International Journal of Statistics and Probability; Vol. 14, No. 1; March 2025
ISSN 1927-7032 E-ISSN 1927-7040
Published by Canadian Center of Science and Education

Comparing Two Independent Groups: Inferences About the Lower and Upper Tails of the Distribution of $D = X - Y$

Rand R. Wilcox[1] & Carla Sanchis-Segura[2]

[1] Dept. of Psychology, University of Southern California, USA

[2] Departament de Psicologia bàsica, clinica i psicobiolgia, Universitat Jaume I., Avda. Sos Baynat, SN12071, Castelló de la Plana, Spain

Correspondence: Rand R. Wilcox, Dept. of Psychology, University of Southern California, USA

Received: December 12, 2024 Accepted: February 20, 2025 Online Published: March 29, 2025

doi:10.5539/ijsp.v14n1p29 URL: https://doi.org/10.5539/ijsp.v14n1p29

Abstract

When comparing two independent groups, a way of getting a more detailed understanding of how the groups compare is to focus on multiple quantiles rather than a single measure of location. There are two distinct approaches regarding how this might be done. The first is to estimate a collection of quantiles for each group and choose an appropriate inferential method for comparing them. This approach has been studied extensively. For example, Doksum and Sievers (1976) derived a nonparametric method for computing confidence intervals for the difference between all quantiles for which the simultaneous probability coverage can be determined exactly assuming random sampling only. An analog of the Doksum–Sievers method was derive by Lombard (2005), which is based on the marginal distributions of two dependent groups. Other methods and applications are described in Wilcox (2023).

Let X and Y denote two independent random variables and let $D = X - Y$. Note that under general conditions, the q quantile of D is not equal to the q quantile of X minus the q quantile of Y. Methods for making inferences about the median of D have been studied that have a close connection to the Wilcoxon–Mann–Whitney test. Here, four methods are suggested and studied via simulations that are aimed at making inferences about the tails of the distribution D with the goal of providing a deeper and more nuanced understanding of how groups compare.

Keywords: typical difference, quantiles, Wilcoxon–Mann–Whitney, Harrell-Davis estimator

1. Introduction

As is evident, there are many methods aimed at comparing two independent groups. Certainly the best-known and most commonly used approach is to use θ_1 and θ_2, where θ_j is some measure of location associated with the jth group ($j = 1, 2$). This approach is informative, but situations are encountered where it is beneficial to get a more detailed understanding of how distributions differ and by how much. Let X and Y denote two independent random variables. One way of accomplishing this goal is to compare multiple quantiles rather than a single measure of location. That is, estimate quantiles associated with the distribution of X, do the same for Y and apply some inferential method that has been found to be reasonably effective. There is a substantial literature on how this might be done (e.g., Wilcox, 2022). Typically, a percentile bootstrap method coupled with a quantile estimator that gives a positive weight to all of the order statistics performs reasonably well. There are known concerns about using estimators summarized by Hyndman and Fan (1996) that are based on only one or two order statistics, particularly when there are tied values. This helps explain why an approach based on the quantile regression estimator derived by Koenker and Bassett (1978) is not recommended. When comparing groups, this corresponds to using only one or two order statistics. Although a percentile bootstrap often performs well in conjunction with an estimator that gives a positive weight to all of the order statistics, there are situations where some alternative approach is preferable. This occurs, for example, when comparing the interquartile range (Greco et al., 2023). For recent advances when dealing with more than two groups, see Wilcox and Rousselet (2024) as well as Özdemir et al. (2024).

Consider random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ and let

$$D_{ij} = X_i - Y_j, \tag{1}$$

($i = 1, \ldots n$; $j = 1, \ldots, m$). Let M_d denote the median of the $N = nm$ D_{ij} values and let M_x and M_y denote the medians based on $X_1, \ldots, X_n$ and $Y_1, \ldots Y_m$, respectively. As is well known, in general, $M_d \neq M_x - M_y$. More broadly, there is no simple connection between the quantiles associated with D and the quantiles associated with X and Y. Moreover, focusing

on M_d provides an informative perspective beyond any method based on M_x and M_y.

