Journal of Statistics and Computer Science
Vol. 1, Issue 2, 2022, pp. 119-134
https://doi.org/10.47509/JSCS.2022.v01i02.02

TWO-WAY ANOVA: INFERENCES ABOUT AN INTERACTION, BASED ON A ROBUST HETEROSCEDASTIC MEASURE OF EFFECT SIZE, WHEN THERE IS A COVARIATE

Rand R. Wilcox
Dept of Psychology
University of Southern California

November 12, 2022

Abstract

Recently, there have been advances on characterizing an interaction in a two-way ANOVA design in a manner that takes into account simultaneously some robust measure of location and some robust measure of dispersion. These methods allow heteroscedasticity. The first goal in this paper is to suggest an analog of one of these measures of effect size when there is a covariate. Extant results related to the proposed method suggest how to make inferences about the proposed measure of effect size, but even for moderately small sample size, they proved to be unsatisfactory. The second goal here is to suggest a method for dealing with this issue.

keywords: robust effect size, heteroscedasticity, quantile regression estimator

Article History
Received : 28 September 2022; Revised : 12 November 2022; Accepted : 16 November 2022; Published : 15 December 2022

To cite this paper
Rand R. Wilcox (2022). Two-Way Anova: Inferences about an Interaction, based on a Robust Heteroscedastic Measure of Effect Size, When there is a Covariate. *Journal of Statistics and Computer Science.* 1(2), 119-134.

1 Introduction

There are two main goals in this paper. The first is to suggest a robust method for characterizing an interaction in a two-way ANOVA design, when there is a covariate, in a manner that takes into account both a measure of location and a measure of dispersion. This is done based on a simple analog of a measure effect size derived by Kulinskaya et al. (2008). The basic idea is to provide an alternative to simply comparing measures of location that helps provide a deeper and more nuanced understanding of how the groups differ. As stressed and illustrated by Steegan et al. (2014), multiple perspectives can be essential when analyzing data. The method used here has certain similarities to the method in Wilcox (2022b) when there is no covariate. But the method here differs in fundamental ways as will be explained.

The second general goal is to suggest a method for making inferences about the proposed measure of effect size. For reasons reviewed below, an obvious speculation is that a percentile bootstrap method, or bootstrap estimate of the standard error, would perform well. But this was not the case when dealing with moderately small sample sizes. A method for dealing with this issue is proposed and studied via simulations.

To put the proposed method in perspective, and to help motivate the approach used, first consider two independent groups having measures of location θ_1 and θ_2 and measures of scale τ_1 and τ_2. Of course, typically groups are compared based on the measures of location with the measures of scale viewed as nuisance parameters. A variety of robust methods are described in WIlcox (2022a). But another approach is to focus on some measure of effect size that takes into account both measures of location and scale. Generally these measures of effect size take the form

$$\frac{\theta_1 - \theta_2}{f(\tau_1, \tau_2)}, \tag{1}$$

where $f(\tau_1, \tau_2)$ is some function of the measures of dispersion to be determined. Seemingly the best-known version of this approach is

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}, \tag{2}$$

where μ_j is the mean of the jth groups, σ_j is the standard deviation and by assumption $\sigma_1 = \sigma_2 = \sigma$ (e.g., Cohen, 1988). But there are two well-known concerns with this measure of effect size. First, it is not robust, roughly meaning that a small change in the distributions

can alter Δ substantially. In particular, a small departure from normality can mask a relatively large effect size among the bulk of the participants as illustrated by Algina et al. (2005). A second concern is that Δ assumes homoscedasticity.

