Sankhyā : The Indian Journal of Statistics
2023, Volume 85-B, Part 2, pp. 330–343

Within Groups Designs: Inferences Based on A Robust Nonparametric Measure of Effect Size

Rand R. Wilcox
University of Southern California, Los Angeles, USA

Abstract

The paper deals with a robust, projection-type measure of effect size when comparing $J > 1$ dependent groups. The measure of effect size is scale invariant and does not assume or require that the underlying multivariate distribution is elliptically contoured. By design the measure of effect size is equal to zero when the corresponding measures of location are equal. A simple method is suggested for testing the hypothesis that this effect size is zero. The method is readily extended to comparing K-variate distributions associated with two independent groups. One of the main goal is to report simulation results on how well the method performs in terms of controlling the Type I error probability. The method, when comparing independent groups, is used to reveal new insights into the connection between depression and cortisol levels.

AMS (2000) subject classification. C12; C18; I31.
Keywords and phrases. Robust methods, Effect size, Heteroscedasticity, Projection distance, ANOVA, Cortisol, Depression.

1 Introduction

Consider J dependent groups and let θ_j denote some measure of location associated with the jth group. Of course, a common goal is to test

$$H_0 : \theta_1 = \cdots = \theta_J. \tag{1}$$

The classic approach to testing this hypothesis is based in part on the variation of the J measures of location. That is, it is based on an estimate of a particular measure of effect size:

$$\Lambda = \sum(\theta_j - \bar{\theta})^2, \tag{2}$$

where $\bar{\theta} = \sum \theta_j/J$. The standard approach to testing (1) is to determine whether the estimate of Λ is sufficiently large to justify rejecting the null hypothesis.

Recent years have seen an increased interest in measures of effect size that reflect differences among measures of location in conjunction with some measure of dispersion. When comparing three or more independent groups, there are now a variety of scale invariant measures of effect size that might be used. That is, they are scale free; multiplying the data by some constant c, $c \neq 0$, does not alter their value. Included are robust, heteroscedastic measures of effect size that take into account differences among some measure of location in conjunction with some robust measure of dispersion (e.g., Wilcox, 2022).

A minor goal here is to suggest a robust, scale invariant measure of effect size for characterizing the extent dependent groups differ. The basic idea is to measure the distance of an estimate of the null vector from an estimate of $(\theta_1, \ldots, \theta_J)$, the center of the data cloud. Perhaps the most obvious approach is to use Mahalanobis distance, but there are two concerns with this approach. First, it is not robust (e.g., Rousseeuw and Leroy, 1987). That is, even a small shift in a distribution can alter its value tremendously. In particular, a relatively large distance assuming normality can be rendered arbitrarily small by even a small shift toward a heavy-tailed distribution where outliers are likely to occur. A known strategy for dealing with this issue is to replace the mean and covariance matrix with some robust analog. There are a variety of possibilities (e.g., Wilcox, 2022). However, there is a second concern: robust analogs of Mahalanobis distance are reasonable provided the unknown multivariate distribution is elliptically contoured. A goal here is to avoid this restriction. This is done via the notion of projection distance (e.g., Wilcox, 2022, section 6.2.5), which takes into account the overall structure of the data cloud.

By design, the measure of effect size considered here, Ξ, is equal to zero when (1) is true. There are two main goals. The first is to report simulation results on how well a method for testing

$$H_0 : \Xi = 0 \tag{3}$$

performs in terms of controlling the Type I error probability. Because the method for testing (3) is sensitive to different features of the data compared to a conventional method for testing (1), there are situations where the power of the proposed method can be substantially higher, as will be illustrated. Not surprisingly, there are situations where the conventional method for (1) has higher power than the proposed method.

Now consider two independent groups where for each group there are K measures for each participant. Let Ξ_ℓ ($\ell = 1, 2$) denote Ξ for the ℓth group. The method for testing (3) is readily extended to testing

$$H_0 : \Xi_1 = \Xi_2. \tag{4}$$

The second goal is to report simulations on how well the proposed method performs when testing (4). As will be evident, testing (4) can detect differences between groups that are missed when comparing the marginal measures of location instead.

