A Heteroscedastic Analog of the Wilcoxon–Mann–Whitney Test When There Is a Covariate

Rand R. Wilcox

Correspondence: Dept of Psychology, University of Southern California, USA

Received: January 23, 2023 Accepted: February 25, 2023 Online Published: March 10, 2023

doi:10.5539/ijsp.v12n2p18 URL: https://doi.org/10.5539/ijsp.v12n2p18

Abstract

A basic method for comparing two independent groups is in terms of the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second group. The Wilcoxon–Mann–Whitney test is based on an estimate of this probability, but it uses an incorrect estimate of the standard error when the distributions differ. Numerous methods have been derived that are aimed at dealing with this issue. The goal here is to suggest a method for estimating this probability, given the value of a covariate. A well-known quantile regression estimator provides a way of dealing with this issue. The paper reports simulation results on how well this method performs.

Keywords: linear model, quantile regression estimator, bootstrap, robust effect size

There is now a range of methods for quantifying the extent two independent groups differ (e.g., Huberty, 2002; Grissom & Kim, 2012; Wilcox, 2022). Some are based on measures of location and others are a function of some measure of location in conjunction with some measure of scale. One of the more basic methods is P, the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second group, which is the focus here. This is not to suggest that P is inherently more interesting or important than other measures of effect size. But surely P provides a useful perspective that is easy to understand by non-statisticians.

As is well known, the classic Wilcoxon (1945) method, which is essentially the same method derived by Mann and Whitney (1947), is based in part on an estimate of P. However, the Wilcoxon–Mann–Whitney test does not provide a satisfactory method for making inferences about P. This stems from the fact that the standard error of the estimator that was used was derived assuming that the distributions are identical. If the distributions differ, an incorrect estimate of the standard error is used. Numerous methods have been proposed for dealing with this issue (Kotz. et al., 2003, Bruner et al., 2002). Twelve methods were compared by Ruscio and Mullen (2012).

Schacht et al. (2008) derived a method that deals with situations where there is some covariate, X. Their method is nonparametric, it makes no assumption about the nature of the underlying distributions. Moreover, the method deals with situations where there are multiple covariates, but for convenience, the method is reviewed when there is a single covariate.

Let Y_j denote the random variable of interest associated with the jth group ($j = 1, 2$) and let X_j denote some covariate. Let

$$P = P(Y_1 < Y_2) \tag{1}$$

and

$$Q = P(X_1 < X_2) \tag{2}$$

Schacht et al. suggest using an adjusted estimate of P based in part on an estimate of Q. The adjustment also depends on the covariances of the estimators for P and Q.

The goal here is to instead focus on $P(x)$, the value of P given that a covariate $X = x$. Note that estimating $P(x)$ for a range x values provides useful information beyond P or the method derived by Schacht et al. The approach used here assumes that a linear model can be used to estimate the quantiles of the distribution of Y given a value for the covariate. In this sense, the method proposed here is a based on a more restrictive assumption compared to the method derived by Schacht et al. But the proposed method provides a perspective that has the potential to provide a more nuanced understanding of the impact of the covariate.

Extant methods for making inferences about P are not readily extended to the problem of making inferences about $P(x)$. To elaborate, first consider the usual estimate of P based on a random sample Y_{ij} from the jth group ($i = 1, \ldots, n_j$; $j = 1$, 2). Let $K_{ik} = 1$ if $Y_{i1} < Y_{k,2}$, otherwise $K_{ik} = 0$. Then the typical estimate of P is

$$\hat{P} = \frac{1}{n_1 n_2} \sum \sum K_{ik}. \tag{3}$$

The difficulty is that when there is a covariate, what is needed is information about the distribution of Y given that $X = x$ in order to estimate $P(x)$. The goal in this paper is to suggest a method for dealing with this issue.

The paper is organized as follows. Section 2 describes a method for estimating $P(x)$ followed by a method for testing

$$H_0 : P(x) = 0.5, \tag{4}$$

as well as a method for computing a confidence interval for $P(x)$. Section 3 reports simulation results and section 4 illustrates the proposed method.

1. The Proposed Method

Here, a linear quantile regression model is assumed. That is, it is assumed that for the jth group, the *qth* quantile of Y_j, given that the covariate $X_j = x$, is

$$Y_{jq} = \beta_{0jq} + \beta_{1jq} x, \tag{5}$$

where β_{0jq} and β_{1jq} are unknown parameters.

