COMPARING TWO DEPENDENT GROUPS VIA QUANTILES

Rand R. Wilcox
Dept of Psychology
University of Southern California
and
David M. Erceg-Hurn
School of Psychology
University of Western Australia

September 14, 2012

ABSTRACT

The paper considers two general ways dependent groups might be compared based on quantiles. The first compares the quantiles of the marginal distributions. The second focuses on the lower and upper quantiles of the usual difference scores. Methods for comparing quantiles have been derived that typically assume sampling is from a continuous distribution. There are exceptions, but generally, when sampling from a discrete distribution where tied values are likely, extant methods can perform poorly, even with a large sample size. One reason is that extant methods for estimating the standard error can perform poorly. Another is that quantile estimators based on a single order statistic, or a weighted average of two order statistics, are not necessarily asymptotically normal. Our main result is that when using the Harrell-Davis estimator, good control over the Type I error probability can be achieved in simulations via a standard percentile bootstrap method, even when there are tied values, provided the sample sizes are not too small. In addition, the two methods considered here can have substantially higher power than alternative procedures. Using real data, we illustrate how quantile comparisons can be used to gain a deeper understanding of how groups differ.

Keywords: Tied values, bootstrap methods, Harrell–Davis estimator, test for symmetry, Well Elderly study

1 Introduction

When comparing two dependent groups, certainly one of the more obvious strategies is to compare the marginal distributions in terms of some measure of location intended to reflect the typical response. However, differences in the tails of the marginal distributions can be of interest as well. Consider, for example, the study by Jackson et al. (2009). One goal was to assess the effectiveness of intervention in terms of reducing depressive symptoms in older adults. Prior to intervention, depressive symptoms were measured with the Center for Epidemiologic Studies Depression Scale (CESD) developed by Radloff (1977) and six months later after receiving intervention. A paired t test indicates a significant difference ($p = .0045$). However, of interest is whether intervention has a differential impact depending on how depressed an individual might be. In particular, is intervention more or less effective among individuals who have a relatively high level of depression? One way of addressing this issue is to compare the differences between the upper quantiles of the marginal distributions. Of course, comparing the lower quantiles can be of interest as well. More formally, let θ_{qj} be the qth quantile corresponding to the jth marginal distribution. Then a goal of interest is testing

$$H_0 : \theta_{q1} = \theta_{q2} \tag{1}$$

or computing a $1 - \alpha$ confidence interval for $\theta_{q1} - \theta_{q2}$.

A review of methods for comparing the median of the marginal distributions can be found in Wilcox (2012). An important point is that when there are tied values, all methods based on an estimate of the standard error of the usual sample median can be unsatisfactory. One reason has to do with obtaining a reasonably accurate estimate of the standard error. Several methods for estimating the standard have been proposed, comparisons of which are reported in Price and Bonett (2001). But under general conditions, when there are tied values, all of these estimators can perform poorly. An alternative strategy is to use a bootstrap estimate of the standard error, but again tied values can wreak havoc. As might be expected, this problem remains when estimating other quantiles with a single order statistic or a weighted average of two order statistics.

Yet another fundamental concern is that when there are tied values, the usual sample median is not necessarily asymptotically normal (e.g., Wilcox, 2012, section 4.6.1). Not surprisingly, this problem persists when estimating other quantiles with a single order statistic or a weighted average of two order statistics. A strategy for dealing with tied values, when testing hypotheses, is to use a percentile bootstrap method, which does not require an estimate of the standard error. When comparing independent groups via the usual sample median, all indications are that the percentile bootstrap method in Wilcox (2012, section 5.4.2) performs reasonably well, even with tied values. But when comparing the marginal medians corresponding to two dependent groups, using an obvious modification of the method used to compare independent groups, the actual level can drop well below the nominal level. For example, when testing at the .05 level, for the discrete distributions used in the simulations reported in section 3 of this paper, the actual level drops as low as .012.

Lombard (2005) derived a distribution free method for comparing the difference between all of the quantiles. The probability of making one or more Type I errors, when comparing all of the quantiles, can be determined exactly assuming random sampling only. In terms of comparing quantiles close to the median, its power compares well to other methods that might be used. However, when there is interest in the quantiles close to zero or one, the lengths of the confidence intervals can be extremely large. Indeed, it is common to have either the lower end of the confidence interval equal to $-\infty$ or the upper end equal to ∞. Also, power can be relatively low, even when comparing the quartiles, as illustrated in section 3. Consequently, there is interest in finding a method that controls the Type I error probability reasonably well, even when there are tied values, and simultaneously competes well with Lombard's method in terms of power.

