International Journal of Statistics and Probability; Vol. 11, No. 2; March 2022
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Inferences About a Quantile Shift Measure of Effect Size When There Is a Covariate

Rand R. Wilcox

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

Received: January 4, 2022 Accepted: February 15, 2022 Online Published: February 17, 2022

doi:10.5539/ijsp.v11n2p52 URL: https://doi.org/10.5539/ijsp.v11n2p52

Abstract

When comparing two independent groups, a possible appeal of the quantile shift measure of effect size is that its magnitude takes into account situations where one or both distributions are skewed. Extant results indicate that a percentile bootstrap method performs reasonably well given the goal of making inferences about this measure of effect size. The goal here is to suggest a method for making inferences about this measure of effect size when there is a covariate. The method is illustrated with data dealing with the wellbeing of older adults.

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

1. Introduction

Consider two independent groups having unknown distributions. Here, the first group is viewed as a control group and the other group is an experimental group. Let δ denote some parameter that characterizes how the distributions differ. There is now a wide range of choices for δ with each providing a different perspective on how the groups compare (e.g., Huberty, 2002; Grissom & Kim, 2012; Wilcox, 2022b).

Note that the median of the experimental group corresponds to the Qth quantile of the control group. That is, Q reflects the extent the median of the experimental group is unusual relative to the control group and is generally known as a quantile shift measure of effect size. A possible appeal of this measure of effect size is that its relative magnitude takes into account whether one or both distributions are skewed. Extant results indicate that a reasonably accurate confidence interval for Q can be computed via a percentile bootstrap method (e.g., Wilcox, 2022b). However, when there is a covariate, there are no results on how to proceed. The goal here is to suggest a method for making inferences about Q, given a value for some covariate, followed by a simulation study that deals with how well the proposed method performs.

To review the motivation for Q as well as some of its properties, first consider the situation where there is no covariate. To begin, let θ_j and τ_j denote some measure of location and scale, respectively, associated with the jth group ($j = 1, 2$). Certainly the most common approach to comparing two distributions is to take the measure of location θ_j to be the population mean or median and to view the measure of scale, τ_j, as a nuisance parameter. More formally use $\delta = \theta_1 - \theta_2$ to characterize how the groups differ and test

$$H_0 : \theta_1 = \theta_2 \tag{1}$$

or compute a confidence interval for $\theta_1 - \theta_2$.

Another general approach is to use a measure of effect size that takes into account both measures of location and some measure of variation. Broadly, this approach uses

$$\delta = \frac{\theta_1 - \theta_2}{f(\tau_1, \tau_2)}, \tag{2}$$

where $f(\tau_1, \tau_2)$ is some function of τ_1 and τ_2 to be determined. Seemingly, the best-known version of (2) is where $\theta_j = \mu_j$, the population mean, $\tau_j = \sigma_j$, the population standard deviation, and by assumption $\sigma_1 = \sigma_2 = \sigma$ (homoscedasticity), in which case (2) becomes

$$\Delta = \frac{\mu_1 - \mu_2}{\sigma}. \tag{3}$$

A common practice (e.g., Cohen, 1988) is to view $\Delta = 0.2$, 0.5 and 0.8 as being small, medium and large, respectively. Presumably, what constitutes a large effect size can depend on the situation. However, for illustrative purposes, Cohen's suggestion is assumed henceforth.

There are two basic concerns with Δ. First, it assumes homoscedasticity. Kulinskaya et al. (2008) derived a heteroscedastic measure of effect size given by