Here, the unknown slopes and intercepts are estimated via the well-known estimator derived by Koenker and Bassett (1978). For convenience, let r_i denote the residuals associated with the first group. Then the Koenker–Bassett method estimates the slope and intercept with the values b_{0q} and b_{1q} that minimize

$$\sum \psi_q(r_i), \tag{6}$$

where

$$\psi_q(w) = w(q - I_{w<0}) \tag{7}$$

and the indicator function $I_{w<0} = 1$ if $w < 0$; otherwise $I_{w<0} = 0$ Of course, the slope and intercept for the second group are estimated in the same manner.

Note that in essence, the Koenker and Bassett (KB) estimator provides a method for estimating the conditional distribution of Y_j given that $X_j = x$. That is, estimating the quantiles of a distribution for a range of values for q provides an estimate of the conditional distribution of Y_j given that $X_j = x$. For convenience, let $u = 100q$. The estimate used here is taken to be

$$D_{uj}(x) = b_{0jq} + b_{1jq} x, \tag{8}$$

which is computed for $q = 0.01(0.01)0.99$.

A simple illustration helps motivate the proposed approach used here to make inferences about $P(x)$. Consider the situation where $Y_j = X_j + \epsilon$ where X_j has a standard normal distribution and ϵ has a lognormal distribution. So given that $X_j = x$, Y_j has a lognormal distribution with mean $x + \sqrt{(}\exp(1))$. A random sample of size $n = 200$ was generated based on this model and the conditional distribution of Y_j, given that $X_j = 0$, was estimated as just described. To demonstrate just how well the D values approximate the true distribution in this particular case, an adaptive kernel density estimate of the conditional distribution was used based on the D values. The solid line in Figure 1 shows the results. The dashed line is the true probability density function, which indicates a close agreement with the estimate based on the D values. Of course, this one illustration is not convincing evidence about the extent a reasonably accurate estimate will be obtained in general. It merely suggests that this approach might be useful for the situation at hand. Simulation results in the next section of this paper are aimed at addressing this issue.

Table 1. The extent estimators given by (3) and (9) give similar results

n	Estimator	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
20	(3)	0.1087	0.4207	0.5007	0.5022	0.5804	0.9751
20	(9)	0.2000	0.4375	0.5050	0.5035	0.5675	0.8425
1000	(3)	0.4408	0.4898	0.5004	0.5005	0.5116	0.5510
1000	(9)	0.4505	0.4920	0.5002	0.5002	0.5086	0.5447

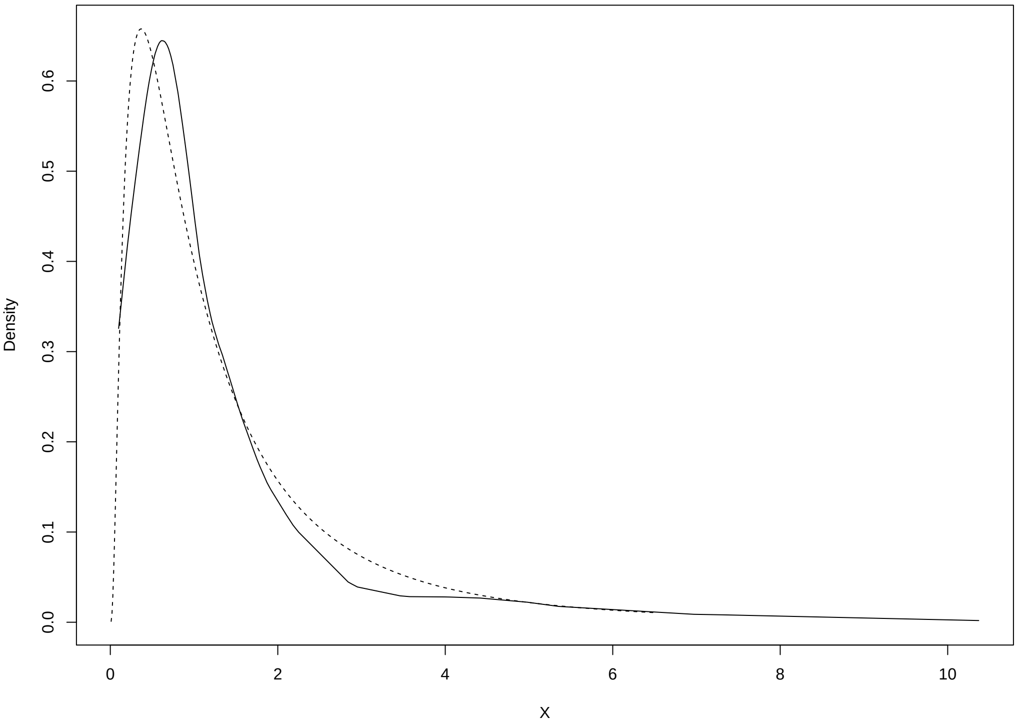

Figure 1. The solid line is an estimate of the conditional distribution of Y given that $X = 0$. The dashed line is the true distribution, which is lognormal

