METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

One-Way and Two-Way ANOVA: Inferences About a Robust, Heteroscedastic Measure of Effect Size

Rand Wilcox [1]

[1] *Department of Psychology, University of Southern California, Los Angeles, CA, USA.*

Methodology, 2022, Vol. 18(1), 58–73, https://doi.org/10.5964/meth.7769

Received: 2021-11-02 • **Accepted:** 2022-03-04 • **Published (VoR):** 2022-03-31

Corresponding Author: Rand Wilcox, Department of Psychology, University of Southern California, 3620 S. McLintoch, Los Angeles, CA 90089-1061, USA. E-mail: rwilcox@usc.edu

Supplementary Materials: Data, Materials [see Index of Supplementary Materials]

Abstract

Consider a one-way or two-way ANOVA design. Typically, groups are compared based on some measure of location. The paper suggests alternative methods where measures of location are replaced by a robust measure of effect size that is based in part on a robust measure of dispersion. The measure of effect size used here does not assume that the groups have a common measure of dispersion. That is, it deals with heteroscedasticity. It is fairly evident that no single method reveals everything of interest regarding how groups differ. Certainly, comparing measures of location provides useful information. But as illustrated, comparing measures of effect size can provide a deeper understanding of how groups compare.

Keywords

ANOVA, non-normality, effect size, multiple comparisons, heteroscedasticity, interactions

Consider the goal of comparing independent groups. A broad issue is how many methods are needed to get a good understanding of how, and by how much, groups differ. Certainly, the best-known approach is to use some measure location such as the mean, median, 20% trimmed mean or even an M-estimator. Classic methods based on means provide some information regarding how groups differ, but it is fairly evident that this approach can miss details that are clinically important. For example, the difference between means might be relatively small when the difference between the medians is large. Of course, the reverse can happen where the difference between the medians is small but the difference between the means is large.

A limitation of methods that test hypotheses based on measures of location only is that they ignore the variation within the groups. The goal in this paper is to suggest methods that are based on a robust measure of effect size that is based in part on a robust measure of dispersion. As is illustrated later in this paper, taking into account the variation within groups can increase power substantially in some situations.

For J independent groups, let τ_j ($j = 1, ..., J$) denote some measure of dispersion associated with the jth group. The better-known measures of effect size assume homoscedasticity, meaning that $\tau_1 = ... = \tau_J$ is assumed. The measures of effect size used here avoid this assumption. That is, they allow heteroscedasticity. The idea is that comparing groups based on a heteroscedastic measure of effect size provides an alternative perspective that helps provide a deeper understanding of how groups differ.

Two situations are considered. The first is where the goal is to perform all pairwise comparisons among J independent groups. More formally, let η_{jk} denote a measure of effect size when comparing groups j and k. The goal is to test

$$H_0: \quad \eta_{jk} = 0 \tag{1}$$

and to compute a confidence interval for η_{jk} for all $j < k$. Typically, as is the case here, η_{jk} is based in part on the difference between two measures of location, say $\theta_j - \theta_k$. Consequently, a decision can be made about whether η_{jk} is greater than or less than zero by simply testing the hypothesis H_0: $\theta_j = \theta_k$. But here, η_{jk} is also based on a measure of dispersion. Consequently, power when testing H_0: $\theta_j = \theta_k$ can differ from the power of testing (1) simply because the method for testing (1) is sensitive to different features of the data. And even when these two approaches have similar power, the problem of computing a confidence interval for η_{jk} remains.

The second goal deals with comparing the levels of the first (or second) factor of a J-by-K ANOVA design. To elaborate a bit, momentarily focus on a 2-by-K design where the goal is to compare the first and second levels of the first factor. To be a bit more precise, consider $2K$ independent random variables. Let ξ_1 be some global measure of effect size associated with the first K random variables and let ξ_2 denote the measure of effect size for the remaining K variables. As will be made evident, the choice for ξ used here is related to the choice for η but it differs in a fundamental way. The goal is to test

$$H_0: \quad \xi_1 = \xi_2 \tag{2}$$

And there is the related goal of computing a confidence interval for $\xi_1 - \xi_2$. The proposed method is readily extended to a J-by-K design where the goal is to make inferences based on all pairwise levels of the first or second factor. The results here extend the results in Wilcox (2022b), which was limited to two-by-two design.