$$\delta_{\text{kms}} = \frac{\mu_1 - \mu_2}{\varsigma}, \tag{4}$$

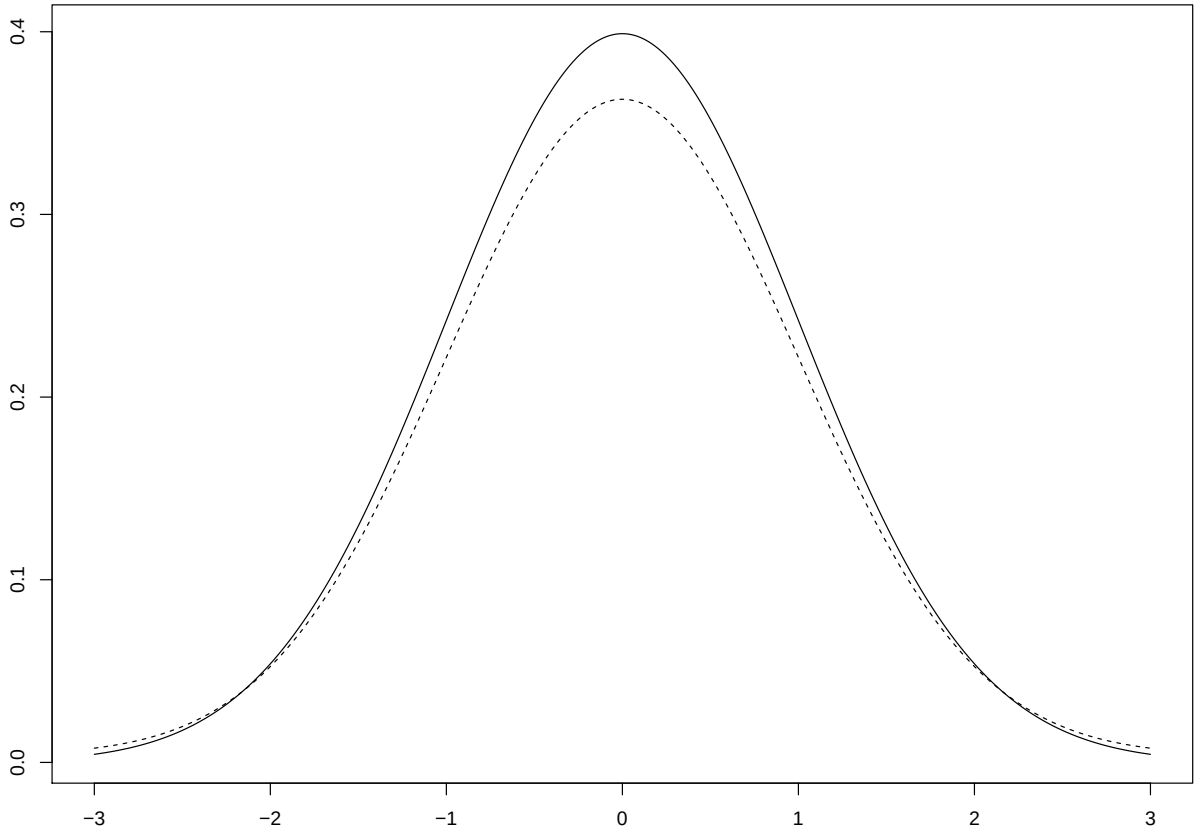

Figure 1. The solid line is a standard normal distribution, $\sigma^2 = 1$. The dashed line is a mixed normal distribution, $\sigma^2 = 10.9$

where

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

$q = n_1/N$, $N = n_1 + n_2$ and n_j are the sample sizes. Wilcox (2022a) reports results using this measure of effect size when dealing with an interaction in a two-way design.

The second concern is that Δ is not robust (e.g., Algina et al., 2005), roughly meaning that even a small departure from normality can alter its value substantially. To be a bit more precise, the standard deviation is not robust (e.g., Hampel et al, 1986; Huber & Ronchetti, 1990; Staudte & Shearer, 1986). It is highly sensitive to the tails of a distribution, the result being that even a slight departure from a normal distribution has the potential of lowering Δ substantially. In particular, a large effect among the bulk of the participants can appear to be small when using Δ.

Following Algina et al. (2005), this issue is illustrated with the mixed normal distribution discussed by Tukey (1960). Its cumulative distribution function (cdf) is given by

$$H(x) = 0.9\Phi(x) + 0.1\Phi(x/10), \tag{5}$$

where $\Phi(x)$ is the cdf of a standard normal distribution. Figure 1 shows a plot of the standard normal and this mixed normal distribution. As is evident, the two distributions appear to be very similar. However, while the standard normal has variance one, the variance of the mixed normal is 10.9.

Now look at Figure 2. In the left panel, are two normal distributions with variance one. The means are 0 and 0.8, so $\Delta = 0.8$, which Cohen characterizes as large. In the right panel are two mixed normals again with means 0 and 0.8.

Now $\Delta = 0.8/\sqrt{10.9} = 0.24$, which is relatively small. Algina et al. (2005) deal with this issue by replacing the mean and variance in (3) with a 20% trimmed mean and Winsorized variance, which is rescaled to estimate the variance when dealing with a normal distribution. A similar modification of δ_{kms} is straightforward. These methods help deal with heavy-tailed distributions such as the mixed normal, but there is an inherent assumption that the distributions are symmetric.