Let $\bar{Y}_j$ denote the sample mean for the jth group. Let $N = n_1 + n_2$, $q = n_1/N$ and note that the squared standard error of $\bar{Y}_1 - \bar{Y}_2$ can be written as ς^2/N, where

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

Kulinskaya et al. (2008) deal with heteroscedasticity by replacing Δ with

$$\xi = \frac{\mu_1 - \mu_2}{\varsigma}, \tag{3}$$

which is called the KMS measure of effect size henceforth. But this method is not robust simply because the mean and variance are not robust for general reasons summarized by Hampel et al. (1986), Huber and Ronchetti (2009) as well as Staudte and Sheather (1990). One way of dealing with this issue is to follow Algina et al. (2005) and replace the mean and variance with a 20% trimmed and Winsorized variance. This is the approach used by Wilcox (2022b) when dealing with an interaction. More precisely, consider a two-by-two ANOVA design. For the first level of the first factor, let ξ_1 denote ξ based on the two levels of the second factor. Similarly, for the second level of the first factor, let ξ_2 denote ξ based on the two levels of the second factor. Wilcox (2022b) used to

$$\delta = \xi_1 - \xi_2. \tag{4}$$

to characterize an interaction.

Further details about the method used in Wilcox (2022b) are omitted because they do not play a direct role here. To provide some indication why, let X denote some covariate. Let $\delta(x)$ denote δ given that the covariate $X = x$. Estimating $\delta(x)$ requires information about the conditional distribution of Y given that $X = x$. An approach to this problem is suggested here that is based on robust measures of location and scale that differ from the trimmed mean and Winsorized variance used by Wilcox (2022b). The reason for not using the conditional trimmed mean and Winsorized variance is explained in section 2.

The paper is organized as follows. Section 2 describes an analog δ when there is a covariate. This is followed by a description of a method for testing

$$H_0 : \delta(x) = 0, \tag{5}$$

no effect, given that $X = x$. The method yields a confidence interval as well. Section 3 reports simulation results on how well a modified bootstrap method performs when dealing with relatively small sample sizes and non-normal distributions. Section 4 illustrates the method using data dealing with the physical and emotional wellbeing of older adults.

For completeness, Tukey (1991) argued that testing for exactly equality, as is done in (5), is nonsensical because surely $\delta(x)$ differs from zero at some decimal place. Jones and Tukey (2000) suggest dealing with this issue via Tukey's three decision rule. If (5) is rejected, make a decision about whether $\delta(x)$ is greater than or less than zero. If (5) is not rejected, make no decision. From this perspective, a p-value reflects the strength of the empirical evidence can be made. But it does not indicate the probability of making a correct decision.

2 The Proposed Method

The proposed method assumes that for jth level of the first factor, and the kth level of the second factor, the qth quantile of the dependent variable Y_{jk}, given that covariate $X_{jk} = x$, is

$$Y_{jkq} = \beta_{0jkq} + \beta_{1jkq}x, \tag{6}$$

where β_{0kjq} and β_{1jkq} are unknown parameters ($j = 1, 2$; $k = 1, 2$). The unknown slopes and intercepts are estimated via the Koenker and Bassett (1978) estimator. Briefly, for any group, let r_i ($i = 1, \ldots, n_{jk}$) denote the residuals. The slope and intercept are taken to be the values that minimize

$$\sum \psi_q(r_i), \tag{7}$$

where

$$\psi_q(u) = u(q - I_{u<0}). \tag{8}$$

Let b_{0kjq} and b_{1jkq} denote the estimates of β_{0kjq} and β_{1jkq}, respectively. For $q = 0.5$, let

$$\hat{\theta}_{jk}(x) = b_{0kjq} + b_{1jkq}x \tag{9}$$

denote the conditional median of Y given that $X = x$.

To mimic the Kulinskaya. et al. measure of effect size, a conditional robust measure of scale is needed. Here, an interquartile range is used that is rescaled to estimate the conditional standard deviation when the conditional distribution of Y, given that $X = x$, is normal. To elaborate, for $q = 0.25$ let

$$\hat{u}_{jk}(x) = b_{0kjq} + b_{1jkq}x \tag{10}$$

and for $q = 0.75$, let

$$\hat{v}_{jk}(x) = b_{0kjq} + b_{1jkq}x. \tag{11}$$

The conditional measure dispersion used here is

$$\hat{w}_{jk}(X) = \frac{\hat{v}_{jk} - \hat{u}_{jk}}{z_{0.75} - z_{0.25}}, \tag{12}$$

where $z_{0.75}$ and $z_{0.25}$ are the 0.75 and 0.25 quantiles of a standard normal distribution, respectively.