When testing (1), a seemingly natural guess is that a percentile bootstrap method will perform well in terms of controlling the Type I error probability. Simulations related to

this issue are reported here. As for testing (2), a percentile bootstrap method was found to perform poorly; the actual level was estimated to be well below the nominal level. An alternative approach is proposed here and studied via simulations.

The paper is organized as follows. The next section describes the robust heteroscedastic measures of effect size that will be used. Both are based on a simple extension of measures of effect size proposed by Kulinskaya et al. (2008) and Kulinskaya and Staudte (2006). This is followed by proposed methods for testing (1) and (2). Then some simulations results are reported followed by two illustrations.

Robust, Heteroscedastic Measures of Effect Size

Momentarily consider two independent groups and let $\bar{X}_j$ denote the sample mean for the jth group. Let n_j denote the corresponding sample size. Let μ_1 and μ_2 denote the population means and let σ_1 and σ_2 denote the population standard deviations. Certainly, one of the better-known measures of effect size is

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}$$

where it is assumed that $\sigma = \sigma_1 = \sigma_2$. That is, homoscedasticity is assumed. Cohen's d, (e.g., Cohen, 1988) provides an estimate of Δ that is biased. Hedge's g (Hedges & Olkin, 1985) provides an unbiased estimator for Δ. Glass et al. (1981) avoid the homoscedasticity assumption by specifying one of the groups as the control group and then use only the estimate of the variance for the control group to measure effect size. Kulinskaya et al. (2008) avoid the homoscedasticity assumption in the following manner. Note that the standard error of $\bar{X}_1 - \bar{X}_2$ can be written as ς^2/N where

$$\varsigma^2 = \frac{(1-q)\sigma_1^2 + q\sigma_2^2}{q(1-q)}$$

$N = n_1 + n_2$ and $q = n_1/N$. As a result, they use

$$\Delta_{KMS} = \frac{\mu_1 - \mu_2}{\varsigma} \tag{3}$$

as a measure of effect size. Kulinskaya and Staudte (2006, p. 101) note that a natural generalization of Δ to the heteroscedastic case does not appear to be possible without taking into account the relative sample sizes. Under normality and when the population variances are equal, $\Delta = 2\Delta_{KMS}$. For $J \geq 2$ independent groups, Kulinskaya and Staudte (2006) generalized this approach by estimating effect size with

$$\sum \frac{(\bar{X}_j - \widetilde{X})^2}{s_j^2} \tag{4}$$

where $\widetilde{X} = \sum q_j \bar{X}_j / \sum q_j$ and s_j^2 is the sample variance associated with the jth group. Their results include a method for computing an interval estimate of the population analog of (4) assuming normality.

A parameter is said to be non-robust if a small change in a distribution has a large impact on its value. Formal mathematical methods for characterizing robust parameters are summarized by Hampel et al. (1986), Staudte and Sheather (1990) as well as Huber and Ronchetti (2009). The point here is that the population mean and variance are not robust. In practical terms, when using the measures of effect size just reviewed, even a small departure from a normal distribution can mask a large effect size among the bulk of the participants. Algina et al. (2005) illustrate this point when using Δ.

Here, robust versions of (3) and (4) are used, which are based on a trimmed mean and a Winsorized variance. This mimics the basic approach used by Algina et al. (2005) to derive a robust version of Δ. For notational convenience, momentarily focus on a single random sample $X_1, \ldots, X_n$ and let $X_{(1)} \leq \ldots \leq X_{(n)}$ denote the values written in ascending order. Let γ denote some specified constant, $0 \leq \gamma < 0.5$, and let $g = \lfloor \gamma n \rfloor$, where $\lfloor \gamma n \rfloor$ is the value of γn rounded down to the nearest integer. The sample γ trimmed mean is computed by removing the g largest and g smallest observations and averaging the values that remain. More formally, the sample trimmed mean is

$$\bar{X}_t = \frac{X_{(g+1)} + \cdots + X_{(n-g)}}{n - 2g}$$

The choice $\gamma = 0.2$ has been studied extensively (e.g., Wilcox, 2022a) and is used here. It has good efficiency under normality and its standard error can be substantially smaller than the standard error of the mean when dealing with heavy-tailed distributions where outliers are likely to occur.

