Journal of Modern Applied Statistical Methods

Volume 11 | Issue 2 Article 2

11-1-2012

Comparing Two Independent Groups Via a Quantile Generalization of the Wilcoxon-Mann-Whitney Test

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

Journal of Modern Applied Statistical Methods
November 2012, Vol. 11, No. 2, 296-302

1538 – 9472/12/$95.00

Invited Article

Comparing Two Independent Groups Via a Quantile Generalization of the Wilcoxon-Mann-Whitney Test

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

The Wilcoxon-Mann-Whitney test, as well as modern improvements, are based in part on an estimate of $p = P(D < 0)$, where $D = X–Y$ and X and Y are independent random variables; a common goal is to test H_0: $p = 0.5$. This corresponds to testing H_0: $\xi_{0.5}$, where $\xi_{0.5}$ is the 0.5 quantile of the distribution of D. If the distributions associated with X and Y do not differ, then D has a symmetric distribution about zero. In particular, $\xi_q + \xi_{1-q} = 0$ for any $q \leq 0.5$, where ξ_q is the q^{th} quantile. Methods aimed at testing H_0: $p = 0.5$ are generalized by suggesting a method for testing H_0: $\xi_q + \xi_{1-q} = 0$, $q < 0.5$

Key words: Bootstrap methods, Harrell-Davis estimator, tests for symmetry, tied values, Well Elderly study.

Introduction

Consider two independent random variables, X and Y, let $D = X–Y$ and let τ_d, τ_x and τ_y be the population medians of D, X and Y, respectively. It is known that, under general conditions, the Wilcoxon-Mann-Whitney (WMW) test does not test H_0: $\tau_x = \tau_y$ (Fung, 1980). The WMW test is based on an estimate of $p = P(X < Y)$, but under general conditions it uses the wrong standard error, in contrast to more modern methods aimed at correcting this problem (Cliff, 1996; Brunner Munzel, 2000; Newcombe, 2006a, 2006b). The explicit goal of these improvements is making inferences about p, which includes the common goal of testing

$$H_0: p = 0.5. \quad (1)$$

Moreover, it is known, and fairly evident, that testing (1) corresponds to testing

$$H_0: \tau_d = 0. \quad (2)$$

Inferences about p and τ_d are important and useful, but a deeper understanding of how two independent groups compare would result by knowing something about the quantiles of the distribution of D.

Rand R. Wilcox is a Professor of Psychology. He is the author of eight textbooks on statistics, including *Modern Statistics for the Social and Behavioral Sciences* (2012, New York, CRC Press). Email him at: rwilcox@usc.edu.

For illustrative purposes, imagine that some experimental method is being compared to a control group and that D > 0 indicates that the experimental method is more effective than no treatment. If D has a skewed distribution, it is possible that p is approximately 0.5 and that testing (1) has relatively low power, yet there is a sense in which the experimental method is beneficial. Let ξ_q be the q^{th} quantile of D and assume, for example, that $\xi_{0.25} = -4$ and $\xi_{0.75} = 6$. Thus, for randomly sampled observations from each group, there is a sense in which the experimental treatment outweighs no treatment. If there are no benefits, then D should have a symmetric distribution about zero. In particular, it should be the case that

$$H_0: \xi_q + \xi_{1-q} = 0 \quad (3)$$

is true for any $q \leq 0.5$; consequently, this article suggests a method for testing (3).

Note that information about $\xi_q + \xi_{1-q}$ for a range of q values provides a more detailed sense about the distribution of D compared to using a single measure of location. For example, a portion of the study conducted by Jackson, et al. (2009) dealt with assessing the extent a particular intervention strategy reduced depression in older adults. An issue is whether the efficacy of the intervention changes as an individual moves from the center of the distribution of D to the tails. For the Jackson, et al. (2009) study, an estimate of the 0.9 quantile is 27.6 and the estimate of the 0.1 quantile is −19.7. That is, the drop in depression, 27.6, as reflected by the 0.9 quantile, exceeds the increase in depression, as reflected by the estimate of the 0.1 quantile, −19.7. For the 0.4 and 0.6 quantiles, the estimates are −1 and 5, again suggesting that intervention is useful, but the impact of intervention is less striking. If the distributions differ in terms of a measure of location only, it would be the case that $\xi_q + \xi_{1-q}$ does not vary with q.