Typically, linear regression models assume that the error term is homoscedastic. Assuming homoscedasticity would make it possible to use a Winsorized variance as a conditional measure of scale. Note that $w_{jk(x)}$ provides a convenient method for avoiding this restriction, which is why it is used rather than a Winsorized variance.

Now consider two independent groups. For convenience, the two groups are taken to be the two levels of the second factor corresponding to the first level of the first factor. The immediate goal is to describe an analog of the KMS measure of effect size given by (3). The measure of effect size is estimated to be

$$\hat{\xi}_1(x) = \frac{\hat{\theta}_{11}(x) - \hat{\theta}_{12}(x)}{\hat{\zeta}_1(x)}, \tag{13}$$

where

$$\hat{\zeta}_1(x) = \frac{(1 - q_1)\hat{w}^2_{11(X)} + q_1\hat{w}^2_{12}(x)}{q_1(1 - q_1)}, \tag{14}$$

where $q_1 = n_{11}(n_{11} + n_{12})$. For the second level of the first factor, a measure of effect size is computed in the same manner and is labeled $\hat{\zeta}_2(X)$. The measure of an interaction is taken to be

$$\hat{\delta}(x) = \hat{\zeta}_1(x) - \hat{\zeta}_2(x). \tag{15}$$

Note that if the rows and columns of the design are interchanged, this will generally alter the value of $\hat{\delta}(x)$.

Now consider the goal of testing (5). Results on methods that have some similarity to the situation at hand (Wilcox, 2022b, c) suggest using a percentile bootstrap method. Often this approach performs well when dealing with robust estimators, but simulations indicated that this is not the case here, even when dealing with moderately large sample sizes. The actual level, when testing at the 0.05 level, was well below 0.05. Consequently, consideration was given to using a bootstrap estimate of the standard error of $\hat{\delta}(x)$, S, and assuming that

$$F = \frac{\hat{\delta}(x)}{S}. \tag{16}$$

has a standard normal distribution.

Let (Y_{ijk}, X_{ijk}) denote a random sample of size n_{jk} from levels j and k of the first and second factor, respectively ($j = 1, 2$; $k = 1, 2$; $i = 1, \ldots, n_{jk}$). The bootstrap estimate of the standard error (e.g., Efron & Tibshirani, 1993) is computed as follows. First, generate a bootstrap sample from each group. That is, for each j and k randomly sample with replacement n_{ijk} values from (Y_{ijk}, X_{ijk}), $i = 1, \ldots, n_{JK}$. Based on these bootstrap samples, compute the estimate of $\delta(x)$ yielding $\hat{\delta}^*(x)$. Repeat this process B times yielding $\hat{\delta}_1^*(x), \ldots, \hat{\delta}_B^*(x)$. The estimated squared standard error is

$$S^2 = \frac{1}{B-1}\sum(\hat{\delta}_b^*(x) - \bar{\delta}^*(x))^2, \tag{17}$$

where $\bar{\delta}^*(x) = \sum \hat{\delta}_b^*(x)/B$. Results in Efron (1987) suggest that generally, $B = 100$ suffices.

This approach performed well in simulations when all sample sizes are at least one hundred. But with relatively small sample sizes, the actual level was found to be less than 0.01 in some cases. Increasing B to 500 and even 1000 did not correct this problem. Some preliminary simulations revealed why. Consider, for example, the case where for all four groups, both X and Y have standard normal distributions. Further consider the case where the goal is to make inferences about $\delta(0)$. That is, the value of the covariate that is of interest is taken to be $X = 0$. For a common sample size of 40, a simulation was used to estimate $\delta(0)$ 3000 times and the standard deviation of the 3000 estimates yielded an estimate of S the true standard error of $\hat{\delta}(0)$, which was 0.206. For each replication the bootstrap estimate of the standard error was computed as well. Both the mean and median of the 3000 estimates

were equal to 0.227. That is, the bootstrap estimate of the standard was found to be biased, it over estimates the true value, which explains why the assumption that F has a standard normal distribution results in actual levels well below the nominal level.