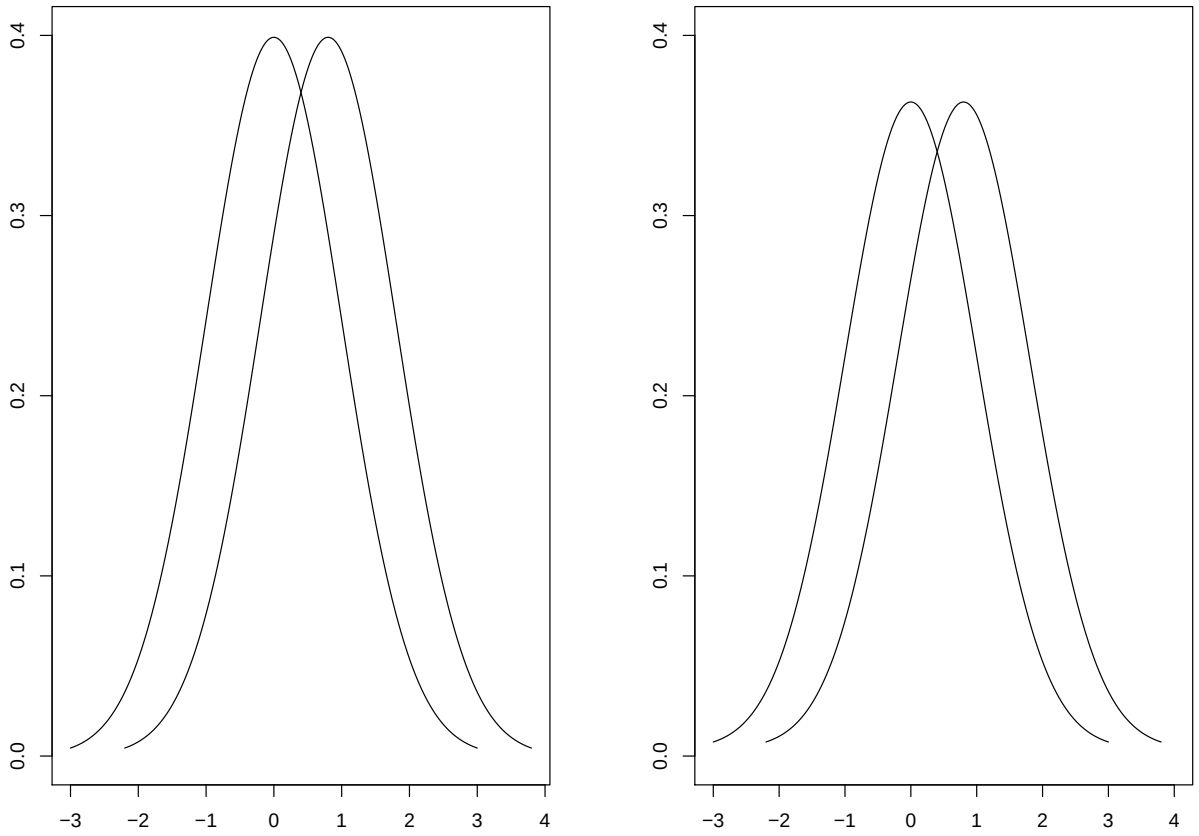

Figure 2. A slight departure from a normal distribution can substantially lower Δ, masking a large effect among the bulk of the participants. The left panel shows two normal distributions where $\Delta = 0.8$. The right panel shows to mixed normals where $\Delta = 0.24$

To underscore some concerns when dealing with skewed distributions, it helps to first note that under normality, $\Delta = 0.2$ indicates that the mean of the experimental group corresponds to the 0.42 quantile of the control group. That is, the experimental group shifts the mean of the control from the $q_1 = 0.5$ quantile to the $q_2 = 0.42$ quantile. Let $\delta_q = q_1 - q_2$, which captures the spirit of a standardized difference, Δ, without imposing any parametric family of distributions. Given that $\Delta = 0.2$ is viewed as a small effect size when dealing with normal distributions, it follows that $\delta_q = 0.08$ is considered small as well. In a similar manner, if $\Delta = 0.5$ and 0.8 are considered medium and large effect size under normality, respectively, this means that $\delta_q = 0.19$ and $\delta_q = 0.29$ are considered medium and large effect size as well.

Wilcox (2022b, section 5.3.4) describes possible concerns about skewed distributions when using Δ or some robust, heteroscedastic version of Δ. Note, for example, that in terms of magnitude, there is no distinction between $\Delta = 0.5$ and $\Delta = -0.5$. Both would be viewed as a median effect size. But consider the situation where the control group has a lognormal distribution, which has mean equal to 1.65, which is the $q_1 = 0.69$ quantile. Suppose the experimental group has a lognormal distribution that has been shifted to have mean θ_2 which is the q_2 quantile associated with the control group. Of course, when the means are equal, $\Delta = \delta_q = 0$. But consider the case where $\Delta = 0.5$. This corresponds to shifting the mean from about the 0.69 quantile to the 0.29 quantile. So $\delta_q = 0.4$, suggesting a very large effect size rather than a medium effect size as suggested by Δ. It is readily verified that the reverse can happen where δ_q suggests a small effect size in contrast to Δ. This same concern occurs for any measure of effect size that implicitly assumes that the

distributions are symmetric.