1.1 An Alternative Perspective

There is an alternative perspective that deserves attention. There are exceptions, but note that under general conditions, the difference between the marginal medians is not equal to median of the difference scores. More formally, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample of size n from some bivariate distribution and let $D_i = X_i - Y_i$, $i = 1, \ldots, n$. Let δ_q denote the qth quantile of the distribution of D and let $\theta_{.5,j}$ be the population median associated with the jth marginal distribution. Then under general conditions $\delta_{.5} \neq \theta_{.5,1} - \theta_{.5,2}$.

To express this property in more practical terms, consider again the Jackson et al. study. If the focus is on the marginal medians, the issue is whether the typical level of depression prior to intervention differs from the typical level after intervention. But another issue is to what extent intervention is effective within each individual. That is, for some participants, depression might decrease after intervention, but for others depression might actually increase. One way of dealing with this is to test

$$H_0 : \delta_{.5} = 0 \tag{2}$$

rather than (1).

More generally, is there some sense in which the decrease in depression outweighs any increase seen in some participants? Note that a broader way of saying that intervention is completely ineffective is to say that the distribution of D is symmetric about zero. Let

θ_q denote the qth quantile of the distribution of D. Then another way of characterizing the difference between time 1 and time 2 measures is in terms of how θ_q compares to θ_{1-q}, $q < .5$. In the Jackson et al. study, for example, the estimate of $\theta_{.75}$ is $\hat{\theta}_{.75} = 5.57$ and the estimate of $\theta_{.25}$ is $\hat{\theta}_{.25} = -3.19$, suggesting that the drop in depression characterized by the .75 quantile is larger than the corresponding increase in depression represented by the .25 quantile. An issue, then, is whether one can reject H_0: $-\theta_{.25} = \theta_{.75}$, which would support the conclusion that there is a sense in which the positive effects of intervention outweigh the negative effects. More generally, there is interest in testing

$$H_0 : \theta_q + \theta_{1-q} = 0,\ q < .5. \tag{3}$$

Of course, one could test the hypothesis that D has a symmetric distribution about zero using the sign test. It is evident, however, that the magnitude of $\theta_{1-q} + \theta_q$ might be sufficiently high as to result in more power testing (3). Also, while $p = P(D > 0)$ is certainly of interest, the extent $\theta_{1-q} + \theta_q > 0$ can provide a useful perspective.

2 Description of the Proposed Methods

When testing (1) or (3), a basic issue is choosing a quantile estimator. A variety of methods for estimating the qth quantile have been proposed, comparisons of which are reported by Parrish (1990), Sheather and Marron (1990), as well as Dielman, Lowry and Pfaffenberger (1994). The simplest approach is to estimate the qth quantile using a single order statistic. Another 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. As might be expected, 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). Sfakianakis and Verginis (2006) show that in some situations the Harrell–Davis estimator competes well with alternative estimators that again use a weighted average of all the order statistics, but there are exceptions. (Sfakianakis and Verginis derived alternative estimators that have advantages for the Harrell–Davis in some situations. But we found that when sampling from heavy-tailed distributions, the standard error of their estimators can be substantially larger than the standard error of the Harrell–Davis estimator.) Preliminary simulations found that in terms of Type I errors, given the goal of comparing quantiles, estimators based on a single order statistic, or a weighted average of two order statistics, are unsatisfactory. In contrast, the Harrell–Davis estimator, which uses a weighted average of all the order statistics, performed well in the simulations, so this estimator is used in the remainder of the paper.

To describe the Harrell–Davis estimate of the qth quantile, let U 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 U is

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

where Γ is the gamma function. Let

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

For the random sample $X_1, \ldots, X_n$, let $X_{(1)} \leq \ldots \leq X_{(n)}$ denoted the observations written in ascending order. The Harrell–Davis estimate of θ_q, the qth quantile, is

$$\hat{\theta}_q = \sum_{i=1}^{n} W_i X_{(i)}. \tag{4}$$

2.1 Method M

First consider the goal of testing (1). Generate a bootstrap sample by resampling with replacement n pairs of observations from $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let $\hat{\theta}_j^*$ be the Harrell–Davis estimate of the qth quantile for jth marginal distribution based on this bootstrap sample and let $d_j^* = \hat{\theta}_1^* - \hat{\theta}_2^*$. Repeat this process B times yielding d_b^*, $b = 1, \ldots, B$. Here, $B = 2000$ is used. Let $\ell = \alpha B/2$, rounded to the nearest integer, and let $u = B - \ell$. Letting $d_{(1)}^* \leq \cdots \leq d_{(B)}^*$ represent the B bootstrap estimates written in ascending order, an approximate $1-\alpha$ confidence interval for $\theta_1 - \theta_2$ is

$$(d_{(\ell+1)}^*, d_{(u)}^*).$$

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

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

a (generalized) p-value is $2\min(\hat{p}^*, 1 - \hat{p}^*)$ (Liu & Singh, 1997). This will be called method M.