The process just described was performed for the common sample sizes 20, 30, 40, 50, 75, 100 and 150, resulting in estimates of the true value of S and the expected value of the bootstrap estimate of the standard error. For each sample size, the ratio of the estimate of S and the mean estimate of the bootstrap estimate of the standard error was computed. Next, a regression line was fitted to these seven ratios and the sample size with the goal of determining a so that $E(S/a)$ is approximately unbiased. The results indicated that approximately, $a = 4.4/n+1.00087$, suggesting that S is asymptotically unbiased as expected. With a common sample size greater than 150, $a = 1$ is used. For unequal sample sizes, $4.4/n_{jk} + 1.00087$ is computed for each sample size and a is taken to be the average of the values.

Based on these results an adjusted test statistic is used:

$$F_a = a\frac{\hat{\delta}(x)}{S}. \tag{18}$$

A $(1 - \alpha)$ confidence interval for $\delta(x)$ is taken to be

$$\hat{\delta}(x) \pm z_{1-\alpha/2}\frac{S}{a}. \tag{19}$$

Of course, there is the issue of how well this approximate solution performs when dealing with non-normal distributions as well as situations where X and Y are dependent. These issues are addressed in section 3.

Note that $\hat{\delta}(x)$ can be computed for a range of x values yielding a curve that provides an additional perspective on the magnitude of the interaction as a function of the covariate. This is illustrated in section 4.

Another issue is how to control the familywise error (FWE) rate, meaning the probability of one or more Type I errors when testing (5) for multiple values for x. Here, results are reported when an improvement on the Bonferroni method derived by Hochberg (1988) is used. No advantage, in terms of Type errors, was found using instead the methods derived by Holm (1979) and Hommel (1986), so these other approaches are not reviewed here.

Hochberg's method is applied as follows. Let α denote the desired FWE rate. Let C denote the number of tests and let $p_{[1]} \geq \cdots \geq p_{[C]}$ denote the resulting p-values written in descending order. Set $k = 1$. If $p_{[k]} \leq \alpha$, reject all C hypotheses. If $p_{[k]} > \alpha$, proceed as follows.

1. Increment k by 1. If
$$p_{[k]} \leq \frac{\alpha}{k},$$
stop and reject all hypotheses having a p-value less than or equal $p_{[k]}$
2. If $p_{[k]} > \frac{\alpha}{k}$, repeat step 1.
3. Repeat steps 1 and 2 until a hypothesis is rejected or all C hypotheses have been tested.

3 Simulation Results

Simulations were used to assess how well F_a performs in terms of a Type I error. The sample sizes were taken to be $(n_{11}, n_{12}, n_{21}, n_{22}) = (20, 20, 20, 20)$, (50, 50, 50, 50), (20, 20, 50, 50), and (20, 20, 100, 100). Let Z denote a standard normal distribution. The distributions for both X and Y were taken to be one of four g-and-h distributions given by

$$\begin{cases} \frac{exp(gZ)-1}{g} exp(hZ^2/2), & \text{if } g > 0 \\ Zexp(hZ^2/2), & \text{if } g = 0 \end{cases} \tag{20}$$

A standard normal distribution corresponds to $g = h = 0$. The parameter g determines skewness and h determines how heavy-tailed the distribution happens to be. As h increases, the distribution becomes more heavy-tailed. Here, the choices for g and h are $(g, h) = (0, 0)$, standard normal; (0.0, 0.2), symmetric and heavy-tailed; (1.0, 0.0), skewed and light-tailed and (1.0, 0.2); skewed and heavy-tailed. Based on a summary of several studies aimed at characterizing the extent distributions differ from normality (Wilcox, 2022a, section 4.2), the distributions used here appear to span most distributions likely to be encountered in terms of skewness and kurtosis. Figure 1 shows these four distributions. Pearson's correlation between Y and X was taken to be $\rho = 0.0$ and 0.5.