Let the indicator function $I_{uv}(x) = 1$ if $D_{u1}(x) < D_{v2}(x)$; otherwise $I_{uv}(x) = 0$. Then an estimate of $P(x)$ is simply

$$\hat{P}(x) = \frac{1}{99}\frac{1}{99}\sum_{u=1}^{99}\sum_{v=1}^{99} I_{uv}(x). \tag{9}$$

To add some perspective on the nature of $\hat{P}(x)$, consider the case where X and Y are independent. Then for any x, $\hat{P}$ given by (3) and $\hat{P}(x)$ are estimating the same quantity. To see how these two estimators compare, a simulation was run where both X and Y have standard normal distributions. Two choices for x were used: 0 and 0.67 (the 0.75 quantile of a standard normal distribution). The results were virtually the same for both of these situations, so only results for $x = 0.67$ are reported. The sample sizes were taken to be $n_1 = n_2 = 20$ and 1000. Table 1 reports the extent these two estimators give similar values based on 2000 replications. As can be seen, there is a very close agreement between these two estimators for small sample sizes, and an even closer agreement for very large sample sizes. The only exception is the two minimum values when $n_1 = n_2 = 20$. Similar results were obtained when 0.8 is added to Y_2, in which case P is approximately 0.72.

The proposed method for testing (4) and computing a confidence interval for $P(x)$ assumes that $\hat{P}(x)$ has a normal distribution. As a partial check on this assumption, a Q-Q plot for $\hat{P}(x)$ was created for $n_1 = n_2 = 20$ and 1000 for the points $x = 0$ and 0.67. Again 0.8 is added to Y_2 to get some sense about the distribution when $P(x)$ differs from 0.5. Figure 2 shows the resulting plot for $n_1 = n_2 = 20$, which indicates a reasonably close agreement with a normal distribution. The plot for $n_1 = n_2 = 1000$, not shown here, indicates an even closer agreement with a normal distribution.

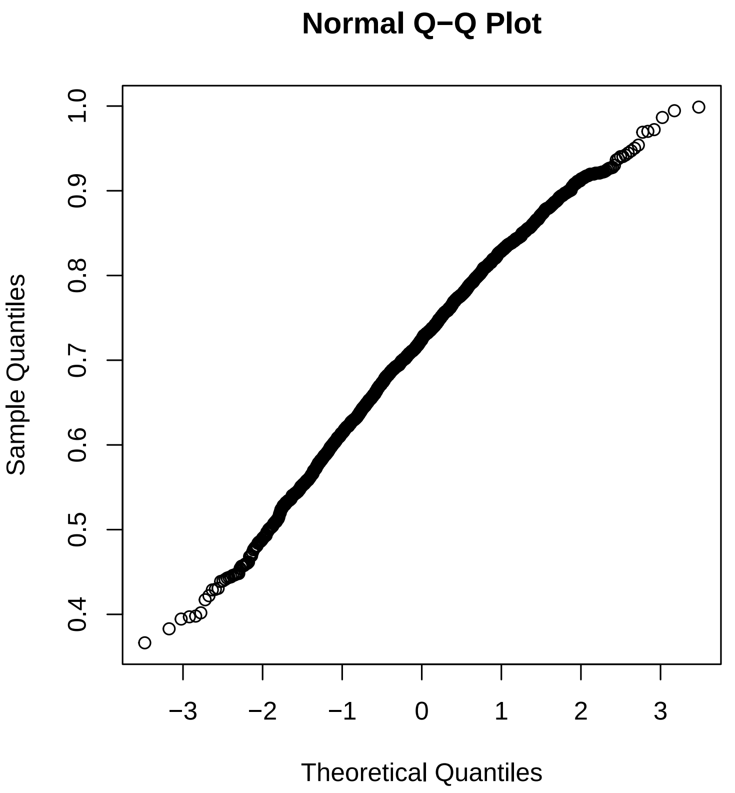

Figure 2. A Q-Q plot based on the sampling distribution of $\hat{P}(x)$, $n_1 = n_2 = 20$ and $x = 0.67$

However, testing (4) based on the D values is invalid because they are dependent. Here, this point is illustrated using the method derived by Cliff (1996), which is one of the methods that performed relatively well in the study by Ruscio and Mullen (2012). Simulations indicate that when the groups have a common sample size of 40, the actual Type I error probability, when testing at the 0.05 level, is 0.231. As the sample size increases, the actual level decreases. For a common sample of 200, the actual level is 0.006.