There are numerous robust location estimators that might be used when dealing with dependent random variables (e.g., Wilcox, 2022). The focus here is on a 20% trimmed mean associated with the marginal distributions with the understanding that arguments can be made for using a variety of alternative estimators. For notational convenience, consider 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 the amount of trimming, $0 \leq \gamma < 0.5$. Let $g = [\gamma n]$, where $[\gamma n]$ is the value of γn rounded down to the nearest integer. The sample 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

$$\bar{X}_t = \frac{X_{(g+1)} + \cdots + X_{(n-g)}}{n - 2g}. \tag{5}$$

Here, $\gamma = 0.2$ is used because it performs nearly as well as the sample mean under normality while guarding against low efficiency when dealing with heavy-tailed distributions where outliers are likely to occur.

The paper is organized as follows. Section 2 describes the details of Ξ. Section 3 describes a simple method for testing (3) as well as (4). Section 4 reports simulation results on how well the method controls the Type I error probability. The power of the proposed method for testing (3) is compared to a standard method for testing (1). The proposed methods are illustrated in Section 5.

2 The Proposed Method

First consider the issue of measuring effect size based in part on the marginal trimmed means but with the additional goal of taking into account the overall dispersion and structure of the data cloud. Let $\hat{\boldsymbol{\mu}}_t = (\bar{X}_{t1}, \ldots, \bar{X}_{tJ})$, where

$\bar{X}_{tj}$ is the sample trimmed mean based on the jth marginal distribution. Let $\tilde{\boldsymbol{\mu}}_t = (\bar{X}_G, \ldots, \bar{X}_G)$ denote the estimate of the measures of location when (1) is true, where $\bar{X}_G = \sum \bar{X}_{tj}/J$. As previously noted, the basic idea is to use an estimate of a projection-type distance of the null vector, $\tilde{\boldsymbol{\mu}}_t$ from an estimate of the center of the data cloud, $\hat{\boldsymbol{\mu}}_t$. Projection distance has a close connection to the Donoho and Gasko (1992) approach to defining the notion of halfspace depth derived by Tukey (1975).

For notational convenience, let $\mathbf{Y}_i = (X_{i1}, \ldots, X_{iJ})$, $i = 1, \ldots, n$ and $\mathbf{Y}_{n+1} = \tilde{\boldsymbol{\mu}}_t$. An outline of the projection distance estimator is as follows. For each i $(i = 1, \ldots, n+1)$, project the data onto the line between $\hat{\boldsymbol{\mu}}_t$ and $\mathbf{Y}_i$. For each projection compute a standardize distance between $\hat{\boldsymbol{\mu}}_t$ and $\mathbf{Y}_{n+1}$. The maximum of these $n+1$ distances is the projection distance between $\hat{\boldsymbol{\mu}}_t$ and $\mathbf{Y}_{n+1}$, which is taken to be $\hat{\Xi}$, the estimate of the effect size, Ξ.

The computational details are as follows. For any i, $i = 1, \ldots, n+1$, let

$$\mathbf{U}_i = \mathbf{Y}_i - \hat{\boldsymbol{\mu}}_t,$$

$$B_i = \mathbf{U}_i \mathbf{U}_i' \tag{6}$$

and for any j $(j = 1, \ldots, n+1)$ let

$$W_{ij} = \sum U_{ik} U_{jk},$$

$$T_{ij} = \frac{W_{ij}}{B_i}(U_{i1}, \ldots, U_{iJ}) \tag{7}$$

and

$$G_{ij} = \|T_{ij}\|,$$

where $\|T_{ij}\|$ is the Euclidean norm. Now let

$$g_{ij} = \frac{G_{ij}}{q_2 - q_1}, \tag{8}$$

where q_1 and q_2 are estimates of the lower and upper quartiles, respectively, based on the values $G_{i1}, \ldots, G_{i,n+1}$. Here, the quartiles are estimated with the ideal fourths estimator (Frigge et al., 1989). The projection distance of Y_{n+1} is $\max g_{i,n+1}$, the maximum being taken over $i = 1, \ldots, n+1$ and is taken to be $\hat{\Xi}$.

3 Methods M and C

This section describes a method for testing (3) and (4). It is noted that when working with a robust estimator, a percentile bootstrap method often performs relatively well when testing hypotheses and computing confidence intervals (e.g., Wilcox, 2022). But this approach was rather unsatisfactory. When testing (4), even when both sample sizes are equal to 200, the actual level was well below the nominal 0.05. An alternative approach was found to be more satisfactory in simulations.