For example, imagine that the goal is to compare males and females based on some measure. Focusing on M_x and M_y, the goal is to compare how a typical male compares to a typical female. Focusing on M_d, the goal is to assess the typical difference between a randomly sampled male and a randomly sampled female. Methods designed to make inferences about the median of D are already available (e.g., Wilcox, 2022). The goal here is to expand on these methods using techniques that deal with the tails of the distribution of D. Two basic approaches are considered coupled with some variations to be described. The basic idea is that multiple perspectives can be required to get a deep and nuanced understanding of data as argued by Steegan et al. (2016).

To elaborate, let ξ_q denote the q quantile of D, $0 < q < .5$. The first goal is to make inferences about

$$\Delta = \xi_q + \xi_{1-q}. \tag{2}$$

If the distributions are identical, $\Delta = 0$. The magnitude of Δ is one way of characterizing the extent the groups differ, and of course another way is to simply focus on ξ_q and ξ_{1-q}. Two related goals are testing

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

and computing a $1 - \alpha$ confidence interval for Δ.

Another point of view is comparing $\omega_1 = P(X - Y <= c_1)$ to $\omega_2 = P(X - Y >= c_2)$, where $c_1 < 0$ and $c_2 > 0$ are constants chosen by the investigator. For example, if $c_2 = 10$, how does the likelihood of a difference of 10 units or more compare to negative outcome where the difference is less than or equal to $c_1 = -10$. Now the goal is to test

$$H_0 : \omega_1 = \omega_2. \tag{4}$$

When D has a symmetric distribution and $|c_1| = c_2$, in essence quantiles are being compared as reflected by (3), but q is not specified, and for a skewed distribution, this approach differs from testing (3). A slight variation is to take c_1 equal to an estimate of some quantile $q < .5$ and set $c_2 = |c_1|$. Again, this differs from method Δ when D has a skewed distribution. Yet another goal is to compute a confidence interval for ω_1 or ω_2. Comments on how this compares to making inferences about ξ_q and ξ_{1-q} are relegated to section 3.

Before proceeding, it is noted that the situation considered here has a connection to the Wilcoxon–Mann–Whitney test, which is based on an estimate of $P(X < Y)$, which is simply the proportion of D_{ij} values less than zero. Note that $H_0 : P(X < Y) = .5$ is the same as $H_0 : \xi_{.5} = 0$.

The remainder of the paper is organized as follows. Section 2 suggests a method for testing (3) and (4) followed by a method for computing a confidence interval for ω_1 or ω_2 as well as ξ_q. Section 3 reports simulation results and section 4 illustrates the methods.

2. The Proposed Methods

There is the issue of choosing a quantile estimator comparisons of which have been made by Parrish (1990), Dielman et al. (1994), as well as Sfakianakis and Verginis (2008). No single estimator dominates in terms of efficiency, but the estimator derived by Harrell and Davis (1982) performed reasonably well, it is based on a weighted average of all the order statistics, so it is used here. However, two recent advances should be noted. First, the estimator derived by Navruz and Özdemir (2020) has better efficiency when dealing the quantiles less than .1 and greater than .9. A possible concern with both the Harrell–Davis and Navruz– Özdemir estimators is that they have a breakdown point of only $1/n$. That is, because all order statistics get a positive weight, it is possible to change a single value and make the estimate arbitrarily large or small. The trimmed Harrell–Davis estimator derived by Akinshin (2024) might deal with this.

For notational convenience, denote the $N = nm$ D_{ij} values as $\mathcal{D}_1, \ldots, \mathcal{D}_N$. Given the goal of estimating the q quantile of D, let U be a random variable having a beta distribution with parameters $a = (N + 1)q$ and $b = (N + 1)(1 - q)$. Let

$$W_i = P\left(\frac{i-1}{N} \leq U \leq \frac{i}{N}\right).$$

The Harrell–Davis estimate of the q quantile of D is

$$\hat{\xi}_q = \sum_{i=1}^{N} W_i \mathcal{D}_{(i)}, \tag{5}$$

where $\mathcal{D}_{(1)} \leq \cdots \leq \mathcal{D}_{(N)}$ are the values written in ascending order.