Here, a bootstrap method is used to estimate the standard error of $\hat{P}(x)$. The method begins by generating bootstrap samples from both groups. More precisely, assume the data for the jth group are stored in a matrix with two columns and n_j rows. A bootstrap sample consists of sampling with replacement n_j rows of data from this matrix. Based on these bootstrap samples, compute $\hat{P}(x)$ and for notational convenience label the result V^*. Repeat this process B times yielding $V_1^*, \ldots, V_B^*$. Then an estimate of the squared standard error of $\hat{P}(x)$ is

$$\hat{\eta}^2 = \frac{1}{B-1} \sum (V_b^* - \bar{V}^*)^2, \tag{10}$$

where $\bar{V}^* = \sum V^*/B$.

The test statistic for testing (4) is

$$W = \frac{\hat{P}(x) - 0.5}{\hat{\eta}}, \tag{11}$$

which is assumed to have a standard normal distribution when the null hypothesis is true. As is evident, still assuming normality, a $1 - \alpha$ confidence interval is given by

$$\hat{P}(x) \pm z_{1-\alpha/2}\hat{\eta}, \tag{12}$$

where $z_{1-\alpha/2}$ quantile of a standard normal distribution.

Here, $B = 100$ is used for two reasons. First, results in Efron (1987) suggest that $B = 100$ suffices. This has proven to be the case for a range of situations where a bootstrap estimate of the standard error is used (Wilcox, 2022). The second reason for using $B = 100$ has to do with execution time when performing simulations: it was found to be extremely high due to having to use the quantile regression estimator with every bootstrap sample. Consider, for example, $n_1 = n_2 = 50$ with $B = 100$. Despite taking advantage of a four-core processor via the R package parallel, execution time is approximately 36 seconds using a MacBook Pro with 2.9 GHz processor. Consequently, a simulation based on 1000 replications requires about ten hours of execution time when using the bootstrap estimate of the standard error. Increasing B to 1000 would require a little over four days of execution time. Increasing the number of replications to 10000, now the execution time is approximately 40 days.

A way of possibly avoiding the assumption that the null distribution of W is standard normal is to use a percentile bootstrap method instead. Typically percentile bootstrap methods are based on $B \geq 500$ bootstrap samples. But now, execution time is an even more serious concern. Switching to the percentile bootstrap method with $B = 500$ with 1000 replications would require about fifty hours of execution time. For this reason, using a percentile bootstrap method was excluded from consideration.

2. Simulation Results

The goal in this section is to gain some sense of how well the proposed method for testing (4) controls the probability of a Type I error when testing at the 0.05 level. Simulations, based on 1000 replications, were run by first generating data according to the model

$$Y_j = X_j + c_j\epsilon, \tag{13}$$

where $c_1 = c_2 = 1$, homoscedasticity, or $(c_1, c_2) = (1, 4)$, heteroscedasticity, were used. Both X_j and ϵ were generated from one of four distributions, which are shown in Figure 3. The upper left panel is a standard normal distribution. The other three are g-and-h distributions Letting Z denote a random variable having a standard normal distribution,

$$M = \begin{cases} \frac{exp(gZ)-1}{g} exp(hZ^2/2), & \text{if } g > 0 \\ Zexp(hZ^2/2), & \text{if } g = 0 \end{cases} \tag{14}$$

has a g-and-h distribution (Hoaglin, 1985), where g and h are parameters that determine the first four moments. In the upper right corner of Figure 3, $g = 0.0$ and $h = 0.2$, which is relatively heavy-tailed. In the lower left panel, $g = 1.0$ and $h = 0$, which is a lognormal distribution that has been shifted to have a median of zero. In the lower right panel $g = 1.0$ and $h = 0.2$. These four distributions were chosen based on a survey of papers that report the extent distributions differ from a normal distribution (Wilcox, 2022, section 4.2). This survey suggests that these distributions appear to span what is generally realistic in terms of skewness and kurtosis.