Winsorizing $X_1, \ldots, X_n$ means that the g smallest values are set equal to $X_{(g+1)}$ and the g largest are set equal to $X_{(n-g)}$. The γ Winsorized mean, $\bar{X}_w$, is the mean of the Winsorized values and the Winsorized sample variance, s_w^2, is the usual sample variance based on the Winsorized data. When sampling from a normal distribution and when $\gamma = 0.2$, $s_{wN}^2 = s_w^2 / 0.642$ estimates the population variance. Let $\bar{X}_{tj}$ denote the trimmed mean for the jth group. Here, the measure of effect size that will be used is estimated with

$$\hat{\xi} = \frac{J}{2} \sqrt{\sum \frac{(\bar{X}_{tj} - \widetilde{X})^2}{s_{jwN}^2}} \tag{5}$$

where $\widetilde{X} = \sum q_j \bar{X}_{tj} / \sum q_j$ and s^2_{jwN} is the Winsorized variance of the jth group rescaled to estimate the variance when dealing with a normal distribution. This will be called the KMS measure of effect size henceforth. The robust version of (3) is estimated with

$$\widehat{\eta} = \frac{\bar{X}_{t1} - \bar{X}_{t2}}{\widehat{\varsigma}} \tag{6}$$

where

$$\widehat{\varsigma}^2 = \frac{(1-q)s^2_{1wN} + qs^2_{2wN}}{q(1-q)}$$

Consider the case where $J-1$ of the groups have a common population trimmed mean and the other group has a population trimmed mean that is larger than the other $J-1$ trimmed means by some specified amount. If the term $J/2$ is excluded from (5), the resulting measure of effect size decreases as J increases. By including this term, $\widehat{\xi}$ remains similar to what would be obtained when $J = 2$.

Testing (1): Method M1

A basic percentile bootstrap method is used to test (1). Momentarily focus on two independent groups. First, generate a bootstrap sample from each group. That is, randomly sample with replacement n_j values from the jth group. Based on these bootstrap samples, compute the measure of effect size corresponding to (6) and for notational convenience label the result D^*. Repeat this process B times yielding $D^*_1, \ldots, D^*_B$. Let $D^*_{(1)} \leq \cdots \leq D^*_{(B)}$ denote the D^* values written is ascending order. Then a $1-\alpha$ confidence interval for η_{12} is

$$\left(D^*_{(\ell+1)}, D^*_{(u)}\right)$$

where $\ell = \alpha B/2$, rounded to the nearest integer, and $u = B - \ell$. Let A denote the number of times $D^* > 0$ and let $P^* = A/B$. From Liu and Singh (1997), a (generalized) p-value is $2\min(P^*, 1-P^*)$. This is called method M1 henceforth.

When $J > 2$, the improvement on the Bonferroni method, derived by Hochberg (1988), is used to control the family wise error rate (FWE) meaning the probability of one or more Type I errors. Given C p-values, say $p_1, \ldots, p_C$, the computational details are as follows. Put the p-values in descending order yielding $p_{[1]} \geq \ldots \geq p_{[C]}$. Suppose the goal is to have FWE equal to α. Set $k = 1$.

1. If $p_{[k]} \leq \alpha/k$, reject all C hypotheses.
2. If $p_{[k]} > \alpha$, increment k by one. If $p_{[k]} \leq \alpha/k$, reject all hypotheses where the p-value is less than or equal to α/k.

3. If $p_{[k]} > \alpha$, repeat step 2.

Testing (2): Method M2

As is evident, the percentile bootstrap method is readily adapted to testing (2) where J=2 and now $\hat{\xi}_j$ ($j = 1, 2$) is given by (5) based on the jth level of the first factor and corresponding K levels of the second factor. However, simulations made it clear that this approach is highly unsatisfactory when the sample sizes are relatively small. The actual level was estimated to be well below the nominal level. A much more satisfactory approach is to use a simulation to estimate the null distribution of $F = \hat{\xi}_1 - \hat{\xi}_2$. Now let n_{jk} denote the sample size corresponding to the jth level of the first factor and the kth level of the second factor. The initial strategy was, for the jth level of the first factor and the kth level of the second, generate n_{jk} observations from a standard normal distribution and then compute F yielding say F^*. This is repeated I times yielding $F_1^*, \ldots, F_I^*$, which yields an estimate of the null distribution of F. Here, $I = 5000$ is used. However, for small sample sizes, control over the Type I error probability was not quite satisfactory when dealing with a skewed, light-tailed distribution.