One way of dealing with this concern in a robust, non-parametric manner is to first take θ_1 and θ_2 to be the population medians of the control group and the experimental group, respectively. Let Y_j denote some random variable of interest associated with the jth group and let

$$Q = P(Y_1 \leq \theta_2). \tag{6}$$

That is, θ_2, the median of the experimental group, is the Qth quantile of the control group. Following Wilcox (2022b), Q is taken to be a measure of effect size. The further Q is from 0.5 the larger the effect. Under normality and homoscedasticity, $\Delta = 0.2$, 0.5 and 0.8 correspond to $Q = 0.58$, 0.69 and 0.79, respectively.

Now consider the situation where there is a covariate X and let $Q(x)$ denote the value of Q given that $X = x$. Section 2 of this paper suggests a method for estimating $Q(x)$. Included is a proposed method for testing

$$H_0 : Q(x) = 0.5, \tag{7}$$

no effect, as well as a method for computing a $1 - \alpha$ confidence interval for $Q(x)$. Section 3 reports the results of a simulation study. Finally, the method is illustrated with data dealing with the physical and emotional wellbeing of older adults.

It is noted that testing (7) is open to the criticism that surely $Q(x)$ differs from 0.5 at some decimal place (Tukey, 1991). Assuming this view is reasonable, the goal is not to test (7), but rather determine the extent it is reasonable to make a decision about whether $Q(x)$ is less than or greater than 0.5 (Jones & Tukey, 2000). From this point of view, a p-value quantifies the strength of the empirical evidence that a decision can be made. But of course a p-value does not indicate the probability of a correct decision.

2. The Proposed Method

Let η_{jqx} denote the qth quantile of Y_j given that $X_j = x$. Here it is assumed that

$$\eta_{jqx} = \beta_{0jq} + \beta_{1jq}x. \tag{8}$$

The unknown slope, β_{1jq} and intercept, β_{0jq}, can be estimated via the well-known Koenker and Bassett (1978) quantile regression estimator yielding say b_{1jq} and b_{0jg}, respectively. Assuming (8) is true provides a straightforward method for estimating $Q(x)$. Let $\hat{\theta}_2 = b_{1,2,0.5}x + b_{0,2,0.5}$ denote the estimate of the conditional median of the experimental group given that $X = x$. As is evident, $\hat{\theta}_2$ corresponds to some quantile of the conditional distribution associated with the control group, given that $X = x$, which is $Q(x)$. An estimate of $Q(x)$, $\hat{Q}(x)$, is the value of q such that

$$b_{1,1,q}x + b_{0,1,q} = \hat{\theta}_2. \tag{9}$$

Here, (9) is solved with the Nelder and Mead (1965) algorithm.

Now consider the goal of testing (7) as well as computing a confidence interval for $Q(x)$. Here, a percentile bootstrap method is used. For theoretical results that motivate the use of this method, see Liu and Singh (1997). Consideration of this approach stems from past studies indicating that it frequently performs well when dealing with robust estimators (Wilcox, 2022b). Briefly, let (X_{ij}, Y_{ij}), $(i = 1, \ldots n_j;\ j = 1,2)$ denote a random sample of size n_j from the jth group. Generate a bootstrap sample from each group by sampling with replacement n_j pairs of values from group j. Based on these bootstrap values, compute the estimate of $Q(x)$ yielding $\hat{Q}^*(x)$. Repeat this process B times and label the results $\hat{Q}^*_b(x)$ $(b = 1, \ldots, B)$.

Let

$$P^* = \sum I(\hat{Q}^*_b(x) < 0.5), \tag{10}$$

where the indicator function $I(\hat{Q}^*_b(x) < 0.5) = 1$ if $\hat{Q}^*_b(x) < 0.5$, otherwise $I(\hat{Q}^*_b(x) < 0.5) = 0$. Then a (generalized) p-value for testing (7) is $2\min(P^*, 1 - P^*)$. To compute a $1 - \alpha$ confidence interval, first put the bootstrap estimates in ascending order and label the results $\hat{Q}^*_{(1)}(x) \leq \cdots \leq \hat{Q}^*_{(B)}(x)$. Let $\ell = \alpha B/2$ and $u = B = \ell$. Then a $1 - \alpha$ confidence interval for $Q(x)$ is

$$(\hat{Q}^*_{(\ell+1)}(x), \hat{Q}^*_{(u)}(x)). \tag{11}$$

This is called method Q henceforth. The choice for B is discussed in the next section of this paper.