When $c_2 = 4$ and skewed distributions are involved, it is noted that the null hypothesis is no longer true, $P(x)$ differs slightly from 0.5. The method for generating the data given by (13) must be altered slightly so that the null hypothesis is true. The method for adjusting the data so that the null hypothesis is true was as follows. Let $\mathcal{D} = Y_1 - Y_2$ and note that $P(Y_1 < Y_2) = 0.5$ corresponds to the situation where the median of $\mathcal{D}$ is zero. Based on sample sizes of 10,000 for both groups, an adjustment was determined so that the median of $\mathcal{D}$ is zero, in which case $P(Y_1 < Y_2) = 0.5$. Now the data are generated according to the model

$$Y_j = X_j + c_j\epsilon + \delta. \tag{15}$$

When $(g, h) = (1, 0)$, $\delta = 0.33$ was used and for $(g, h) = (1, 0.2)$, $\delta = 0.36$.

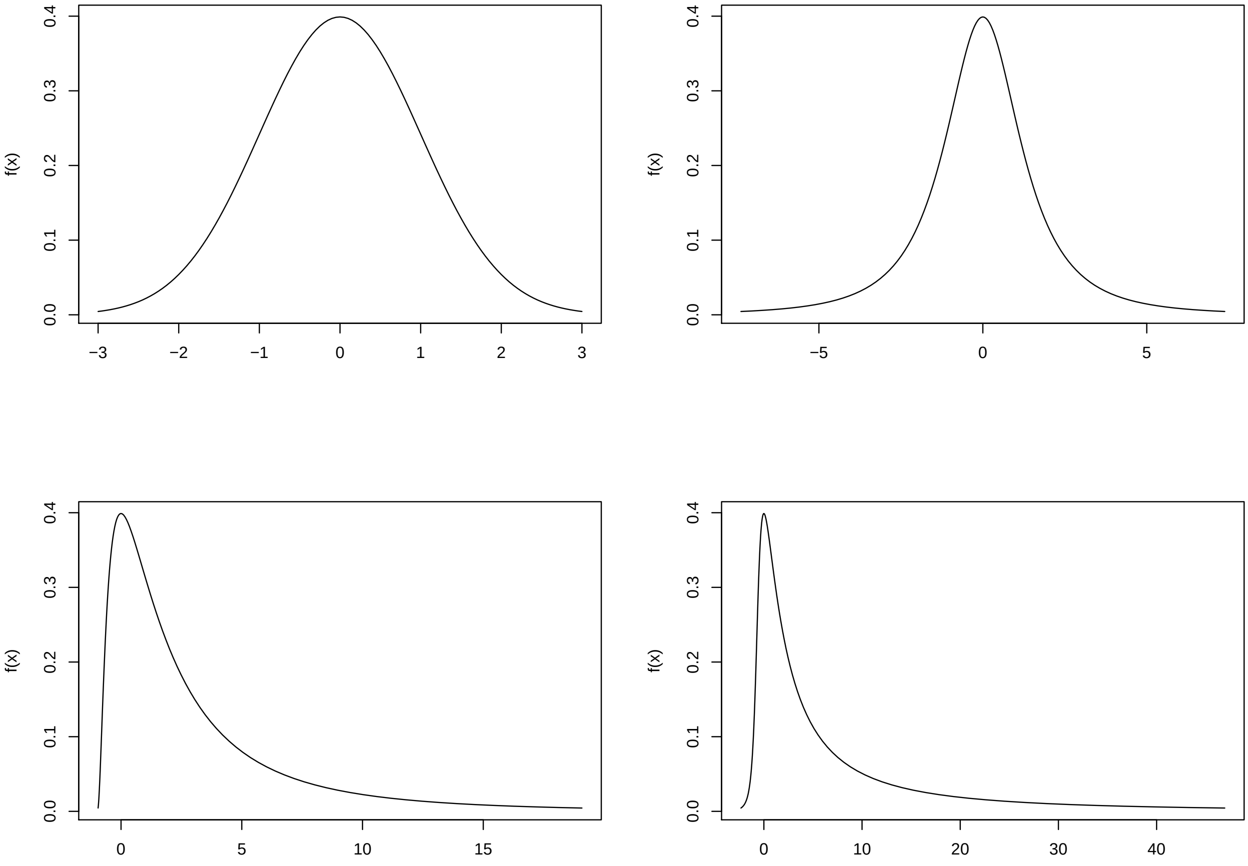

Figure 3. Distributions used in the simulations. Upper left is standard normal. Upper right is a symmetric, heavy-tailed distribution. Lower left is lognormal, shifted to have a median equal to zero. Lower right is a skewed, heavy-tailed distribution

There is the issue of choosing x. Imagine that for the first group, the covariate values range between 10 and 20, while for the second covariate they range between 15 and 22. One could assume that even though there are no data for the second group where the covariate is say 12, the linear model for estimating the quantiles of Y, given that $X = 12$ will provide reasonably accurate results. But this seems problematic at best. Moreover, a speculation was that if x is unusually large or small relative to the bulk of the covariate values, the proposed method would perform poorly, and preliminary simulations confirmed that this was the case.