As previously noted, a speculation is that an effective approach to testing (3) and computing a confidence interval for Δ is to use a percentile bootstrap. Briefly, for each group generate a bootstrap sample by sampling with replacement n values from the first group and m values from the second group. Let $\hat{\xi}_q^*$ and $\hat{\xi}_{1-q}^*$ denote the bootstrap estimates of the q and $(1-q)$ quantiles of D, and let $\Delta^* = \hat{\xi}_q^* + \hat{\xi}_{1-q}^*$. Repeat this process B times yielding $\Delta_1^*, \ldots, \Delta_B^*$. Here, $B = 1000$ is used. A $1-\alpha$ confidence interval for Δ is

$$(\hat{\Delta}_{(\ell+1)}^*, \hat{\Delta}_{(u)}^*), \tag{6}$$

where $\ell = \alpha B/2$ rounded to the nearest integer and $u = B - \ell$. Let $p^* = A/B$, where A is the number of bootstrap estimates that are less than zero. A (generalized) p-value is

$$2\min(p^*, 1-p^*) \tag{7}$$

(Liu & Singh, 1993). This approach is called method Q henceforth.

As for testing (4), a simple modification of the bootstrap method just described is used. This approach is called method R henceforth. A simple modification of the bootstrap method provides a confidence interval for ω_1 and ω_2, which is called method T. It is evident that method T is readily modified to make inferences about ξ_q and ξ_{1-q} based on the Harrell–Davis estimator, which is called method SQ.

3. Simulation Results

Several sets of simulations were used to get some sense of how well methods Q, R, T and SQ perform. The first set of simulations focused on three types of continuous distributions: normal, symmetric and heavy-tailed, and a lognormal distribution that is a skewed distribution with relatively heavy tails. The lognormal distribution was chosen because its skewness and kurtosis appear to be an extreme but realistic departure from a normal distribution based on results in Cain et al. (2017), Micceri (1989) and Wu (2002).

More precisely, data were generated from a g-and-h distribution. If Z has a standard normal distribution,

$$\begin{cases} \frac{\exp(gZ)-1}{g}\exp(hZ^2/2), & \text{if } g > 0 \\ Z\exp(hZ^2/2), & \text{if } g = 0 \end{cases} \tag{8}$$

has a g-and-h distribution (Hoaglin, 1985), where g and h are parameters that determine the first four moments. The choice $(g,h) = (1,0)$ corresponds to a lognormal distribution shifted to have a median of zero, while $(g,h) = (0,.2)$ is a symmetric distribution about zero having kurtosis approximately equal to 24.8.

Simulations based on 3000 replications were used to assess the extent the Type I error probability is controlled when testing at the $\alpha = .05$ level. The sample sizes were taken to be $(n_1, n_2) = (20, 20)$, (50, 50) and (20, 50). For method R, $|c_1| = c_2 = c = 1$ was used. For method Q, $q = .75$ and .9 were used. As for method T, $1-\alpha = .95$ confidence intervals for ω_1 were computed when c_1 is taken to be an estimate of the $q = .1$ quantile of D. Within each of the 3000 replications, the true value of ω_1 was determined by sampling 100,000 values from each group and determining the proportion of times $X < Y$ is less than c_1. The results are reported in Table 2.