For completeness, Wilcox and Erceg-Hurn (in press) considered the case where X and Y are dependent with two goals. The first is to compare the quantiles of the marginal distributions and the other is to test (3) but with D corresponding to the usual paired differences. Note that this differs from the situation at hand. For dependent groups, the goal is to assess changes within a subject in terms of the quantiles of D; here, the goal is make inferences about the difference between two randomly sample participants. A crude description of the method by Wilcox and Erceg-Hurn is that it generalizes the sign test for dependent groups. The suggestion is that a similar generalization of the Wilcoxon-Mann-Whitney test might be of interest. (Note that control over the Type I error probability is a function of both q and the sample sizes.) It was found that conditions under which good control over the Type I error probability is achieved differ to some degree from those when comparing dependent groups.

Description of the Proposed Method

A variety of methods for estimating the q^{th} quantile have been proposed, comparisons of which are reported by Parrish (1990), Sheather and Marron (1990) and Dielman, Lowry and Pfaffenberger (1994). The simplest approach is to estimate the q^{th} quantile using a single order statistic. Another approach is to use an estimator based on a weighted average of two order statistics while other estimators are based on a weighted average of all the order statistics. Regarding the issue of which estimator is best, the only certainty is that no single estimator dominates in terms of efficiency. For example, the Harrell and Davis (1982) estimator has a smaller standard error than the usual median when sampling from a normal distribution or a distribution that has relatively light tails, but for sufficiently heavy-tailed distributions, the reverse is true (Wilcox, 2012, p. 87).

Consider the special case where the goal is to estimate the population median. Currently all methods that are based in part on an estimate of the standard error of the usual sample median can perform poorly when tied values occur (Wilcox, 2006).

There are two problems: The first is obtaining a reasonably accurate estimate of the standard error. Many estimators have been proposed, all of which can be highly inaccurate when there are tied values. The second general concern is that, when tied, values occur the usual sample median is not necessarily asymptotically normal. Wilcox (2012) illustrated this result

when the cardinality of a sample space is relatively small. To date, the only method known to perform reasonably well in simulations is a slight generalization of the standard percentile bootstrap method (Wilcox, 2006). Thus, an obvious speculation is that when the goal is to make inferences about the quantiles of the distribution associated with D, the same percentile bootstrap method might perform well. However, simulations indicate that this is not necessarily the case.

Let n_j be the sample size for the j^{th} group (j = 1, 2). Consider, for example, the situation where $n_1 = 20$, $n_2 = 30$ and observations are generated from a binomial distribution with probability of success 0.4 and when the sample space is 0(1)7. When testing at the 0.05 level, simulations indicate that the actual level is approximately 0.102. Due to the difficulty of not being able get a reasonably accurate estimate of the standard error when sampling from a discrete distribution, bootstrap methods based in part on an estimate of the standard error hold little promise.

Here, the one method that performed well in simulations was based in part on the estimator derived by Harrell and Davis (1982) that estimates the q^{th} quantile using a weighted average of all the order statistics. More precisely, let Y be a random variable having a beta distribution with parameters a = (n + 1)q and b = (n + 1)(1 − q). That is, the probability density function of Y is

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1},$$

where Γ is the gamma function.

Let

$$W_i = P((i-1)/n \le Y \le i/n).$$

For the random sample $X_1, \ldots, X_n$, let $X_{(1)} \le \ldots \le X_{(n)}$ denote the observations written in ascending order. The Harrell-Davis estimate of ξ_q is $\hat{\xi}_q = \sum W_i X_{(i)}$. In terms of its standard error, Sfakianakis and Verginis (2006) show that in some situations the Harrell-Davis estimator competes well with alternative estimators that use a weighted average of all the order statistics, but there are exceptions. For example, Sfakianakis and Verginis (2006) derived alternative estimators that have advantages over the Harrell-Davis in some situations, but it was found that when sampling from heavy-tailed distributions the standard errors of their estimators can be substantially larger than the standard error of Harrell-Davis estimator.