For example, suppose x is taken to be the maximum observed value among the covariate values for the first group. For $n_1 = n_2 = 20$, $g = h = 0$ (normality) and $c_1 = c_2 = 1$ (homoscedasticity), the estimated Type I error probability, when testing at the 0.05 level, was 0.12. Consequently, when performing a simulation, x was chosen to be well within the range of the observed values. Here, two values were used, which were determined as follows.

Let $U_j = \hat{x}_{j,0.8}$ denote an estimate of the 0.8 quantile associated with the jth group. And let $L_j = \hat{x}_{j,0.2}$. Let $L = \max(L_1, L_2)$ and $U = \min(U_1, U_2)$. The first choice for x was $(L + U)/2$ and the second choice was U. The sample sizes were taken to be (n_1, n_2) = (20, 20), (50, 50) and (50, 100).

The results for the homoscedastic case are shown Table 2 and results for the heterosceastic case are shown in Table 3. 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, all of the results satisfy this criterion. A few additional simulations were run with (n_1, n_2) = (20, 40) and (n_1, n_2) = (200, 200). The results were consistent with those reported in Tables 2 and 3. Bradley also suggests that ideally, the actual level should be between 0.045 and 0.055. Even this more restrictive criterion is met in

Table 2. Estimated Type I errors, homoscedastic case, $\alpha = 0.05$

n_1	n_2	g	h	(L+U)/2	U
20	20	0.0	0.0	0.052	0.064
		0.0	0.2	0.051	0.059
		1.0	0.0	0.050	0.043
		1.0	0.2	0.052	0.047
50	50	0.0	0.0	0.053	0.053
		0.0	0.2	0.052	0.048
		1.0	0.0	0.049	0.044
		1.0	0.2	0.050	0.042
50	100	0.0	0.0	0.053	0.047
		0.0	0.2	0.055	0.045
		1.0	0.0	0.056	0.048
		1.0	0.2	0.056	0.054

Table 3. Estimated Type I errors, heteroscedastic case, $\alpha = 0.05$

n_1	n_2	g	h	(L+U)/2	U
20	20	0.0	0.0	0.041	0.070
		0.0	0.2	0.042	0.062
		1.0	0.0	0.042	0.043
		1.0	0 .2	0.047	0.040
50	50	0.0	0.0	0.050	0.056
		0.0	0.2	0.048	0.052
		1.0	0.0	0.064	0.053
		1.0	0.2	0.062	0.050
50	100	0.0	0.0	0.048	0.057
		0.0	0.2	0.047	0.050
		1.0	0.0	0.051	0.047
		1.0	0.2	0.050	0.045

most situations. When both sample sizes are at least 50, the lowest estimate is 0.042 and the highest estimate is 0.064. For (n_1, n_2) = (20, 20) the lowest estimate is 0.040 and the highest is 0.070.

There are many robust methods for comparing regression lines beyond the approach used here (e.g., Wilcox, 2022). A natural issue is how the method proposed here compares to these other techniques in terms of power. The answer is fairly evident: it depends on the nature of unknown distributions being compared simply because different methods are sensitive to different features of the data. The only certainty is that different methods can yield substantially different p-values. This is illustrated in the next section. Nevertheless, some power comparisons might help.

Consider the case where both slopes are zero, the intercept for the first group is zero and the intercept for the second group is 0.5. Under normality and when $x = 0$, the power of the proposed method was estimated to be 0.32. For the situation at hand it suffices to compare the intercepts. This was done using the robust heteroscedastic method in Wilcox (2022, section 11.12.1). Briefly, the method uses the robust regression estimator derived by Theil (1950) and Sen (1968) coupled with a percentile bootstrap method. Now the power was estimated to be 0.16.

Now suppose the covariate is ignored and the method derived by Cliff (1996) is used to test $H_0 : P = 0.5$. Power was estimated to be 0.31. However, when x is taken to be U, the power of the proposed method drops to 0.24. But regardless of any power issues, what seems more important is that the proposed method provides the potential of a more detailed indication of how groups compare as demonstrated in the next section.

3. An Illustration

The proposed method is illustrated based on two groups, both of which consist of older adults. The first group consisted of participants who did not complete high school. The second group is participants who completed four years of college. The sample sample sizes are 135 and 48, respectively. The dependent variable is a measure of life satisfaction (LSIZ) and the covariate is a measure of depressive symptoms (CESD).