3.1 Method M Method M, aimed at testing (3), is similar in spirit to Student's t-test and the ANOVA F test: determine the null distribution under normality and investigate the impact of non-normality on the actual level of the test statistic. More precisely, momentarily assume multivariate normality with a mean $\mathbf{0}$ and covariance matrix equal to the identity matrix. Next, estimate the null distribution via a simulation. That is, given n and J, generate n vectors of observations from a J-variate normal distribution and compute an estimate of Ξ, $\hat{\Xi}^*$. Repeat this B times yielding $\hat{\Xi}^*_1, \ldots, \hat{\Xi}^*_B$. Here, $B = 2000$ is used. Let $\hat{\Xi}^*_{(1)} \leq \cdots \leq \hat{\Xi}^*_{(B)}$ denote $\hat{\Xi}^*_1, \ldots, \hat{\Xi}^*_B$ written in ascending order. The null hypothesis is rejected at the α level if $\hat{\Xi}$ is greater than or equal to the $1 - \alpha$ quantile of the estimated null distribution. Here, the $1 - \alpha$ quantile of the null distribution is simply taken to be $\hat{\Xi}^*_{(c)}$, where $c = (1 - \alpha)B$ rounded to the nearest integer. A p-value for testing (3) is

$$\sum \frac{1}{B} I(\hat{\Xi} \geq \hat{\Xi}^*_b), \tag{9}$$

where the indicator function $I(\hat{\Xi} \geq \hat{\Xi}^*_b) = 1$ if $\hat{\Xi} \geq \hat{\Xi}^*_b$, otherwise $I(\hat{\Xi} \geq \hat{\Xi}^*_b) = 0$. This will be called method M henceforth.

3.2 Method C As for testing (4), method C mimics method M. That is, for each group, momentarily assume multivariate normality with a mean $\mathbf{0}$ and covariance matrix equal to the identity matrix and use a simulation to estimate the null distribution of $D = \hat{\Xi}_1 - \hat{\Xi}_2$. That is, generate n_1 vectors of values for the first group, n_2 vectors for the second group, compute the estimated difference yielding D^*, and repeat this process B times yielding $D^*_1, \ldots, D^*_B$. Let $D^*_{(1)} \leq \cdots \leq D^*_{(B)}$ denote the $D^*_1, \ldots, D^*_B$ values written in ascending order. Let $\ell = \alpha B/2$, rounded to the nearest integer and let $u = (1 - \alpha/2)B$ rounded to the nearest integer. Now reject at the α level if

$D \leq D_{(\ell)}$ or if $D \geq D_{(u)}$. Let

$$P = \frac{1}{B}\sum I(D < D_b^*) \tag{10}$$

A p-value is

$$2\min\{P, 1-P\}. \tag{11}$$

4 Simulation Results

Simulations were used to assess the ability of methods M and C to control the probability of a Type I error when dealing with non-normal distributions as well as situations where the the correlations among the random variables differ from zero. All of the simulation results are based on 2000 replications. The marginal distributions were taken to have one of four g-and-h distributions (Hoaglin, 1985) that contains the standard normal distribution as a special case. If Z has a standard normal distribution, then by definition

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

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 1 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). Data were generated from a multivariate normal distribution with a common Pearson's correlation, ρ, equal to zero or 0.5. Then the marginal distributions were transformed to a g-and-h distribution. This transformation alters slightly the value of Pearsons's correlation, but the R function rngh in the R package WRS adjusts for this. The sample sizes were taken to be $n = 25$, 50, 100 and 200

Table 2 shows the estimated Type I error probability when $J = 4$ and when testing at the 0.05 level. To add perspective, results are reported on a method for testing (1), based on a 20% trimmed mean, which is labeled method RMT. The computational details are summarized in Wilcox (2022, section 8.1.1), but for brevity they are not described. The results for $n = 100$ provided no new insights so they are not reported.

Although the seriousness of a Type I error depends on the situation, 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. RMT satisfies Bradley's criterion. The largest estimates for methods M and RMT are 0.058 and 0.053, respectively. As for method M, when dealing with a relatively heavy-tailed distribution, the estimates drop below 0.025 for $n = 25$ and 50. The lowest estimate was 0.016. For $n = 200$, the estimates indicate that M satisfies Bradley's criterion except for $\rho = 0.5$, $g = 0$ and $h = 0.2$, the estimate being 0.022.