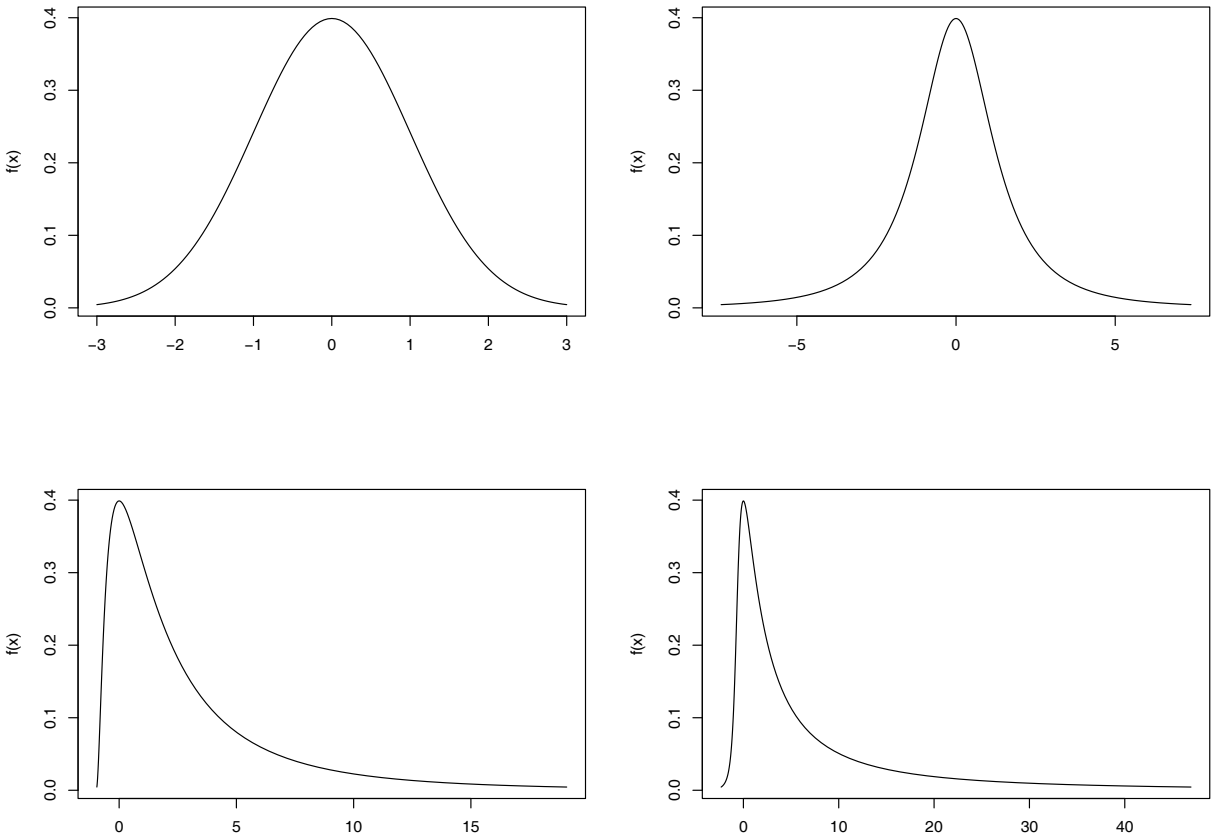

Figure 1: Distributions used in the simulations

Table 1: Estimated Type I errors, $\alpha = 0.05$, $x = 0$, $\rho = 0$

n	g	h	$\hat{\alpha}$
(20, 20, 20, 20)	0.0	0.0	0.043
	0.0	0.2	0.045
	1.0	0.0	0.044
	1.0	0.2	0.042
(50, 50, 50, 50)	0 0	0.0	0.044
	0.0	0.2	0.042
	1.0	0.0	0.053
	1.0	0.2	0.044
(20, 20, 50, 50)	0.0	0.0	0.038
	0.0	0.2	0.040
	1.0	0.0	0.032
	1.0	0.2	0.031
(20, 20, 100, 100)	0.0	0.0	0.025
	0.0	0.2	0.028
	1.0	0.0	0.029
	1.0	0.2	0.028