2.2 Method D

When testing (3), proceed in a manner similar to method M. More precisely, generate a bootstrap sample as before and let D_i^* $(i = 1, \ldots, n)$ be the resulting difference scores. Let $\hat{\delta}_q^*$ be the Harrell–Davis estimate of the qth quantile based on this bootstrap sample and let $\hat{\Delta}^* = \hat{\delta}_q^* + \hat{\delta}_{1-q}^*$ $(q < .5)$. Repeat this process B times yielding $\hat{\Delta}_b^*$ $(B = 1, \ldots, B)$. Then an approximate $1 - \alpha$ confidence interval for $\theta_q + \theta_{1-q}$ is

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

and a p-value is given by (5) only now A is the number of times $\hat{\Delta}^*$ is less than zero and C is the number of times $\hat{\Delta}^* = 0$.

Note that if the two groups being compared differ in a shift in location only, then $\hat{\Delta}^*$ should differ very little as a function of q. That is, $\hat{\Delta}^*$, computed for a range of q values, provides some sense in which a simple shift in location model adequately describes how two groups compare.

3 Simulation Results

Simulations were used to study the small-sample properties of methods M and D. The sample sizes considered were 20, 30 and 40. Estimated Type I error probabilities, $\hat{\alpha}$, were based on 2000 replications. Both continuous and discrete distributions were used with data generated via the R function rmul, which is stored in the R package WRS and can be installed with the R command install.packages("WRS", repos="http://R-Forge.R-project.org"). The four continuous (marginal) distributions were normal, symmetric and heavy-tailed, asymmetric

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

g	h	κ_1	κ_2
0.0	0.0	0.00	3.0
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

and light-tailed, and asymmetric and heavy-tailed. The correlation between X and Y was taken to be either 0 or .7. More precisely, the marginal distributions were taken to be one of four g-and-h distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. (When using the R function rmul, setting the argument mar.fun=ghdist results in observations being generated from a g-and-h distribution.) If Z has a standard normal distribution, then

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

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 ($g = h = 0.0$), 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).

To elaborate on how the correlation between X and Y was simulated, let R be the correlation matrix and form the Cholesky decomposition $U'U = R$, where U is the matrix of factor loadings of the principal components of the square-root method of factoring a correlation matrix, and U' is the transpose of U. Next, an $n \times 2$ matrix of data, say V, was randomly generated from a specified distribution. Then the matrix product VU produces an $n \times 2$ matrix of data that has population correlation matrix R.

Table 2 reports the estimated Type I error probabilities when testing at the .05 level and $n = 20$. As indicated, when comparing the quartiles ($q = .25$ and .75), control over the Type I error probability is estimated to be close to the nominal level. Note that the estimates barely change among the continuous distributions considered. However, when the

Table 2: Estimated Type I Error Probability, Method D, continuous case, $\alpha = .05$

				$\hat{\alpha}$	
q	n	g	h	$\rho = 0.0$	$\rho = 0.7$
0.75	20	0.0	0.0	0.050	0.048
		0.0	0.2	0.049	0.046
		0.2	0.0	0.051	0.053
		0.2	0.2	0.050	0.048
0.25		0.2	0.0	0.059	0.048
0.25		0.2	0.2	0.061	0.044
0.90	30	0.0	0.0	0.061	0.062
		0.0	0.2	0.069	0.068
		0.2	0.0	0.064	0.062
		0.2	0.2	0.071	0.065
0.10		0.2	0.0	0.062	0.062
		0.2	0.2	0.068	0.069

goal is to compare the .9 quantiles, not shown in Table 2, now $\hat{\alpha}$ can exceed .1. As indicated in Table 2, increasing the sample size to 30 improves matters and as expected, increasing n to 40 improves the control over the Type I error probability even more. Although the seriousness of a Type I error can depend on the situation, Bradley (1978) has suggested that as a general guide, when testing at the .05 level, the actual level should not exceed .075 and all indications are that this goal is achieved with $n \geq 30$ and the goal is to compare the .9 quantiles.

When dealing with the .1 and .9 quantiles, increasing the sample size to 40 results in estimated Type I error probabilities close to the nominal level. Note, for example, that in Table 2 the highest estimate is .071 and occurs when comparing the .9 quantiles with $n = 30$ and $g = h = .2$. Increasing the sample size to 40, the estimate is .049.

To gain perspective on the effects of tied values, data were generated from a bivariate distribution with the marginal distributions having the g-and-h distributions indicated in Table 1, then each value was multiplied by 5 and rounded down to the nearest integer.