Table 3 shows the results when $J = 6$. Again, both methods perform well in terms of avoiding Type I errors well above the nominal level. The main difficulty is that for method M, estimates drop below 0.025, the lowest being 0.014. As in Table 3, it is heavy-tailed distributions that cause problems, especially when the variables have a relatively high correlation.

Table 4 compares the power of method M and RMT when $n = 50$. This was done by generating the data as done in Table 2, and then transforming the first marginal distribution to $\sigma X_{i1} + \delta_1$ $(i = 1, \ldots, n)$. The second marginal distribution was transformed to $X_{i2}+\delta_2$. As might be expected, the difference in power can be quite large because the two methods are sensitive to different features of the data. The main point is that there are situations where each method offers a distinct advantage over the other. That is, in practical terms, given the goal maximizing power, the choice can make a substantial difference with the optimal choice depending on the unknown nature of the distributions.

Finally, Tables 5 and 6 report results on the ability of method C to control the Type I error probability. Now there are situations where the estimates exceed 0.06, the largest estimate being 0.069. For equal sample sizes, the lowest estimate is 0.015, which occurred when $n = 25$, $\rho = 0.5$ and $g = h = 0.2$. For unequal sample sizes, there is only one estimate less than 0.025, namely 0.022, which occurred for $n_1 = 25$, $n_2 = 50$, $\rho = 0.5$ and $g = h = 0.2$.

Table 1: Some properties of the g-and-h distribution

g	h	κ_1	κ_2
0.0	0.0	0.5	3.0
0.0	0.2	0.5	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

Table 2: Estimated Type I errors using methods M and RMT, $J = 4$

n	ρ	g	h	M	RMT
25	0.0	0.0	0.0	0.050	0.037
25	0.0	0.0	0.2	0.032	0.047
25	0.0	0.2	0.0	0.055	0.042
25	0.0	0.2	0.2	0.039	0.045
25	0.5	0.0	0.0	0.033	0.046
25	0.5	0.0	0.2	0.019	0.036
25	0.5	0.2	0.0	0.031	0.038
25	0.5	0.2	0.2	0.016	0.039
50	0.0	0.0	0.0	0.050	0.053
50	0.0	0.0	0.2	0.034	0.048
50	0.0	0.2	0.0	0.058	0.046
50	0.0	0.2	0.2	0.041	0.040
50	0.5	0.0	0.0	0.036	0.044
50	0.5	0.0	0.2	0.023	0.041
50	0.5	0.2	0.0	0.041	0.049
50	0.5	0.2	0.2	0.019	0.037
200	0.0	0.0	0.0	0.050	0.051
200	0.0	0.0	0.2	0.032	0.052
200	0.0	0.2	0.0	0.055	0.044
200	0.0	0.2	0.2	0.031	0.044
200	0.5	0.0	0.0	0.046	0.051
200	0.5	0.0	0.2	0.022	0.045
200	0.5	0.2	0.0	0.052	0.046
200	0.5	0.2	0.2	0.025	0.043

5 An Illustration

The methods are illustrated with data from the Well Elderly 2 study (Clark et al., 2011) where the general goal was to assess the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. The focus here is on cortisol levels measured at four different times: upon awakening, 30-45 minutes after awakening, just before lunch and just before dinner. Past studies indicate that cortisol has a connection with psychosocial factors (Kirschbaum et al. 1995; Chida and Steptoe, 2009) including depression and anxiety disorders (e.g., Stetler and Miller, 2005; Bhattacharyya et al., 2008). These studies have focused on the cortisol awak-

ening response (CAR), which is just the difference between a participant's cortisol level upon awakening and measured again 30-45 minutes later.

It is briefly noted that both methods M and RMT reject at the 0.05 level indicating that the cortisol levels differ over time. What is more interesting are results based on method C when comparing two groups based on a measure of depressive symptoms. The first group consisted of participants with a score greater than 15, which generally is taken to indicate someone with minor depressive symptoms or worse. The sample size is 61. The second group consisted of the participants with a score less than or equal to 15 and the sample size is 114.