The first set of simulations used $x = 0$. The idea is that at a minimum, the method should perform well when x is taken to be some value near the center of the distributions.

Table 1 shows the results for $x = 0$ and $\rho = 0.0$. The results for $\rho = 0.5$ were very similar and are not reported for brevity. As can be seen, for equal sample sizes, the actual level is estimated to be very close to nominal level. The highest estimate among all results is 0.053. For unequal sample sizes, the actual level depends on how much the sample sizes differ as well as how small the smallest values happen to be. Here, the lowest estimate was 0.025, which occurred when $(n_{11}, n_{12}, n_{21}, n_{22}) = (20, 20, 100, 100)$. Note that given the sample sizes, the actual level does not depend very much on the nature of the distribution used to generate the data.

Next, simulations were run for two situations based on five choices for x. The first approach is based on choosing x in manner that avoids extrapolation. If, for example, $x = 4$

Table 2: Estimated Type I error probability, $\alpha = 0.05$ and $x = -1.28$

n	g	h	FWE	$x = -1.28$
20	0.0	0.0	0.015	0.010
	0.0	0.2	0.020	0.019
	1.0	0.0	0.028	0.041
	1.0	0.2	0.031	0.037
50	0.0	0.0	0.030	0.037
	0.0	0.2	0.033	0.043
	1.0	0.0	0.028	0.035
	1.0	0.2	0.033	0.026

is used, but for one of the groups there are no covariate values greater than 3, in effect there are no data that provide information about the nature of the association when $X > 3$. This concern is addressed by choosing three values for x as follows. Let $U_{jk} = \hat{x}_{jk,0.9}$ denote an estimate of the 0.9 quantile associated with the jth level of the first factor and the kth level of the second factor. And let $L_{jjk} = \hat{x}_{jk,0.1}$ denote the estimate of the 0.1 quantile. Let $L = \max(L_{11}, L_{12}, L_{21}, L_{22})$ and $U = \min(U_{11}, U_{12}, U_{21}, U_{22})$. The simulations used five values for x evenly space between L and U inclusive. The second approach used five choices for x evenly space between the 0.1 and 0.9 quantiles of a standard normal distribution. That is, the five values are between -1.28 and 1.28, inclusive. Results for the second situation, where $x = -1.28$, are shown in Table 2. Results for the first situation were very similar to the second situation, so they are not reported.

Bradley (1978) suggests that as a general guide, when testing at the 0.05 level, the actual level should be between 0.025 and 0.075. As can be seen, when the groups have a common sample size of 20, both the FWE rate and the probability of a Type I error when $x = -1.28$ do not satisfy Bradley's criterion when dealing with a symmetric distribution, the estimates are less than 0.025. For the skewed distributions, Bradley's criterion is satisfied. For a common sample size of 50, now Bradley's criterion is met for all of the situations considered.

Some additional simulations were run where $x = 2$, the idea is to make it likely that extrapolation occurs in one or more groups. This lowered somewhat the estimate of α.

Table 3: Results for CESD = 15, 17, 19 and 21

CESD	$\delta(\hat{x})$	p.value	ci.low	ci.up
15	−0.242	0.083	−0.516	0.032
17	−0.267	0.061	−0.546	0.0124
19	−0.292	0.054	−0.589	0.005
21	−0.317	0.056	−0.643	0.007

4 An Illustration

The method is illustrated with data dealing with the physical and emotional wellbeing of older adults. (The data are publicly available at https://osf.io/nvd59/quickfiles and stored in the file A1B1C.) The first factor is education. Level one consists of participants who did not complete high school and level two are those who did complete high school or higher. The second factor is the cortisol awakening response, which refers to the change in a participant's cortisol level upon awakening and measured again 30-45 minutes later. The two levels here are whether cortisol increases or decreases when measured the second time. This factor has been found to be related to measures of stress (e.g., Wilcox, 2022a, section 11.1.13). The dependent variable is a measure of meaningful activities and the covariate is a measure of depressive symptoms (CESD).