Figure 4 shows the 0.5 quantile regression lines for the two groups. The solid line corresponds to participants who did not complete high school. Points indicated by a + correspond to participants who completed four years of college. The regression lines cross suggesting the possibility that for high CESD scores, typical LSIZ scores for group 1 are higher than for group 2. However, testing (4) for CESD values 30, 35 and 40, the corresponding p-values are 0.400, 0.303 and 0.245, respectively. In contrast, for CESD values 6, 8 and 10, the p-values are 0.003, 0.009 and 0.047, respectively. The corresponding estimates of $P(x)$ are 0.657, 0.626 and 0.598. That is, the results indicate that among participants with low CESD scores, participants who did not complete high school are more likely to report lower life satisfaction than participants who completed four years of college. In contrast, comparing the conditional medians corresponding to CESD values 6, 8 and 10, using the bootstrap method in Wilcox (2022, section 12.1), the p-values range between 0.218 and 0.377. That is, different methods are sensitive to different features of the data, so not surprisingly the choice of method can make a practical difference in the conclusions reached.

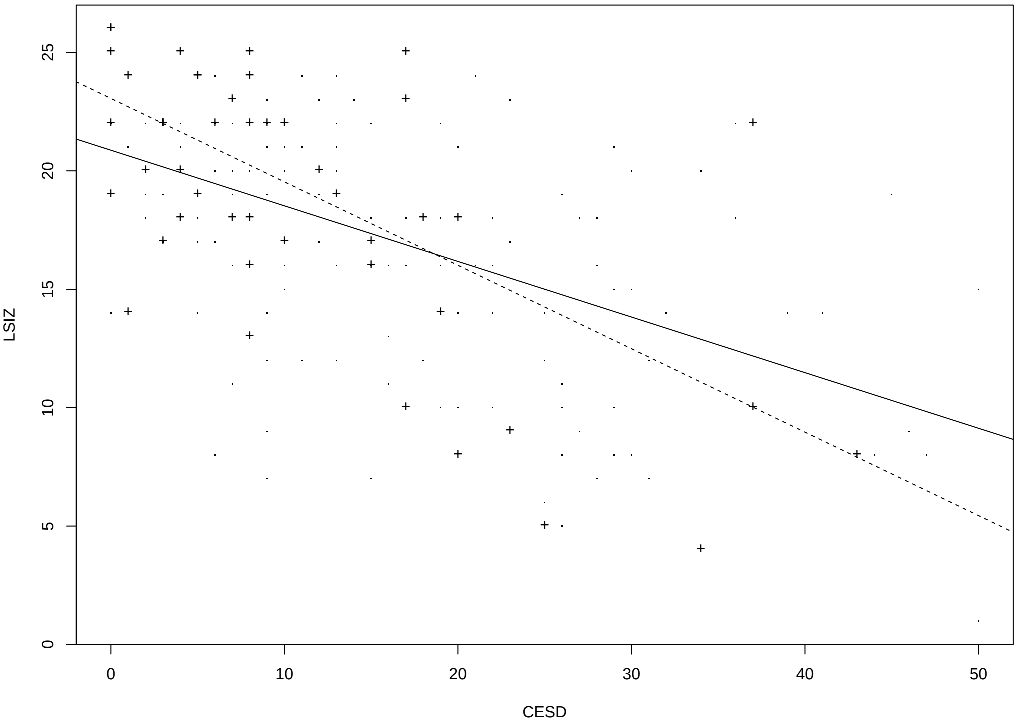

Figure 4. Quantile regression lines. The solid line corresponds to participants who did not complete high school. Points indicated by a + are participants who completed four years of college

If the covariate is ignored and (4) is tested with the Brunner and Munzel (2000) method, the estimate of P is 0.655 and the p-value is 0.003. However, consider Figure 5, which shows a plot of $\hat{P}(x)$ for twenty values of x (CESD scores) ranging from 5 to 23. As can be seen, the plot indicates that as depressive symptoms increase, the estimates of $P(x)$ decrease. CESD scores greater than 15 are often taken to indicate mild depression. For a CESD score equal to 16, $\hat{P}(x) = 0.52$. That is, the more detailed information revealed by $P(x)$ provides a useful perspective that is missed when the covariate is ignored.

4. Concluding Remarks

Steegan et al. (2016) point out and illustrate that multiple perspectives can be needed to get a deep and relatively nuanced understanding of data. $P(x)$ is one way of achieving this goal. Put another way, it is not being suggested that $P(x)$ is a a better measure of effect size than other measures of effect size that might be used. Other measures of effect size, such as the difference between measures of location, are obviously sensitive to different features of the data and provide alternative perspectives that have practical value.