First focus on the first three cortisol measures. Viewing the data in the context of a between-by-within ANOVA design, the between group main effect yielded a p-value equal to 0.214 based on a Welch-type analog for trimmed means, which is described in Wilcox (2022, section 8.6). For each of the three measures taken, the two groups did not differ significantly at the 0.05 level based on a method for comparing 20% trimmed means derived

Table 3: Estimated Type I errors using methods M and RMT, $J = 6$

n	ρ	g	h	M	RMT
20	0.0	0.0	0.0	0.050	0.040
20	0.0	0.0	0.2	0.036	0.045
20	0.0	0.2	0.0	0.059	0.046
20	0.0	0.2	0.2	0.043	0.046
20	0.5	0.0	0.0	0.027	0.037
20	0.5	0.0	0.2	0.015	0.030
20	0.5	0.2	0.0	0.039	0.037
20	0.5	0.2	0.2	0.017	0.031
200	0.0	0.0	0.0	0.050	0.057
200	0.0	0.0	0.2	0.017	0.050
200	0.0	0.2	0.0	0.054	0.046
200	0.0	0.2	0.2	0.029	0.045
200	0.5	0.0	0.0	0.051	0.051
200	0.5	0.0	0.2	0.014	0.049
200	0.5	0.2	0.0	0.049	0.048
200	0.5	0.2	0.2	0.021	0.054

Table 4: Estimated power using methods M and RMT, $J = 4$, $n = 50$

ρ	g	h	M	RMT
$\sigma = 1$, $\delta_1 = 0.5$, $\delta_2 = 0$				
0.0	0.0	0.0	0.519	0.625
0.0	0.0	0.2	0.382	0.564
0.0	0.20	0.0	0.526	0.619
0.0	0.20	0.2	0.386	0.559
0.5	0.0	0.0	0.774	0.894
0.5	0.0	0.2	0.614	0.843
0.5	0.2	0.0	0.777	0.890
0.5	0.2	0.2	0.620	0.828
$\sigma = 3$, $\delta_1 = 1$, $\delta_2 = 0.2$				
0.0	0.0	0.0	0.931	0.656
0.0	0.0	0.2	0.810	0.564
0.0	0.2	0.0	0.929	0.643
0.0	0.2	0.2	0.820	0.538
0.5	0.0	0.0	0.989	0.864
0.5	0.0	0.2	0.950	0.767
0.5	0.2	0.0	0.986	0.849
0.5	0.2	0.2	0.946	0.752

by Yuen (1974). The corresponding p-values are 0.124, 0.299 and 0.740. Comparing the groups based on CAR, again the groups did not differ significantly; the p-value is 0.464. That is, past studies suggest that these two groups would differ based on CAR, but this was not verified here. However, based on method C, again using the first three cortisol measures, the p-value is less than 0.001. The estimated effect sizes are 2.01 for the depressive group and 1.23 for the non-depressive group. The ratio of these two estimates is 1.6. Repeating the analysis using all four cortisol measures, now the p-value based on method C is 0.022.

Method C was applied again for a separate group where measures were taken after intervention. Now the sample sizes for the depressed and not depressed groups are 65 and 169, respectively. Comparing the depressed group to the not depressed group based on the first three cortisol measures, the p-value is 0.048. But using all four measures the p-value is 0.574.

Table 5: Estimated Type I errors using method C, equal sample sizes, $K = 4$

n	ρ	g	h	C
25	0.0	0.0	0.0	0.051
25	0.5	0.0	0.0	0.052
25	0.0	0.0	0.2	0.021
25	0.5	0.0	0.2	0.025
25	0.0	0.2	0.0	0.047
25	0.5	0.2	0.0	0.047
25	0.0	0.2	0.2	0.031
25	0.5	0.2	0.2	0.015
50	0.0	0.0	0.0	0.050
50.	0.5	0.0	0.0	0.062
50	0.0	0.0	0.2	0.021
50	0.5	0.0	0.2	0.028
50	0.0	0.2	0.0	0.039
50	0.5	0.2	0.0	0.028
50	0.0	0.2	0.2	0.021
50	0.5	0.2	0.2	0.027
200	0.0	0.0	0.0	0.055
200	0.5	0.0	0.0	0.051
200	0.0	0.0	0.2	0.031
200	0.5	0.0	0.2	0.022
200	0.0	0.2	0.0	0.049
200	0.5	0.2	0.0	0.053
200	0.0	0.2	0.2	0.020
200	0.5	0.2	0.2	0.029

6 Concluding Remarks

All indications are that methods M and C avoid Type I errors well above the nominal level. The main difficulty occurs when dealing with distributions that are skewed and relatively heavy tailed: there are situations where the actual level was estimated to be below 0.025 when testing at the 0.05 level.