The estimates satisfy Bradley's (1978) suggestion that when testing at the .05 level, the actual level should be between .025 and .075. The largest estimate is .067 and the lowest is .051. For method T, even when $n_1 = n_2 = 20$, the estimate of α, when computing a $1-\alpha = .95$ confidence interval, is very close to the nominal .05 level. Some additional simulations were run using method T when $q = .2$ and .4 yielding results similar to those in Table 2. No results are reported for method SQ (inferences about ξ_q), which was found to be rather unsatisfactory. For $q = .1$, estimates of the actual level were greater than .075 in all situations, the highest estimate being .124. For $n_1 = n_2 = 50$ and $q = .2$, SQ performed well. When making inferences about tail probabilities, method T is clearly preferable to method SQ in terms of getting a reasonably accurate confidence interval.

The next set of simulations is aimed at getting a more detailed understanding of method R. In contrast to the results in Table 2, now four c_2 values are used that are taken to be the estimates of the .6(.1).9 quantiles of the distribution of D. Again, $c_1 = -c_2$ is used. Two methods for controlling the familywise error (FWE) rate are considered. The first is the method derived by Hochberg (1988). Let $p_1, \ldots, p_C$ be the p-values associated with C tests. Put these p-values in descending yielding $p_{[1]} \geq p_{[2]} \geq \cdots \geq p_{[C]}$. Set $k = 0$ and 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 to $p_{[k]}$.

Table 1. Estimated Type I error probabilities for methods R and Q and estimates of α when computing a $1 - \alpha = .95$ confidence using method T

				R	Q		T
n_1	n_2	g	h	$c = 1$	$q = .75$	$q = .9$	$q = .1$
20	20	0	0	.064	.067	.067	.053
20	20	0	.2	.064	.063	.067	.052
20	20	1	0	.064	.056	.062	.057
50	50	0	0	.051	.052	.053	.054
50	50	0	.2	.053	.054	.053	.049
50	50	1	0	.058	.055	.052	.060
20	50	0	0	.064	.060	.067	.053
20	50	0	.2	.063	.060	.067	.057
20	50	1	0	.064	.058	.062	.045

Table 2. FWE rate using method R based on the .6(.1).9 quantiles

n_1	n_2	g	h	BH	HOCH
20	20	0	0	.057	.052
20	20	0	.2	.063	.057
20	20	1	0	.068	.060
50	50	0	0	.046	.042
50	50	0	.2	.048	.043
50	50	1	0	.057	.051
20	50	0	0	.055	.050
20	50	0	.2	.062	.056
20	50	1	0	.051	.047

2. If $p_{[k]} > \alpha/k$, repeat step 1.

3. Repeat steps 1 and 2 until a significant result is obtained or all C hypotheses have been tested.

The second method is Benjamini and Hochberg (1995), which is aimed at controlling the false discovery rate (FDR), the expected proportion of Type I errors among the null hypotheses that are correct. The Benjamini–Hochberg method simply replaces $p_{[k]} \leq \alpha/k$ in step 1 of Hochberg's method with

$$p_{[k]} \leq \frac{(C - k + 1)\alpha}{C}. \tag{9}$$

A possible appeal of BH is that it always has as much or more power than Hochberg. However, BH does not necessarily control the FWE rate (Hommel, 1988). A minor goal is to get some sense of how BH compares to Hochberg's method despite this result.

Table 3 shows the results. The estimated FWE rate satisfies Bradley's criterion in all of the situations considered. For $(n_1, n_2) = (20, 20)$, Hochberg always has an FWE rate closer to the nominal level. For $(n_1, n_2) = (50, 50)$, there is very little separating the two methods. Now there are situations where BH is closer to the nominal level than Hochberg's method. Moreover, the largest estimate based on BH is .057 for this special case. As for $(n_1, n_2) = (20, 50)$, again the FWE rate using Hochberg is closer to the nominal level.

To get some sense of how tied values impact methods Q, R, T and SQ, data were generated from beta-binomial distributions where the number of bins, M, is 10 or 15. That is, the possible values are the integers that range from 0 to 10 or from 0 to 15. The two parameters for the beta-binomial distribution were taken to be $(r, s) = (3, 3)$, a symmetric distribution, and (4, 2), a distribution skewed to the left. Again, when using method R, c_2 was taken to be the estimate of the q quantile. The actual value of the quantile when using SQ was based on the mean of 100,000 estimates of ξ_q.