3. Simulation Results

Simulations were used to get some sense of how well the percentile bootstrap performs when making inferences about $Q(x)$. First, some comments about choosing B are required. Racine and MacKinnon (2007) discuss this issue at length and proposed a method for choosing the number of bootstrap samples. Davidson and MacKinnon (2000) proposed a pretest procedure for choosing B. Typically $B \geq 500$ is used. However, a practical problem was that execution time using the Nelder-Mead method to solve (9) was much higher than expected. Even with $B = 100$ and $n_1 = n_2 = 20$, execution time was over 18 seconds on a MacBook Pro using a 2.9 GHz processor. The problem is that running a simulation with 1000 replications and $B = 500$ would require over 52 hours. Switching to alternative minimization functions in the R package optim did not improves matters. Here, the execution time was reduced by taking advantage of a quad core processor via the R package parallel. Now with $B = 200$, execution time for a single replication was a little over 26 seconds. That is, for 1000 replications, the execution time is a little over seven hours. Consequently, $B = 200$ was used in the simulations with 1000 replications.

Data were generated from four distributions: normal, symmetric and heavy tailed, skewed and relatively light-tailed, and skewed with heavy tails. Roughly, heavy-tailed distributions are characterized by outliers. More precisely, data were generated from four g-and-h distributions. Let Z denote a random variable having a standard normal distribution. Then

$$V = \begin{cases} \frac{exp(gZ)-1}{g} exp(hZ^2/2), & \text{if } g > 0 \\ Zexp(hZ^2/2), & \text{if } g = 0 \end{cases} \tag{12}$$

has a g-and-h distribution (Hoaglin, 1985), where g and h are parameters that determine the first four moments. The four distributions considered here are the standard normal distribution ($g = h = 0$), a symmetric heavy-tailed distribution ($g = 0$, h= 0.2), an asymmetric distribution with relatively light tails ($g = 1$, $h = 0$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). The g-and-h distribution with $g = 1$ and $h = 0$ corresponds to a lognormal distribution that has been shifted to have a median of zero. Figure 3 shows plots of the four distributions used here. A review of five papers aimed at characterizing the extent distributions are non-normal (Wilcox, 2022b, section 4.2) suggests that the g-and-h distributions used here span what typically encountered in practice.

Inferences about $Q(x)$ were made based on two choices for x. 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.

Estimates of the actual Type I error probability are reported in Table 1. 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, the highest estimate is 0.053. Bradley's criterion is satisfied for the point $(L+U)/2$ with one exception, which occurred when $(n_1, n_2) = (20, 20)$, $g = 1$ and $h = 0.2$. The estimate is 0.020. For U there are situations where the estimate drops below 0.025 when one or both sample sizes are less than or equal to 50. The lowest estimate is 0.017. For $(n_1, n_2) = (100, 100)$, the estimated Type I error probability satisfies Bradley's criterion in all of the situations considered.

There is the issue of how the power of the proposed method compares to situations where the covariate is ignored or not available. Power can be higher or lower depending on the nature of the association. Consider, for example, $n = 50$, $g = h = 0$ and suppose the groups are compared for the covariate value corresponding U. If there is no association, $\beta_{01} = 0.5$ and $\beta_{02} = 0$, the power of the proposed method is 0.31. But if $H_0 : Q = 0.5$ is tested ignoring the covariate, power is 0.66. However, if $\beta_{01} = \beta_{02} = 0$, $\beta_{11} = 1$ and $\beta_{12} = 0$, the proposed method has power 0.568. In contrast, ignoring the covariate, the power is only 0.050 because in effect the hypothesis $H_0 : Q = 0.5$ is true.

4. Illustration

The proposed method is illustrated with data from the Well Elderly 2 study (Clark et al., 2011). Generally, this study dealt with an intervention program aimed at improving the physical and emotional well being of older adults. The focus here is on a measure of meaningful activities (MAPA). For each participant, cortisol was measured upon awakening and again 30-45 minutes later. The change in cortisol, generally known as the cortisol awakening response (CAR) has been found to be associated with measures of stress (e.g., Clow et al., 2004; Chida & Steptoe, 2009). Consequently, the goal is to compare MAPA measures with CAR taken as the covariate. The sample sizes are 232 for the control group and 141 for the intervention group.

Figure 4 shows a plot of the data and the 0.5 quantile regression lines. For the control group, the data points are indicated by a + and the solid line is the regression line. Table 2 summarizes the results for CAR=-0.2, -0.1 and 0.1. As can be seen, the first two p-values are less than or equal to 0.02. At CAR=-0.2, the estimate of Q is 0.711, which is moderately large.