Let Z be a random variable having a standard normal distribution. Then

$$V = \frac{\exp(gZ) - 1}{g} \exp\left(h\frac{Z^2}{2}\right)$$

has a g-and-h distribution, where g and h are parameters that determine the first four moments (Hoaglin, 1985). When $g = 0$, the expression for V is taken to be

$$V = Z \exp\left(h\frac{Z^2}{2}\right)$$

Here, an estimate of the null distribution of F is based on data randomly sampled from a g-and-h distribution with $g = 0.75$ and $h = 0$.

Let $\hat{\delta}$ denote the estimate of $\xi_1 - \xi_2$. Let $H = \sum \mathbf{I}(\hat{\delta} < F_i^*)/I$ where the indicator function $\mathbf{I}(\hat{\delta} < F_i^*) = 1$ if $\hat{\delta} < F_i^*$, otherwise $\mathbf{I}(\hat{\delta} < F_i^*) = 0$. A p-value is $2\min(H, 1 - H)$. A $1 - \alpha$ confidence interval can be computed by assuming that the null distribution of F is reasonably well approximated as described above. Let f_q be an estimate of the qth quantile of the null distribution of F based on $F_1^*, \ldots, F_I^*$. Then a $1 - \alpha$ confidence interval for $\xi_1 - \xi_2$ is readily shown to be

$$\left(\hat{\delta} - f_{1-\alpha/2}, \hat{\delta} - f_{\alpha/2}\right)$$

That is, an estimate of the upper quantile of the distribution of F is used to compute the lower end of the confidence interval. (This result is similar to how confidence intervals are computed via a bootstrap-t method. See Wilcox, 2022a, for details.) When $J > 2$ and

the goal is to perform all pairwise comparisons among the J levels of the first factor, FWE is controlled via Hochberg's method.

Some comments about a 2-by-2 design are helpful. For this special case, let θ_{jk} denote some measure of location associated with the jth level of the first factor and the kth level of the second factor. Let $\xi_{j.}$ denote ξ when comparing the groups associated with the jth level of the first factor and the first and second levels of the second factor. In a similar manner, let $\xi_{.k}$ denote ξ when comparing the groups associated with the kth level of the second factor and the first and second levels of the first factor. As is evident, when dealing with an interaction, testing

$$H_0: \quad \theta_{11} - \theta_{12} = \theta_{21} - \theta_{22}$$

is the same as testing

$$H_0: \quad \theta_{11} - \theta_{21} = \theta_{12} - \theta_{22}$$

That is, it is irrelevant whether differences are based within rows rather than within columns. Note that testing

$$H_0: \quad \xi_{1.} = \xi_{2.} \tag{7}$$

is an analog of testing for an interaction. Rather than comparing measures of effect size based on the difference between measures of location only, a measure of effect size is used that is based in part on the Winsorized variance within each group. But testing (7) is not necessarily the same as testing

$$H_0: \quad \xi_{.1} = \xi_{.2} \tag{8}$$

That is, comparing measures of effect size corresponding to the levels of the first factor differs from comparing measures of effect size corresponding to the levels of the second factor.

Simulation Results

Simulations were used to assess the small sample properties of methods M1 and M2. Data were generated from four types of distributions: normal, symmetric and heavy-tailed symmetric and relatively light-tailed, and asymmetric and relatively heavy-tailed, roughly meaning that outliers tend to be common. More specifically, data were generated from four g-and-h distributions. The four distributions used here are the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0$), a skewed distribution with relatively light tails ($g = 1$, $h = 0$), and a skewed distribution with heavy tails ($g = 1$, h = 0.2). Table 1 summarizes the skewness (κ_1) and kurtosis (κ_2) of these distributions. All

g-and-h distributions have a median equal to zero. For $g = 0$ and $h = 0.2$ the variance is 2.15. For g = 1 and h = 0, the mean and variance are 0.648 and 4.67, respectively. For g = 1 and h = 0.1, the mean and variance are 0.97 and 30.6. Figure 1 shows plots of these four distributions.