Table 4 shows the results. Both methods Q and R performed reasonably well; Bradley's criterion was met in all situations. In terms of controlling the Type I error probability, there is little separating these two techniques. Once again method T performed very well. When the smallest sample size was equal to 20, the actual level of T was estimated to be closer to nominal level than methods Q and R. The estimates for T ranged between .041 and .063 with the bulk of the estimates

Table 3. Estimated Type I error probabilities and estimates of α when computing a $1 - \alpha = .95$ confidence using method T and SQ and there are tied values

					R		Q		T	SQ
n_1	n_2	M	r	s	$q = .75$	$q = .9$	$q = .75$	$q = .9$	$q = .1$	$q = .1$
20	20	10	3	3	.057	0.60	.057	.063	.057	.075
20	20	15	3	3	.058	.062	.057	.062	.054	.078
20	20	10	4	2	.066	.071	.063	.068	.041	.085
20	20	15	4	2	.062	.067	.063	.062	.043	.068
50	50	10	3	3	.052	.054	.055	.056	.052	.067
50	50	15	3	3	.052	.054	.052	.056	.055	.072
50	50	10	4	2	.051	.049	.051	.051	.050	.099
50	50	15	4	2	.052	.052	.051	.051	.051	.059
20	50	10	3	3	.062	.063	.062	.063	.061	.073
20	50	15	3	3	.061	.067	.065	.073	.057	.073
20	50	10	4	2	.069	.075	.065	.067	.051	.087
20	50	15	4	2	.070	.078	.068	.066	.052	.083

Table 4. Results dealing with depressive symptoms using method Q

q	q.est	(1-q).est	Sum	ci.low	ci.up	p.value	p.adj
.1	-16.03	22.38	6.35	-.47	12.88	.066	.066
.2	-8.55	15.60	7.05	.97	12.89	.030	.040
.3	-3.79	10.65	6.86	1.03	12.86	.012	.024
.4	.017	6.81	6.83	1.79	12.60	.006	.024

between .05 and .057. Method SQ is unsatisfactory when the smallest sample size is 20. Even with $n_1 = n_2 = 50$, it can be unsatisfactory when $M = 10$. All indications are it should be used only when the smallest sample size is 50 and $M \geq 15$. Generally, given the goal of making inferences about the tail probabilities of D, again method T is preferable to method SQ.

4. Some Illustrations

The first illustration is based on data dealing with the emotional and physical wellbeing of older adults. The two groups of interest consist of participants who did not complete high school versus those that did complete high school. The corresponding sample sizes are 72 and 160. The goal is to compare these two groups based on a measure of depressive symptoms, CESD. Table 5 shows the results based on method Q. The column headed by q indicates the quantile that was used. The next two columns are the estimates of the q and 1-q quantile respectively, followed by their sums. The columns headed by ci.low and ci.up are the .95 confidence intervals for Δ. The p-values when testing (3) are in the next column followed by adjusted p-values using the Benjamini–Hochberg method. The estimates of Δ range between 6.35 and 7.05. Generally, a randomly sampled participant from the first group will have a higher CESD score than a randomly sampled participant from the second group by about 6 to 7 units.

Table 6 shows the results using method R. The column headed by Pts indicates the values for c_2 that were used, which correspond to the .6(.1).9 quantiles reported in Table 5. For example, the first value is $c_2 = 6.81$ in which case $c_1 = -6.81$. The results in Table 6 agree with the results Table 5 in the sense that the first group tends to have higher CESD scores. Table 6 merely adds perspective on the extent this is the case. The estimate of $P(X < Y)$ is .0194 and the p-value when testing $H_0 : P(X < Y) = 0$, based on the method derived by Cliff (1996), is .0001. Table 6 indicates that $P(X - Y < -15.6) = .11$, which provides more information regarding the extent the groups differ. But there is an estimated .2 probability that the difference is greater than 15.6, which reflects the extent the first group is more likely to have higher depressive symptoms.

