of weighted effect sizes in the one-way heteroscedastic ANOVA*, British J. Math. Statist. Psychol. 59 (2006), pp. 97–111.

[12] D.A. Lax, *Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions*, J. Amer. Statist. Assoc. 80 (1985), pp. 736–741.

[13] J.A. Randal, *A reinvestigation of robust scale estimation in finite samples*, Comput. Statist. Data Anal. 52 (2008), pp. 5014–5021.

[14] D.M. Rocke, *Robustness properties of S-estimators of multivariate location and shape in high dimension*, Ann. Statist. 24 (1996), pp. 1327–1345.

[15] J.L. Rosenberger and M. Gasko, *Comparing location estimators: Trimmed means, medians, and trimean*, in *Understanding Robust and exploratory data analysis*, D.F. Hoaglin, F. Mosteller, and J. Tukey, eds., Wiley, New York, 1983, pp. 297–336.

[16] P.J. Rousseeuw and A.M. Leroy, *Robust Regression & Outlier Detection*, Wiley, New York, 1987.

[17] L.H. Shoemaker and T.P. Hettmansperger, *Robust estimates and tests for the one- and two-sample scale models*, Biometrika 69 (1982), pp. 47–54.

[18] R.G. Staudte and S.J. Sheather, *Robust Estimation and Testing*, Wiley, New York, 1990.

[19] A. Thompson and R. Randall-Maciver, *Ancient Races of the Thebaid*, Oxford University Press, Oxford, 1905.

[20] J.W. Tukey, *A survey of sampling from contaminated normal distributions*, in *Contributions to Probability and Statistics*, I.S. Olkin, W. Ghurye, W. Hoeffding, W. Madow, and H. Mann, eds., Stanford University Press, Stanford, CA, 1960, pp. 448–485.

[21] J.E. Ware Jr., *SF-36 Health Survey update*, Spine 25 (2000), pp. 3130–3139.

[22] J. Ware, M. Kosinski, and J. Dewey, *How to score version 2 of the SF-36(r) health survey*, RI QualityMetric Inc., Lincoln, 2000.

[23] R.R. Wilcox, *Applying Contemporary Statistical Techniques*, Academic Press, New York, 2003.

[24] R.R. Wilcox, *Introduction to Robust Estimation and Hypothesis Testing*, 2nd ed., Academic Press, San Diego, CA, 2005.

[25] V.J. Yohai and R.H. Zamar, *High breakdown point estimates of regression by means of the minimization of an efficient scale*, J. Amer. Statist. Assoc. 83 (1988), pp. 406–414.