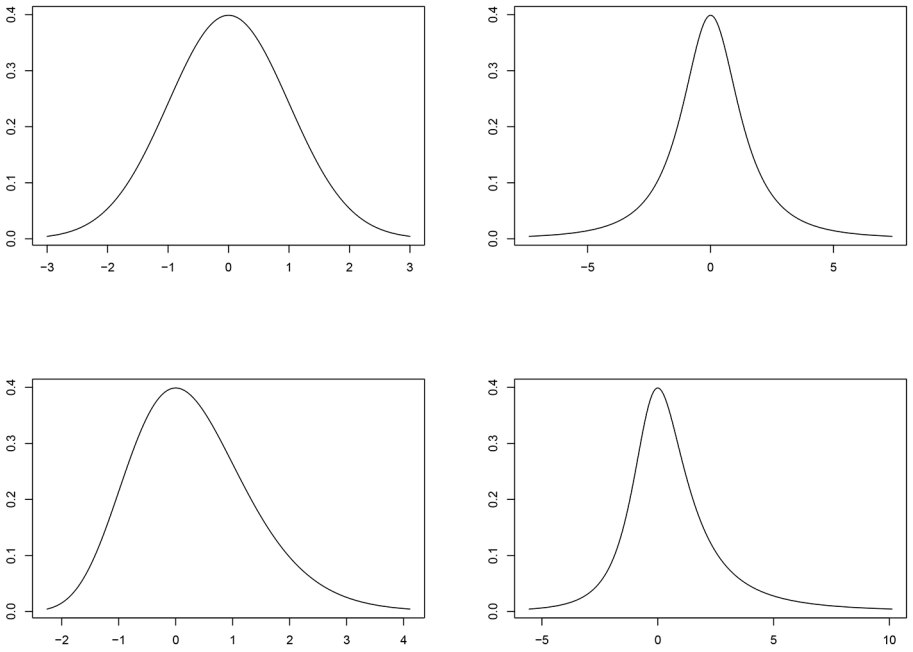
Table 1

Some Properties of the g-and-h Distribution

g	h	κ_1	κ_2
0.0	0.0	0.00	3.00
0.0	0.2	0.00	21.46
1.0	0.0	0.61	3.68
1.0	0.2	32.81	2295.98

Figure 1

The Four g-and-h Distributions Used in the Simulations



Note. In the upper left panel, $g = h = 0$, in the upper right, $g = 0$ and $h = 0.2$, in the lower left $g = 1$ and $h = 0$ and in the lower right $g = 1$ and $h = 0.2$.

The range of distributions used here is motivated by a review of several studies aimed at charactering the extent distributions differ from normality (Wilcox, 2022a, section 4.2). It is noted that the reported skewness and kurtosis for $g = h = 0.2$ are estimates of the actual skewness and kurtosis based on a sample size of one million. Repeating this process yielded even larger estimates. This particular distribution might appear to be extreme. The point here is that if a method performs reasonably well when dealing with this distribution, this provides some assurance that it will perform reasonably well for distributions that are likely to be encountered.

For method M1, simulations were run for $J = 2$ and 4 groups. For $J = 4$, three sample size configurations were used: (20, 20, 20, 20), (40, 40, 40, 40) and (20, 20, 40, 40). For $J = 2$ the sample sizes were (20, 20), (40, 40) and (20, 40). For convenience these three sample size configurations are labeled N1, N2 and N3, respectively. The results for N1 and N2 are reported in Table 2 and are based on 1000 replications. The estimates for N3 were very similar and so for brevity are not reported. Note that the largest estimate occurs for N2, $J = 4$, g= 0 and $h = 0.2$. Increasing the common sample size to 60, the estimate is 0.040. Some additional simulations were run where the first group has a standard deviation four times larger than the other groups. Again, control over the Type I probability was similar to the results in Table 2. Bradley (1978) has suggested that as a general guide, when testing at the 0.05 level, the actual level should be between 0.025 and 0.075. All indications are that M1 satisfies this criterion. Replacing the KMS method with the robust version of Δ derived by Algina et al., 2005, gives very similar results.