All indications are that the proposed method is asymptotically correct. But a limitation of this study is that there is no formal proof that this is the case. Such a proof needs to take into account the dependence among the D values.

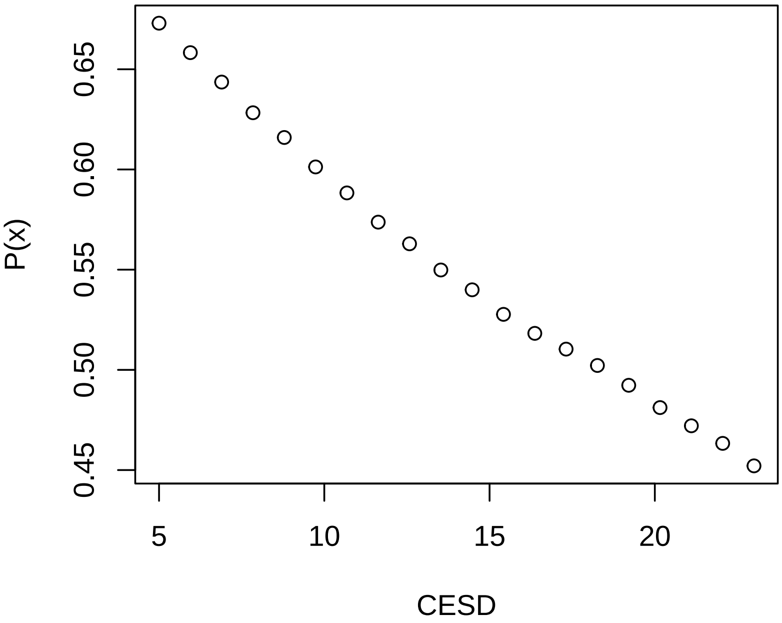

Figure 5. Estimates of $P(x)$ for CESD scores ranging from 5 to 23

Note that (9) is based on 99*2=198 parameters for each group. But in the illustration in section 4, the sample sizes are 135 and 48, which at some level might seem concerning because the number of parameters being estimated exceeds the sample sizes. Note, however, that this situation has a close similarity to the sample mean. Consider, for example, any random variable X that has a finite sample space having cardinality N. Of course, from a theoretical point of view it is convenient to assume that X is continuous with an infinitely large sample space. But typically, the reality is that N is finite albeit possibly quite large. For example, in a weight gain study, the possible outcomes are limited to the precision of the scale that is used. Let $X_1, \ldots, X_N$ denote the sample space in which case the population mean is

$$\mu = \sum X_i p_i, \tag{16}$$

where p_i is the probability that X is equal to X_i. In effect, the sample mean is based on an estimate of the N unknown parameters, $p_1, \ldots, p_N$. Letting n denote the sample size, often n is much smaller than N, the number of parameters in (16). Of course, there are situations where the sample mean performs poorly. The only point is that there are situations where, despite the number of parameters being estimated, the sample mean performs well. All indications are that for the situation at hand, the proposed estimator also performs well despite the number of parameters being estimated.

Presumably the confidence interval for $P(x)$ breaks down when $P(x)$ is close to zero or one. At the moment, studying this issue is difficult due to the execution time issue previously noted. Hopefully, future studies will be able to address this issue.

The proposed method is not limited to a single covariate. But it remains to be determined whether the method continues to perform well when there is more than one covariate..

Finally, the R function wmw.ancbse applies the proposed method and is stored in the file Rallfun-v40, which can be downloaded from https://osf.io/xhe8u/. Figure 5 was created by the R function wmw.anc.plot.

Disclosure Statement

The author reports there are no competing interests to declare

References

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

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and small-sample approximation. *Biometrical Journal, 42*, 17–25.

Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric Analysis of Longitudinal Data in Factorial Experiments*. New York: Wiley.

Cliff, N. (1996). *Ordinal Methods for Behavioral Data Analysis*. Erlbaum, Mahwah, NJ.

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

Hoaglin, D. C. (1985). Summarizing shape numerically: the g-and-h distribution. In: Hoaglin, D., Mosteller, F., Tukey, J. (Eds.), *Exploring Data Tables Trends and Shapes*. New York: Wiley, pp. 461–515.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrika, 46*, 33–50.

Kotz, S., Lumelskii, Y., & Pensky, M. (2003). *The Stress-Strength Model and Its Generalizations*. World Scientific Publishing Company.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18*, 50–60.

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research, 47*, 201–223.

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

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

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1*, 80– 83.

Copyrights