There are many possible variations of methods M and C. For example, replace the 20% trimmed mean with the median or perhaps a robust M-estimator. There are also numerous affine equivariant estimators that effectively deal with outliers in a manner that takes into account the

Table 6: Estimated Type I errors using method C, unequal sample sizes, $K = 4$

ρ	g	h	C
$n_1 = 25$, $n_2 = 50$			
0.0	0.0	0.0	0.064
0.5	0.0	0.0	0.050
0.0	0.0	0.2	0.030
0.5	0.0	0.2	0.028
0.0	0.2	0.0	0.061
0.5	0.2	0.0	0.064
0.0	0.2	0.2	0.026
0.5	0.2	0.2	0.022
$n_1 = 25$, $n_2 = 100$			
0.0	0.0	0.0	0.060
0.5	0.0	0.0	0.056
0.0	0.0	0.2	0.025
0.5	0.0	0.2	0.038
0.0	0.2	0.0	0.051
0.5	0.2	0.0	0.069
0.0	0.2	0.2	0.029
0.5	0.2	0.2	0.031

overall structure of the data cloud (e.g., Wilcox, 2022, section 6.3.13). And there are alternative ways of computing the depth of a point in a data cloud. The practical advantages of these variations remain to be determined.

It is not being suggested that inferential methods based on measures of location should be replaced by methods that compare scale invariant measures of effect size. The idea is that different methods provide different perspectives and that multiple perspectives can provide a more nuanced understanding of data. The illustration based on method C demonstrates this point.

Finally, R functions for applying methods M and C are stored in the file Rallfun-v40, which can be downloaded from https://osf.io/nvd59/quickfiles.

The R function rmES.pro applies method M and bwESP.GLOB.B applies method C.

Funding Open access funding provided by SCELC, Statewide California Electronic Library Consortium.

Availability of data and material Publicly available at https://dornsife.usc.edu/cf/labs/wilcox/wilcox-faculty-display.cfm.

Code Availability Stored at https://osf.io/nvd59/quickfiles.

Declarations

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

Bhattacharyya, M. R. & Molloy, G. J., Steptoe, A., (2008). Depression is associated with flatter cortisol rhythms in patients with coronary artery disease. *Journal of Psychosomatic Research, 65*, 107–113. https://doi.org/10.1016/j.jpsychores.2008.03.012

Bradley, J. V. (1978) Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144–152.

Chida, Y. & Steptoe, A., (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology, 80*, 265–278

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. (2011). 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*, 782–790. https://doi.org/10.1136/jech.2009.099754.

Donoho, D. L. & Gasko, M. (1992). Breakdown properties of the location estimates based on halfspace depth and projected outlyingness. *Annals of Statistics, 20*, 1803–1827.

Frigge, M., Hoaglin, D. C. & Iglewicz, B. (1989). Some implementations of the Boxplot. *American Statistician, 43*, 50–54.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) Exploring data tables, trends, and shapes. (pp. 461–515). New York: Wiley.

Kirschbaum, C., Prussner, J. C., Stone, A. A., Federenko, I., Gabb, J., Lintz, D., Schommer, N., Hellhammer, D.H. (1995). Persistent high cortisol responses to repeated psychological stress in a subpopulation of healthy men. *Psychosomatic Medicine. 57*, 468–474. https://doi.org/10.1097/00006842-199509000-00009

Pruessner, J. C., Hellhammer, D. H., & Kirschbaum, C. (1999). Burnout, perceived stress, and cortisol responses to awakening. *Psychosomatic Medicine, 61*, 197–204. https://doi.org/10.1097/00006842-199903000-00012

Rousseeuw, P. J. & Leroy, A. M. (1987). *Robust Regression & Outlier Detection*. New York: Wiley.

Stetler, C. & Miller, G. (2005). Blunted cortisol response to awakening in mild to moderate depression: Regulatory influences of sleep patterns and social contacts. *Journal of Abnormal Psychology, 114*, 697–705.

Tukey, J. W. (1975). Mathematics and the picturing of data. *Proceedings of the International Congress of Mathematicians, 2*, 523–531.

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

Yuen, K. K. (1974). The two sample trimmed t for unequal population variances. *Biometrika, 61*, 165–170. https://doi.org/10.2307/2334299

Rand R. Wilcox
Department of Psychology,
University of Southern California,
Los Angeles CA 90089, USA
E-mail: rwilcox@usc.edu

Paper received: 3 June 2021; accepted 1 August 2023.