Table 2

Estimated Type I Errors for Method M1

g	h	$J = 2$	$J = 4$
		N1	
0.0	0.0	0.055	0.055
0.0	0.2	0.053	0.051
1.0	0.0	0.056	0.057
1.0	0.2	0.041	0.053
		N2	
0.0	0.0	0.048	0.053
0.0	0.2	0.047	0.068
1.0	0.0	0.040	0.051
1.0	0.2	0.047	0.048

Note. $\alpha = 0.05$.

There is a well-established heteroscedastic method for performing all pairwise comparisons based on trimmed means (e.g., Wilcox, 2022a, section 7.4.1). For convenience, this

method is labeled T1. In some cases, T1 will have about the same amount of power as M1, but the expectation is that the choice of method might make a practical difference simply because they are sensitive to different features of the data. Consider, for example, normal distributions having variance one and suppose the mean of the first group is 0.5 while the remaining groups have a mean of zero. For N2, the probability of one or more significant results was estimated to be 0.553 for method M1 and 0.543 for method T1: Hochberg's method was also used to control FWE when using T1. Increasing the common sample size to 60, the estimates were 0.734 and 0.693, respectively. If instead the first random variable is divided by 4 and 0.8 is added to the first and second random variables, the estimates are 0.945 and 0.865.

As for method M2, simulations were run for $J = 2$ and $K = 4$. Four sample size configurations were used. The first three had a common sample size of 20, 50 and 100. Now N1, N2 and N3 are used to denote these three sample configurations, respectively. The fourth, N4, consisted of a common sample size of 20 for the first four groups (corresponding to the first level of the first factor) and a common sample size of 50 for the remaining four groups. The Type I error probability was estimated with 5000 replications. Table 3 shows the estimated probability of a Type I error, $\widehat{\alpha}$, when testing at the α= 0.05 level.

As can be seen, the estimates satisfy Bradley's criterion in all situations except N1 and when dealing with a symmetric distribution; the estimates are less than 0.025, the lowest estimate being 0.016.

Method T1 is readily extended to a two-way ANOVA design where the goal is to perform all pairwise comparisons of the levels of the first or second factor, which is called method T2 henceforth. The details are in Wilcox (2022a, section 7.2.1; here, the R function t2way in the R package WRS was used to apply method T2). As was the case for a one-way design, M2 and T2 are sensitive to different features of the data, so which method has more power depends on the situation. A few simulations are reported here to provide some indication of the extent the choice of method matters, where again $J = 2$ and $K = 4$ and the goal is to compare the two levels of the first factor. Power was estimated for situations where, excluding the first group (the group corresponding to the level 1 of both factors), the remaining groups have a standard normal distribution. The first group also had a normal distribution but a mean $\mu > 0$. Values for the standard deviation for the first group were taken to be $\sigma = 0.5$, 1 and 2. Table 4 reports the results. As can be seen, M2 generally has the highest power. But as indicated, this is not always the case. The only point is that the power of the two methods can differ substantially. Moreover, the method that has the most power depends on how the groups differ which is not known. But perhaps the more important point is that the two methods provide different perspectives on how the groups differ.

Table 3

Estimated Type I Errors for Method M2

g	h	$\widehat{\alpha}$
	N1	
0.0	0.0	0.021
0.0	0.2	0.016
1.0	0.0	0.071
1.0	0.2	0.070
	N2	
0.0	0.0	0.032
0.0	0.2	0.029
0.2	0.0	0.063
1.0	0.2	0.063
	N3	
0.0	0.0	0.043
0.0	0.2	0.042
1.0	0.0	0.055
0.2	0.2	0.057
	N4	
0.0	0.0	0.029
0.0	0.2	0.028
1.0	0.0	0.065
1.0	0.2	0.065

Note. $\alpha = 0.05$.

Table 4

Estimated Power for Methods M2 and T2

μ	σ	**M2**	**T2**
	N1		
1.0	1.0	0.447	0.301
1.0	0.5	0.798	0.229
1.0	2.0	0.284	0.206
	N3		
0.5	1.0	0.712	0.361
0.5	0.5	0.993	0.407
0.5	3.0	0.290	0.213

Two Illustrations

Methods M1 and M2 are illustrated based on data stemming from a study of an intervention program aimed at improving the physical and mental health of older adults (Clark et al., 2012). The first illustration is based on a measure of depressive symptoms (CESD) taken after intervention. The goal is to compare five groups corresponding to a participant's educational level: less than high school, high school graduate, some college or technical school, four years of college completed, post-graduate study. The total sample size is 328. Table 5 reports the p-values for methods M1 and T1. The p-values were adjusted via Hochberg's method. As indicated, these two methods are consistent in terms of whether a significant result is obtained at the 0.05 level. Note, however, that the p-values differ substantially in some cases indicating that the two methods can paint a decidedly different picture about which groups differ significantly.