To describe the details of the proposed test of (3), let $X_1, \ldots$, and $Y_1, \ldots$, be random samples of size n_1 and n_2, respectively, and let $D_{ik} = X_i - Y_k$ (i = 1, …, n_1; k = 1, …, n_2). The q^{th} quantile of distribution of D, δ_q, is estimated via the Harrell-Davis estimator, applied to the D_{ik} values, yielding $\widehat{D}_q$. Next, generate a bootstrap sample from the j^{th} group by resampling with replacement n_j observations from group j. Let $\widetilde{D}_q$ be the estimate of q^{th} quantile of D based on these bootstrap samples and let $d = \widetilde{D}_q + \widetilde{D}_{1-q}$. Repeat this process B times yielding d_b, b = 1, …, B; here, B = 1,000 is used. Let $\ell = \alpha$ B/2, rounded to the nearest integer, and let $u = B - \ell$. Letting $d_{(1)} \le \ldots \le d_{(B)}$ represent the B bootstrap estimates written in ascending order, an approximate $1 - \alpha$ confidence interval for $\delta_q + \delta_{1-q}$ is $(d_{(\ell+1)}, d_{(u)})$. This will be called method DHD.

Let A denote the number of times d is less than zero and let C be the number of times d = 0. Letting

$$\hat{p} = \frac{A + .5C}{B},$$

a (generalized) p-value is $2\min(\hat{p}, 1-\hat{p})$ (Liu & Singh, 1997).

Results

Simulations were used to study the small-sample properties of method DHD. The sample sizes considered were (n_1, n_2) = (10, 10), (20, 20), (10, 30) and (20, 30). Estimated Type I error probabilities were based on 2,000 replications. Two values for q were considered: 0.25 and 0.1. Both continuous and discrete distributions were used. The four continuous distributions were normal, symmetric and heavy-tailed, asymmetric and light-tailed and asymmetric and heavy-tailed. More precisely, four g-and-h distributions

were used (Hoaglin, 1985) that contain the standard normal distribution as a special case. If Z has a standard normal distribution, then

$$W = \frac{\exp(gZ)-1}{g}\exp(h\frac{Z^2}{2}), \text{ if } g > 0,$$
$$= Z\exp(h\frac{Z^2}{2}), \text{ if } g = 0$$

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).

Table 1: Some Properties of the g-and-h Distribution

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

To gain perspective on the effects of tied values, data were generated from a discrete distribution having a sample space consisting of the integers 0 through 7; more precisely, data were generated from a binomial distribution with probability of success equal to 0.4. First consider the four g-and-h distributions when testing at the 0.05 level and $n_1 = n_2 = 10$. As indicated in Table 2, if q = 0.25, in which case the goal is to test (3) with q = 0.25, then $\hat{\alpha}$, the probability of a Type I error, is estimated to be close to the nominal level. Note that the estimates barely change among the continuous distributions considered. However, when q = 0.1, the estimated Type I error probability can exceed 0.1. Increasing one of the sample sizes to 30 improves the estimate, but it still exceeds 0.075. Although the seriousness of a Type I error can depend on the situation, Bradley (1978) suggested that, as a general guide, when testing at the 0.05 level the actual level should not exceed 0.075. With $n_1 = 20$ and $n_2 = 40$, again the estimate can exceed 0.1. With $n_1 = n_2 = 30$ (not shown in Table 2), reasonably accurate control over the probability of Type I error is achieved. Increasing both sample sizes to 40, the probability of Type I error is estimated to be between 0.045 and 0.051 among all situations considered.

Generating data from the binomial distribution gave results similar to those in Table 2. For $n_1 = n_2 = 10$ and q = 0.25, $\hat{\alpha} = 0.065$. For $n_1 = n_2 = 20$ $\hat{\alpha} = 0.056$ and 0.063 for q = 0.25 and 0.1, respectively. For $n_1 = 20$ and $n_2 = 30$ the estimates are 0.056 for both q = 0.25 and q = 0.1.

How the power of method DHD compares to other methods depends in part on the nature of the distributions being compared. As is evident, different methods are sensitive to different features of the data. However, to provide at least some perspective, some results are reported when distributions differ in location only. In particular, consider $D = X - Y + \lambda$ for some constant λ where both X and Y have mean zero and variance one. Under normality, it can be seen that $\delta_q + \delta_{1-q} = 2\lambda$. Thus, when comparing means, rather than testing (3), this suggests that method DHD might have relatively high power under normality despite the sample mean having a smaller standard error than the Harrell-Davis estimator. Table 3 reports some simulation power estimates when q = 0.25. The column headed by Welch indicates the estimated power when using the method from Welch (1938) to test the hypothesis of equal means. As shown, the power of method DHD compares well to Welch's method – and that DHD seems to have a slight advantage.