Figure 2 shows the estimates of $\delta(x)$ for a range of CESD scores. The plot suggests that as depressive symptoms increase, the interaction between education level and cortisol becomes more pronounced. Table 3 reports the result when testing (5) for CESD equal to 15, 17, 19 and 21. The results suggest that for the more depressed individuals, there is a more pronounced interaction, but arguably the strength of this result is not strongly compelling.

It is noted that if the covariate is ignored and a two-way ANOVA method based on the medians is used, the p-value for no interaction is 0.226. This was done via the percentile bootstrap method in Wilcox (2022a)

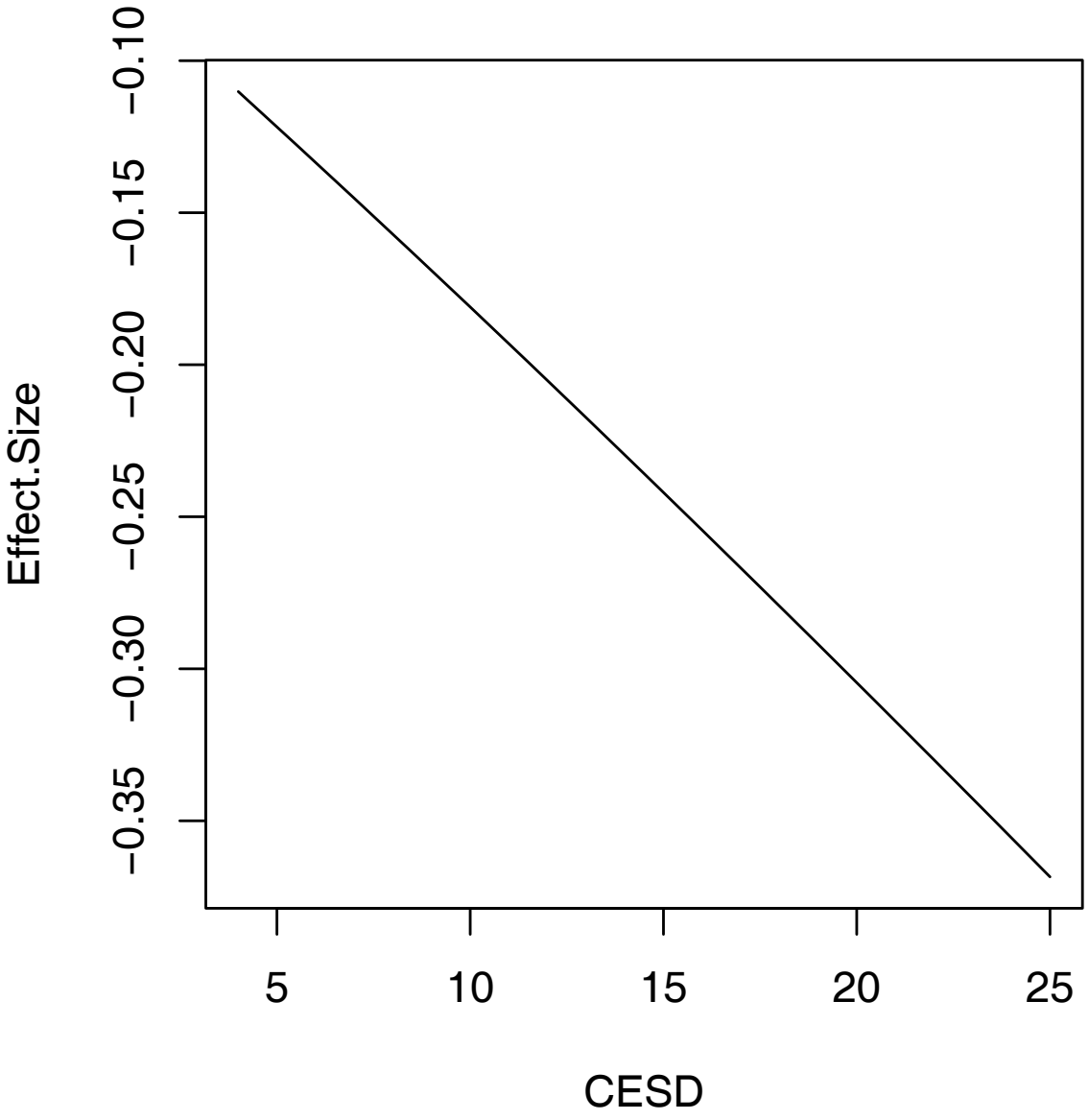

Figure 2: Effect size as a function of depressive symptoms, CESD

5 Concluding Remarks

In summary, when the covariate value is near the center of the distributions, all indications are that non-normality has at most a minor impact on the ability of the method used here to control the Type I error probability. The correlation between the dependent variable and the covariate also appears to have little or no impact on the Type I error probability. If the sample sizes are large, or small and approximately equal, the actual Type I error was found to be close to the nominal level. If some sample sizes are very small and others quite large, the actual level drops below the nominal but not by extreme amount.

When the covariate value is relatively far from the center of the distributions, the actual level can drop well below the nominal level when all sample sizes are equal to twenty. For sample sizes of at least fifty in all four groups, reasonably accurate control over the Type I error probability was obtained in the simulations.

There are related issues that might be addressed in future investigations. For example, computing an approximate $(1-\alpha)$ confidence band for the curve in Figure 2 would be useful. Another issue is how the method used here might be generalized to do deal with a J-by-K design where K, say, is greater than two. Of course, all relevant interactions can be estimated for any two levels of the first factor and any two levels of the second factor. But another approach would be to use some measure effect that reflects the overall extent the levels of second factor differ. That is, rather than use a measure effect size for each pair of groups, use a global measure of effect size that characterizes the extent the groups differ.