Table 5. Result dealing with depressive symptoms using method R

Pts	P($x - y <$ −Pts)	P($x - y >$Pts)	Dif	ci.low	ci.up	p.value	p.adj
6.81	.24	.40	.16	.03	.30	.018	.043
10.65	.17	.30	.13	.02	.24	.032	.043
15.60	.11	.20	.09	.10	.18	.032	.043
22.38	.06	.11	.05	-.014	.12	.112	.112

The next illustration compares the BMI of females and males. The data are a subset of the 1206 Subject Release of the Human Connectome Project (Van Essen et al., 2013). Both sample sizes are 400. Table 7 shows the results using method Q where the first group corresponds to females. For example, the probability that the difference between the BMI for a randomly sampled female, minus the BMI for a randomly sampled male, is less than −9.12, is estimated to be .1. But there is no compelling evidence that Δ is less than zero. Significant results for Δ are obtained for $q = .3$ and .4 after adjusting the p-values.

Table 6. Results for the BMI data using method Q

q	q.est	(1-q).est	Sum	ci.low	ci.up	p.value	p.adj
.1	-9.12	8.82	-0.20	-2.05	1.47	.682	.682
.2	-6.27	4.79	-1.48	-2.91	-0.01	.050	.067
.3	-4.34	2.34	-2.00	-3.18	-0.59	.014	.028
.4	-2.69	.49	-2.20	-3.60	-0.96	.000	.000

Table 8 shows the results using method R where c_2 is again taken to be the upper quantiles used by method Q. Now three of the four tests reject at the .05 level after adjusting the p-values using Hochberg's method. This is in contrast to method Q that rejects only when $q = .3$ and .4. The results indicate that to some extent, there is a tendency for females to have a lower BMI than males, but there is considerable variation in the extent to which this is true. For example, column 2 in Table 8 indicates there is a .423 probability that a male would have a BMI greater than female's BMI by 2.34. But there is a .3 probability that a female's BMI is greater than a male's BMI by 2.34. The estimate of $P(X < Y)$ is .57 and the p-value when testing $H_0 : P(X < Y) = 0$ is .0001.

Table 7. Results for the BMI data using method R with c_2 values matching method Q

Pts	P($x - y <$ −Pts)	P($x - y >$Pts)	Dif	ci.low	ci.up	p.value	p.adj
8.82	.108	.100	-.008	-.045	.030	.626	.626
4.79	.274	.200	-.074	-.134	-.013	.012	.016
2.34	.423	.300	-.123	-.190	-.046	.004	.008
.49	.541	.399	-.142	-.231	-.065	.000	.000

5. Concluding Remarks

As seems fairly evident, there is no single method that tells us everything we would like to know about how two distributions differ. The only suggestion here is that methods Q, R and T provide details that help provide a deeper and more nuanced understanding of the data at hand.

As previously noted, extant results suggest that using the Harrell-Davis estimator helps deal with tied values compared to using an estimator based on one or two order statistics. The Harrell-Davis estimator performed well when using methods R, S and T, but in some situations this was not the case when using method SQ. That is, the Harrell-Davis estimator is not a panacea for dealing with tied values. Here, additional simulations were run where quantiles were estimated using a single order statistic. Control over the Type I error probability was highly unsatisfactory consistent with past studies.

Finally, access to R functions that apply the methods in this paper can be achieved by sourcing the file Rallfun-v43.txt, which can be downloaded from https://osf.io/xhe8u/ as well as https://zenodo.org/records/10420647. The R functions wmw.ref.dif and wmw.ref.mul perform method R. The first function is based on a single choice for c_1, in which case c_2 is taken to be $-c_1$, and it contains an option for plotting an estimate of the distribution of D. If no value for c_1 is specified, it is set equal to the estimate of the .25 quantile of D. The quantile that is used can be altered via the argument q. The second function can deal with multiple choices for the reference point c_1 or multiple choices for q. The R function wmw.QC.mul performs method Q and can handle multiple quantiles. The R function tailci.mul makes inferences about $P(D \leq c)$ using method T.