Table 2: Estimated Type I Error Probability, $\alpha = 0.05$

q	n_1	n_2	g	h	$\hat{\alpha}$
0.25	10	10	0.0	0.0	0.069
			0.0	0.2	0.066
			0.2	0.0	0.072
			0.2	0.2	0.073
	20	20	0.0	0.0	0.060
			0.0	0.2	0.056
			0.2	0.0	0.060
			0.2	0.2	0.058
	10	30	0.0	0.0	0.082
	20	30	0.0	0.0	0.062
0.10	10	10	0.0	0.0	0.092
			0.0	0.2	0.104
			0.2	0.0	0.091
			0.2	0.2	0.108
	20	20	0.0	0.0	0.065
			0.0	0.2	0.069
			0.2	0.0	0.065
			0.2	0.2	0.067
	20	30	0.0	0.0	0.060

An Illustration

Consider the Jackson, et al. (2009) study described in the introduction that used sample sizes of 232 and 140. Figure 1 shows an estimate of $\delta_q + \delta_{1-q}$, indicated by *, as a function of q, where the q values are 0.05(0.05)0.40. The corresponding p-values are 0.002, 0.004, 0.008, 0.010, 0.016, 0.020, 0.020 and 0.020. The + above and below the * indicate a 0.95confidence interval. These results suggest that intervention is effective and that this is the case particularly in terms of more extreme quantiles.

Conclusion

In terms of controlling the probability of a Type I error, method DHD generally performs well in simulations. The restriction is that as q approaches zero larger samples size are needed, particularly when the sample sizes are unequal. For $n_1 = n_2 = 10$, all indications are that method DHD performs reasonably well for $q \geq 0.2$. For $n_1 = 10$ and $n_2 = 30$, this is not the case, however, for $\min(n_1, n_2) \geq 20$, control over the Type I error probability was found to be reasonably satisfactory.

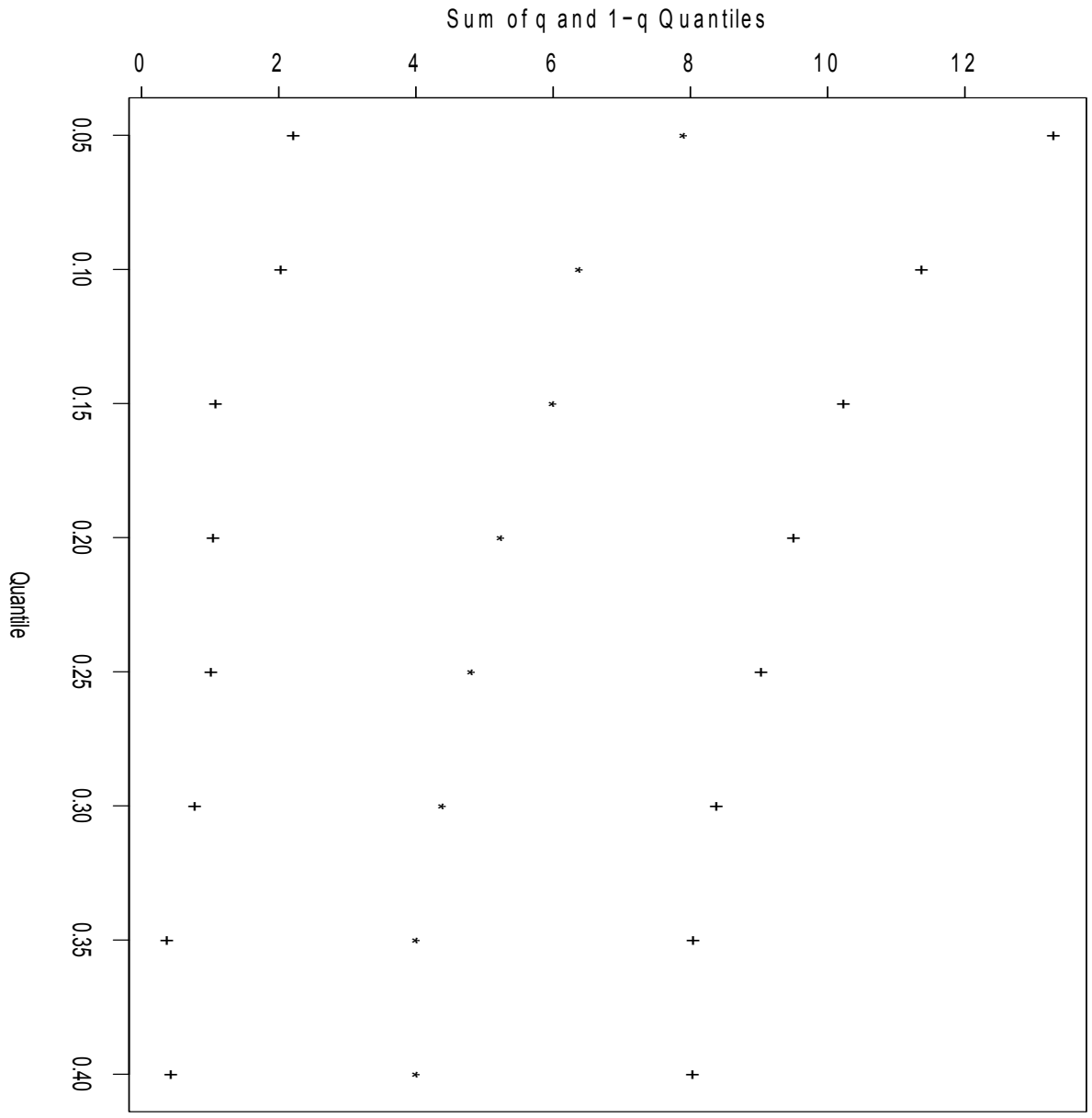
It is not suggested that method DHD should be used to the exclusion of all other techniques aimed at comparing two independent groups. Rather, the suggestion is that multiple techniques are needed to obtain a good understanding of how two groups compare and the DHD method helps achieve this goal.