To provide perspective, the groups were compared again based on the conditional median of the MAPA scores given a

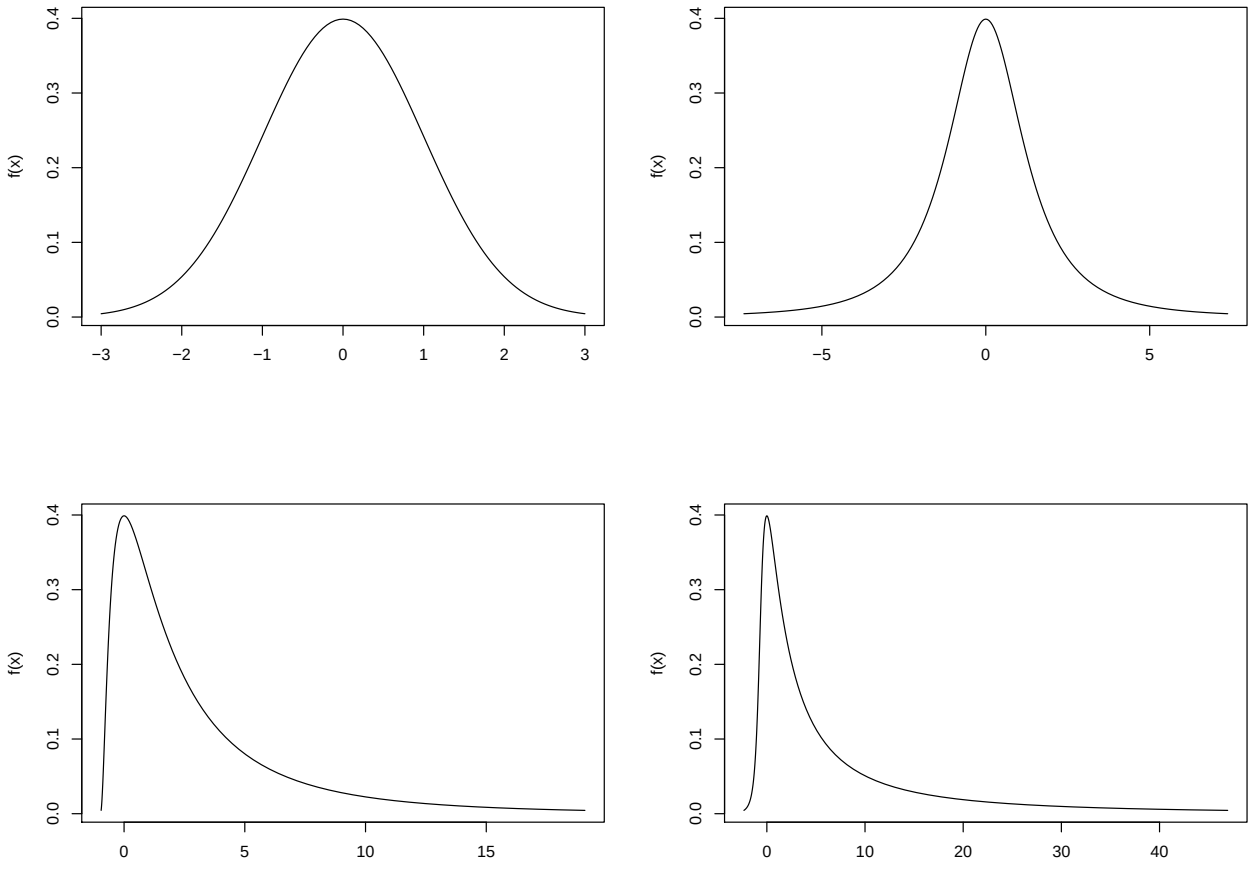

Figure 3. Distributions used in the simulations. Upper left, a standard normal distribution; upper right $g = 0, h = 0.2$; lower left, $g = 1, h = 0$; lower right, $g = 1, h = 0.2$

Table 1. Estimated Type I error probabilities, $\alpha = 0.05$

(n_1, n_2)	g	h	$(L + U)/2$	U
(20, 20)	0.0	0.0	0.030	0.019
	0.0	0.2	0.035	0.026
	1.0	0.0	0.030	0.021
	1.0	0.2	0.020	0.021
(20, 50)	0.0	0.0	0.028	0.031
	0.0	0.2	0.026	0.023
	1.0	0.0	0.029	0.026
	1.0	0.2	0.029	0.021
(50, 50)	0.0	0.0	0.031	0.034
	0.0	0.2.	0.035	0.017
	1.0	0. 0	0.038	0.021
	1.0	0.2	0.040	0.020
(100, 100)	0.0	0.0	0.049	0.034
	0.0	0.2	0.035	0.037
	1.0	0.0	0.039	0.030
	1.0	0.2	0.058	0.036