Finally, the R function t2way.KMS.interbt applies the method used here and is stored in the file Rallfun-v40, which can be downloaded from https://osf.io/xhe8u/. By default, the values for the covariate are taken to be L, $(L+U)/2$ and U. The values used for the covariate can be specified via the argument pts. The curve in Figure 2 was created with the function t2way.KMS.curve, which is stored in the file Rallfun-v40 as well. Both of these functions are being added to the R package WRS.

Disclosure statement: The author reports there are no competing interests to declare

Acknowledgement: The author thanks the reviewers for their constructive comments.

References

Benjamini, Y. Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57*, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x. https://www.jstor.org/stable/2346101.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics, 29*, 1165–1188. doi: 10.1214/aos/1013699998.

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

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association, 82*, 171–185.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap* New York: Chapman and Hall.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. New York: Academic Press.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70. https://www.jstor.org/stable/4615733.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*, 383–386. doi: 10.2307/2336190.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800–803. doi: 10.2307/2336325.

Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics*, 2nd Ed. New York: Wiley.

Jones, L. V. and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological Methods, 5*, 411–414. DOI: 10.1037/1082-989x.5.4.411

Kulinskaya, E., Morgenthaler, S. and Staudte, R. (2008). *Meta Analysis: A guide to calibrating and combining statistical evidence.* New York: Wiley. DOI: 10.1348/000711005X68174

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing.* New York: Wiley.

Steegen, S., Tuerlinck, F., Gelman, A. and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702–712.

Wilcox, R. R. (2022a). Introduction to Robust Estimation and Hypothesis Testing. 5th Edition. San Diego, CA: Academic Press.

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