Finally, method DHD can be applied with the R function cbmhd. The R function qwmwhd applies the method using a range of q values. The plot in Figure 1 was created with the latter function.

Table 3: Estimated Power, $\alpha = 0.05$, $\lambda = 1$

q	n_1	n_2	g	h	DHD	WELCH
0.25	10	10	0.0	0.0	0.62	0.55
			0.0	0.2	0.60	0.54
			0.2	0.0	0.43	0.36
			0.2	0.2	0.42	0.35

Figure 1: Estimates of $\xi_q + \xi_{1-q}$



These R functions are included in a package that can be downloaded from http:// college.usc.edu/labs/rwilcox/home.

References

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

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

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 Statistics-Simulation and Computation*, *23*, 355-371.

Fung, K. Y. (1980). Small sample behaviour of some nonparametric multi-sample location tests in the presence of dispersion differences. *Statistica Neerlandica*, *34*, 189-196.

Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika*, 69, 635-640.

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

Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S., Chou, C.-P., Jordan-Marsh, M., Forman, T., White, B., Granger, D., Knight, B., & Clark, F. (2009). Confronting challenges in intervention research with ethnically diverse older adults: The USC Well Elderly II trial. *Clinical Trials*, *6*, 90-101.

Liu, R. G. & Singh, K. (1997). Notions of limiting P values based on datadepth and bootstrap. *Journal of the American Statistical Association*, 92, 266-277.

Newcombe, R. G. (2006a). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 1: General issues and tail-area-based methods. *Statistics in Medicine*, *25*, 543-557.

Newcombe, R. G. (2006b). Confidence intervals for an effect size measure based on the Mann-Whitney statistic. Part 2: Asymptotic methods and evaluation. *Statistics in Medicine*, *25*, 559-573.

Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, *46*, 247-257.

Sfakianakis, M. E. & Verginis, D. G. (2006). A New Family of Nonparametric Quantile Estimators. *Communications in Statistics-Simulation and Computation*, *37*, 337-345.

Sheather, S. J. & Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85, 410-416.

Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, *29*, 350-362.

Wilcox, R. R. (2006). Comparing medians. *Computational Statistics & Data Analysis*, *51*, 1934-1943.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (*3rd Edition*). San Diego, CA: Academic Press.