Table 5

The Hochberg Adjusted p-Values

Group		M1	T1
1	**2**	0.084	0.060
1	**3**	0.000	0.002
1	**5**	0.000	0.001
2	**3**	0.352	0.534
2	**4**	0.260	0.325
2	**5**	0.564	0.103
3	**4**	0.564	0.999
3	**5**	0.564	0.825
4	**5**	0.564	0.997

Note. When comparing education groups based on a measure of depressive symptoms.

Presumably, what constitutes a large effect size can depend on the situation. For illustrative purposes, suppose $\Delta = 0.2$, 0.5 and 0.8 are viewed as small, medium and large effect size, respectively, as is sometimes suggested (e.g., Cohen, 1988). Then under normality and homoscedasticity, $2\Delta_{KMS} = 0.2$, 0.5 and 0.8 correspond to small, medium and large effect sizes as well. For the three significant results reported here, the estimates of $2\Delta_{KMS}$ range between 0.66 and 0.77. That is, the results indicate a rather substantial difference between participants who did not complete high school versus those who have some educational training beyond high school.

The second illustration deals with the goal of understanding the association between a measure of meaningful activities (MAPA) and two independent variables: a measure

of life satisfaction (LSIZ) and a participant's cortisol awakening response (CAR), which is the difference between cortisol measured upon awakening and measured again about 30-45 minutes later. The focus here is on measures taken after intervention.

For illustrative purposes, the data are split into four groups based on the medians of the two independent variables, resulting in a two-by-two ANOVA design. Wilcox (2019) demonstrates that this approach can reveal details that are missed by nonparametric regression methods. The median of the CAR values is −0.033, so the two CAR groups reflect approximately groups where cortisol increases versus decreases soon after awakening. Analyzing the data with method T2, for the low versus high LSIZ groups the p-value is 0.001. That is, there is reasonably strong evidence that the low LSIZ has a lower trimmed mean when CAR is ignored. For the low versus high CAR groups, ignoring LSIZ, the p-value is 0.509. Testing the hypothesis of no interaction, the p-value is 0.004. For low LSIZ, the 20% trimmed means for low versus high CAR groups are 33.03 and 30.64, respectively. The difference between these estimates differs at the 0.05 level; the p-value is 0.014. For the high LSIZ group the estimates are 34.31 and 35.74, which do not differ at the 0.05 level; the p-value is 0.186. That is, the evidence suggests that there is a disordinal interaction, but the strength of the evidence is not very strong.

As for method M2, for low LSIZ, the estimate of ξ, $\xi_{1.}$, when comparing the groups corresponding to low versus high CAR values, is 0.248. For high LSIZ scores, the estimate of $\xi_{2.}$, is 0.129. Testing (6), the p-value is 0.160. In contrast, for low CAR values, comparing the low and high LSIZ groups, the estimate of $\xi_{.1}$ is 0.131 and for high CAR values the estimate of $\xi_{.2}$ is 0.468. The p-value when testing (8) is 0.002 and the 0.95 confidence interval for $\xi_{.1} - \xi_{.2}$ is (-0.510, -0.146). That is, testing (8) yields a significant result at the 0.05 level in contrast to testing (7).

As previously noted, based on trimmed means only, a significant interaction was obtained. The results based on ξ also indicate an interaction but with the added benefit of an alternative perspective regarding the nature and relative importance of the interaction. In particular, the results demonstrate that differences within rows can yield a different perspective compared to differences within columns when using the KMS measure of effect size. Roughly, knowing whether cortisol increases or decreases appears to provide information about how low and high life satisfaction groups differ in terms of meaningful activities. When cortisol decreases, the results indicate that there is a much more pronounced difference between the high and low LSIZ groups compared to when cortical increases.