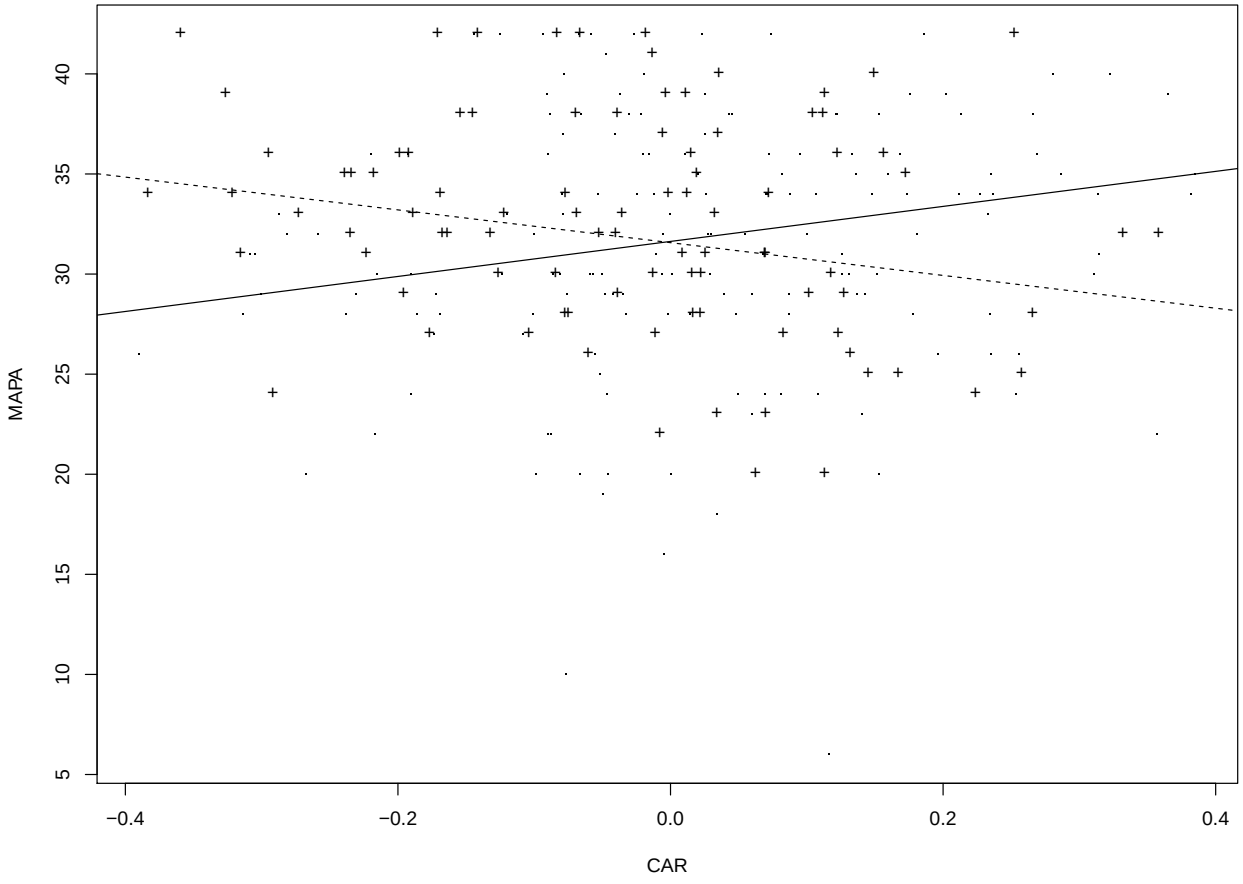

Figure 4. Solid line is the regression line for the control group

Table 2. Results for the Well Elderly data using method Q

CAR	p.value	Conf.Inter	$\hat{Q}$
-0.2	0.00	(0.588, 0.814)	0.711
-0.1	0.02	(0.509, 0.726)	0.586
0.1	0.36	(0.211, 0.632)	0.374

Table 3. Results when comparing the conditional medians

CAR	Est.1	Est.2	DIF	Conf.Inter	p.value
−0.2	30.8	33.5	−2.7	(−5.76, −0.28)	0.030
−0.1	31.3	32.6	−1.3	(−3.90, 1.21)	0.208
0.1	32.3	30.9	1.4	(−1.63, 4.51)	0.261

value for CAR. This was done via the method in Wilcox (2022b, section 12.1, method S1). The R function ancJN in the R package WRS was used. The results are reported in Table 3. As can be seen, the p-values differ substantially from those reported in Table 2, especially for CAR=−0.1 and 0.1, illustrating that the choice of method can make a practical difference. Of course, this is not surprising because the two methods used here are sensitive to different features of the data.