References

Akinshin, A. (2024). Trimmed Harrell-Davis quantile estimator based on the highest density interval of the given width. *Communications in Statistics-Simulation and Computation, 53*, 1565-1575. https://doi.org/10.1080/03610918.2022.2050396

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*(1), 289-300.

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

Cain, M., Zhang, Z., & Yuan, K.-H. (2017). Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence, and estimation. *Behavioral Research, 49*(5), 1716-1735. https://doi.org/10.3758/s13428-016-0814-1

Cliff, N. (1996). *Ordinal methods for behavioral data analysis*. Mahwah, NJ: Erlbaum.

Dielman, T., Lowry, C., & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in StatisticsłSimulation and Computation, 23*(2), 355-371.

Doksum, K. A., & Sievers, G. L. (1976). Plotting with confidence: Graphical comparisons of two populations. *Biometrika, 63*(3), 421-434. https://doi.org/10.2307/2335720

Greco, L., Luta, G., & Wilcox, R. (2023). On testing the equality between interquartile ranges. Computational Statistics. https://doi.org/10.1007/s00180-023-01415-8

Harrell, F. E., & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika, 69*(3), 635-640. https://doi.org/10.1093/biomet/69.3.635

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

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

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika, 75*(2), 383C386. https://doi.org/10.2307/2336190

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician, 50*(4), 361-365. https://doi.org/10.2307/2684934

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica, 46*(1), 33-49. 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*(437), 266-277. https://doi.org/10.2307/2291471

Lombard, F. (2005). Nonparametric confidence bands for a quantile comparison function. *Technometrics, 47*(3), 364-369.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. https://doi.org/10.1037/0033-2909.105.1.156

Özdemir, A. F., Yildiztepe, E., Paksoy, T., & Navruz, G. (2024). Searching the differences through the tails of distributions using an approach based on Mahalanobis distance and percentile bootstrap. Communications in Statistics - Simulation and Computation. https://doi.org/10.1080/03610918.2024.2333347

Navruz, G., & Özdemir, A. F. (2020). A new quantile estimator with weights based on a subsampling approach. *British Journal of Mathematical and Statistical Psychology, 73*(3), 506-521. https://doi.org/10.1111/bmsp.12198

Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics, 46*(1), 247-257. https://doi.org/10.2307/2531649

Salk, L. (1973). The role of the heartbeat in the relations between mother and infant. *Scientific American, 235*(4), 26-29.

Sfakianakis, M. E., & Verginis, D. G. (2008). A new family of nonparametric quantile estimators. *Communications in StatisticsłSimulation and Computation, 37*(2), 337-345.

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

Van Essen, D. C., Smith, S. M., Barch, D. M., Behrens, T. E., Yacoub, E., Ugurbil, K., & Wu-Minn HCP Consortium. (2013). The WU-Minn Human Connectome Project: An overview. *NeuroImage, 80*, 62-79. https://doi.org/10.1016/j.neuroimage.2013.05.041

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

Wilcox, R. R. (2023). *A guide to robust statistical methods*. New York: Springer. https://doi.org/10.1007/978-3-031-41713-9

Wilcox, R. R., & Rousselet, G. A. (2024). A quantile shift approach to main effects and interactions in a 2-by-2 design.

Methodology, 20. https://doi.org/10.5964/meth.12271

Wu, P.-C. (2002). Central limit theorem and comparing means, trimmed means, one-step M-estimators, and modified one-step M-estimators under nonnormality (Doctoral dissertation). University of Southern California.

Copyrights

Copyright for this article is retained by the author(s), with first publication rights granted to the journal.

This is an open-access article distributed under the terms and conditions of the Creative Commons Attribution license (http://creativecommons.org/licenses/by/4.0/).