Concluding Remarks

Steegen et al. (2016) discuss and illustrate an extremely important issue: no single method reveals everything of interest when comparing groups. Multiple methods can be required to get a deep and nuanced understanding of data. The goal here is to suggest

a method for comparing groups, via a robust, heteroscedastic measure of effect size, that helps achieve this goal. Here, the illustration involving a 2-by-2 ANOVA design demonstrates this point.

Perhaps it should be stressed that it is not being suggested that methods based on measures of location only should be abandoned. Surely, they provide useful information about how groups compare. Again, the only suggestion is that comparing groups based on a robust measure of effect size that includes some measure of variation can provide insights about how groups compare.

There are several other ways for dealing with an interaction, based on a heteroscedastic measure of effect size, beyond the approach used here (Wilcox, 2022a, section 7.4.17). One approach is to use a quantile shift measure of effect size and another is to use the notion of explanatory power. Inferences based on these methods are under investigation.

Funding: The author has no funding to report.

Acknowledgments: The author has no additional (i.e., non-financial) support to report.

Competing Interests: The author has declared that no competing interests exist.

Data Availability: For this article, data is freely available (see Index of Supplementary Materials).

Supplementary Materials

For this article, the following supplementary materials are available (see Index of Supplementary Materials below):

- The file TWO-WAY_KMS_illustration.pdf contains the R code that was used in the illustrations plus some additional analyses of the data.
- The R functions for applying the methods in this paper are stored in the file Rallfun-v39. The function KMSmcp.ci applies method M1 and AN2GLOB.KMS applies method M2. In the illustration involving LSIZ and CAR, method M2 was applied via the R function KMSgridRC, which provides a convenient way of splitting the data as was described. The function KMS2way performs all relevant pairwise comparisons and tests all tetrad interactions.
- The code used in the simulations is stored in the files KMSmcp_ci_sim.tex (method M1) and ANOG2KMS_sim.tex (method M2).

Index of Supplementary Materials

Wilcox, R. (2022). *Supplementary materials to "One-way and two-way ANOVA: Inferences about a robust, heteroscedastic measure of effect size"* [Data, code, materials]. OSF. https://osf.io/xhe8u/.

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. https://doi.org/10.1037/1082-989X.10.3.317

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical & Statistical Psychology, 31*(2), 144–152. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh, M., Knight, B. G., Mandel, D., Blanchard, J., Granger, D. A., Wilcox, R. R., Lai, M. Y., White, B., Hay, J., Lam, C., Marterella, A., & 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 and Community Health, 66*(9), 782–790. https://doi.org/10.1136/jech.2009.099754

Cohen, J. (1988), *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum.

Glass, G. V., McGaw, B., & Smith, M. L. (1981), *Meta-analysis in social research*. Sage.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. Wiley.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller, & J. Tukey (Eds.), *Exploring data tables trends and shapes* (pp. 461-511). Wiley.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*(4), 800–802. https://doi.org/10.1093/biomet/75.4.800

Huber, P. J., & Ronchetti, E. (2009). *Robust statistics* (2nd ed.). Wiley.

Kulinskaya, E., Morgenthaler, S., & Staudte, R. (2008). *Meta analysis: A guide to calibrating and combining statistical evidence*. Wiley

Kulinskaya, E., & Staudte, R. G. (2006). Interval estimates of weighted effect sizes in the one-way heteroscedastic ANOVA. *British Journal of Mathematical & Statistical Psychology, 59*(1), 97–111. https://doi.org/10.1348/000711005X68174

Liu, R. G., & Singh, K. (1997). Notions of limiting p values based on data depth and bootstrap. *Journal of the American Statistical Association, 92*(437), 266–277. https://doi.org/10.2307/2291471

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. Wiley.

Steegen, S., Tuerlinck, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Wilcox, R. R. (2019). Regression when there are two covariates: Some practical reasons for considering quantile grids. *Journal of Modern Applied Statistical Methods, 18*(1), Article eP3227. https://doi.org/10.22237/jmasm/1556670120

Wilcox, R. R. (2022a). *Introduction to robust estimation and hypothesis testing* (5th ed.). Academic Press.

Wilcox, R. R. (2022b). Two-way ANOVA: Inferences about interactions based on robust measures of effect size. *British Journal of Mathematical & Statistical Psychology, 75*(1), 46–58. https://doi.org/10.1111/bmsp.12244

Methodology is the official journal of the European Association of Methodology (EAM).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.