5. Concluding Remarks

An alternative to Q that reflects the approach given by (2) can be outlined as follows. Let $\theta_j(x)$ denote the conditional median of Y_j given that $X = x$. Let $\tau_j(x)$ denote the interquartile range of Y_j given that $X = x$, rescaled to estimate the standard deviation when the conditional distribution of the Y_j is normal. Using $\tau_j(x)$ as a robust measure of scale is convenient because it is readily estimated by the Koenker–Bassett regression estimator. Then an analog of (4) is readily derived, which is labeled ξ. However, when dealing with skewed distributions, this approach might be deemed unsatisfactory for reasons previously described.

A possible appeal of ξ is that it provides a measure of effect size without having to specify one of the groups as a control group. But perhaps this is not a serious concern when using Q. Imagine, for example, males and females are compared. One could use females as the control group, estimate Q, and then use males as the control group, which in general would yield a different estimate of Q.

It is not being suggested that Q should be used to the exclusion of other measures of effect size. The suggestion is that multiple perspective can be useful and that Q supplements other measures that might be deemed reasonable. A possible appeal of Q is that it provides a flexible way of characterizing the extent an experimental group improves upon a control group regardless of the shape of the distribution of the control group.

Finally, the R function anclin.QS.CIpb performs method Q. It is contained in the file Rallfun-v39, which can be downloaded from https://osf.io/dashboard. Simply source the file to gain access to anclin.QS.CIpb. By default, the covariate values are taken to be L, $(L+U)/2$ and U. The covariate values can be specified via the argument pts. Setting the argument MC=TRUE, the function will take advantage of a multicore processor if one is available provided the R package parallel has been installed. It is noted that the Nelder-Mead method was applied via the R function nelderv2, which is in the R package WRS as well as the Rallfun-v39 file. When using the R function optim instead, situations were found where for x sufficiently large, nonsensical estimates of Q were obtained in some instances. The reason for this is unknown.

References

Algina, J., Keselman, H. J. & Penfeld, 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*, 317–328. https://doi.org/10.1037/1082-989X.10.3.317

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

Chida, Y., & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology, 80*, 265–278. https://doi.org/10.1016/j.biopsycho.2008.10.004

Clark, F., Jackson, J., & Carlson, M., et al. (2011). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomise Controlled Trial. *Journal of Epidemiology and Community Health, 66*, 782–790.

Clow, A., Thorn, L., Evans, P., & Hucklebridge, F. (2004). The awakening cortisol response: Methodological issues and significance. *Stress, 7*, 29–37. https://doi.org/10.1080/10253890410001667205

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

Davidson, R., & MacKinnon, J. G. (2000). Bootstrap tests: how many bootstraps? *Econometric Reviews, 19*, 55–68. https://doi.org/10.1080/07474930008800459

Grissom, R. J., & Kim, J. J. (2012). *Effect Sizes for Research: Univariate and Multivariate Applications*. UK: Routledge. https://doi.org/10.4324/9780203803233

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

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.

Huber, P. J., & Ronchetti, E. (2009). *Robust Statistics*, 2nd Ed. New York: Wiley. https://doi.org/10.1002/9780470434697

Huberty, C. (2002). A history of effect size indices. *Educational and Psychological Measurement, 62*, 227–240. https://doi.org/10.1177/0013164402062002002

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

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrika, 46*, 33–50. https://doi.org/10.2307/1913643

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

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *Computer Journal, 7*, 308–313. https://doi.org/10.1093/comjnl/7.4.308

Racine, J., & MacKinnon, J. G. (2007). Simulation-based tests than can use any number of simulations. *Communications in Statistics–Simulation and Computation, 36*, 357–365. https://doi.org/10.1080/03610910601161256

Staudte, R. G., & Sheather, S. J. (1990). *Robust Estimation and Testing*. New York: Wiley. https://doi.org/10.1002/9781118165485

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In I. Olkin, S. Ghurye, W. Hoeffding, W. Madow & H. Mann (Eds.) *Contributions to Probability and Statistics. Stanford*, CA: Stanford University Press (pp. 448–485).

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100–116. https://doi.org/10.1214/ss/1177011945

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

Wilcox, R. R. (2022b). *Introduction to Robust Estimation and Hypothesis Testing*. 5th Edition. San Diego, CA: Academic Press. https://doi.org/10.1016/B978-0-12-820098-8.00007-5