Table 3: Estimated probability of a Type I error, discrete case, $\alpha = .05$

				$\hat{\alpha}$	
q	n	g	h	$\rho = 0.0$	$\rho = 0.7$
0.75	20	0.0	0.0	0.061	0.051
		0.0	0.2	0.058	0.046
		0.2	0.0	0.058	0.049
		0.2	0.2	0.062	0.046
0.25		0.2	0.0	0.052	0.048
0.25		0.2	0.2	0.044	0.047
0.90	30	0.0	0.0	0.062	0.058
		0.0	0.2	0.068	0.066
		0.2	0.0	0.064	0.065
		0.2	0.2	0.070	0.065
0.10		0.2	0.0	0.062	0.060
		0.2	0.2	0.070	0.066

Table 3 summarizes the estimated Type I error probabilities, again testing at the .05 level.

Tables 4 and 5 report the simulation results using method D. Altering the correlation does not alter the estimated Type I error probability, so for brevity only the results for $\rho = 0$ are reported. As can be seen, the results are similar to those in Tables 2 and 3: for $q = .25$, control over the Type I error probability is very good with $n = 20$. For $q = .1$, the Type I error probability is unsatisfactory for $n = 20$ (not shown in Tables 4 and 5), but increasing n to 30, reasonably good control is achieved.

To provide some sense of how the power of methods M and D compare to the power of the method derived by Lombard (2005), data were generate from two normal distributions both having variance one, $\rho = 0$, the first marginal distribution had a mean of 0 and the second a mean of 1. Comparing the .25 quantiles at the .05 level, power was estimated to be 0.81 using method M with $n = 25$. For method D, power was estimated to be 0.88. Power using Lombard's method was estimated to be 0. Again, Lombard's method performs relatively well, in terms of power, given the goal of detecting differences between the quantiles close to

Table 4: Estimated Type I Error Probability, Method D, $\alpha = .05$

q	n	g	h	$\rho = 0.0$
0.25	20	0.0	0.0	0.057
		0.0	0.2	0.054
0.25		0.2	0.0	0.052
0.25		0.2	0.2	0.053
0.10	30	0.0	0.0	0.066
		0.0	0.2	0.068
		0.2	0.0	0.062
		0.2	0.2	0.067

Table 5: Estimated probability of a Type I error, Method D, discrete case, $\alpha = .05$

q	n	g	h	$\rho = 0.0$
0.25	20	0.0	0.0	0.059
		0.0	0.2	0.056
		0.2	0.0	0.052
		0.2	0.2	0.053
0.10	30	0.0	0.0	0.067
		0.0	0.2	0.067
		0.2	0.0	0.062
		0.2	0.2	0.068

the population median. But in terms of detecting differences when comparing quartiles or when q is relatively close to zero or one, power is poor.

4 Some Illustrations

Consider again the study comparing depressive symptoms that was described in the introduction. The sample size is $n = 326$. Figure 1 shows the estimated difference between the deciles of the marginal distributions. These estimates suggest that among participants with higher CESD (depression) scores, the more effective is intervention. (The pluses indicate a .95 confidence interval for the differences between the corresponding decile.) Comparing the .9 quantiles, the p-value is less than .001 and for the .8 quantile the p-value is .036; the p-values for the other deciles are greater than .05.

Using method D instead, for the quantiles .05(.05).40, the p-values range between .010 and .052 as indicated in Table 6, where $\hat{\Delta}$ indicates the estimate of $\theta_q+\theta_{1-q}$. So by Hochberg's (1988) improvement on the Bonferroni inequality, all eight tests are significant with a familywise Type I error rate of .052. Moreover, for each of the eight quantiles that were compared, the estimate of $\hat{\Delta}$ is positive and decreases as q increases indicating that within subjects, intervention is typically effective at reducing depression, particularly among participants who have relatively high depression prior to intervention. In contrast, for a control group measured again after six months, the estimate is negative for q=.05(.05).20 but not significant. That is, there are nonsignificant indications that typically, among the more depressed individuals depression gets a bit worse without intervention. Also note that for q=.25(.05).40, the estimate of $\hat{\Delta}$ is very close to zero suggesting that among the participants in the control group with relatively low depression, levels of depression change very little six months later.

A further illustration of the utility of the new methods comes from an experiment conducted by the second author. The purpose of the experiment was to learn about the effects of pharmaceutical companies advertising antidepressant medications to the public. These ads try to convince people that depression is caused by an imbalance of chemicals in the brain, such as serotonin, that can be corrected by using antidepressant medications. How effective the ads are at convincing people that depression is caused by a "chemical imbal-

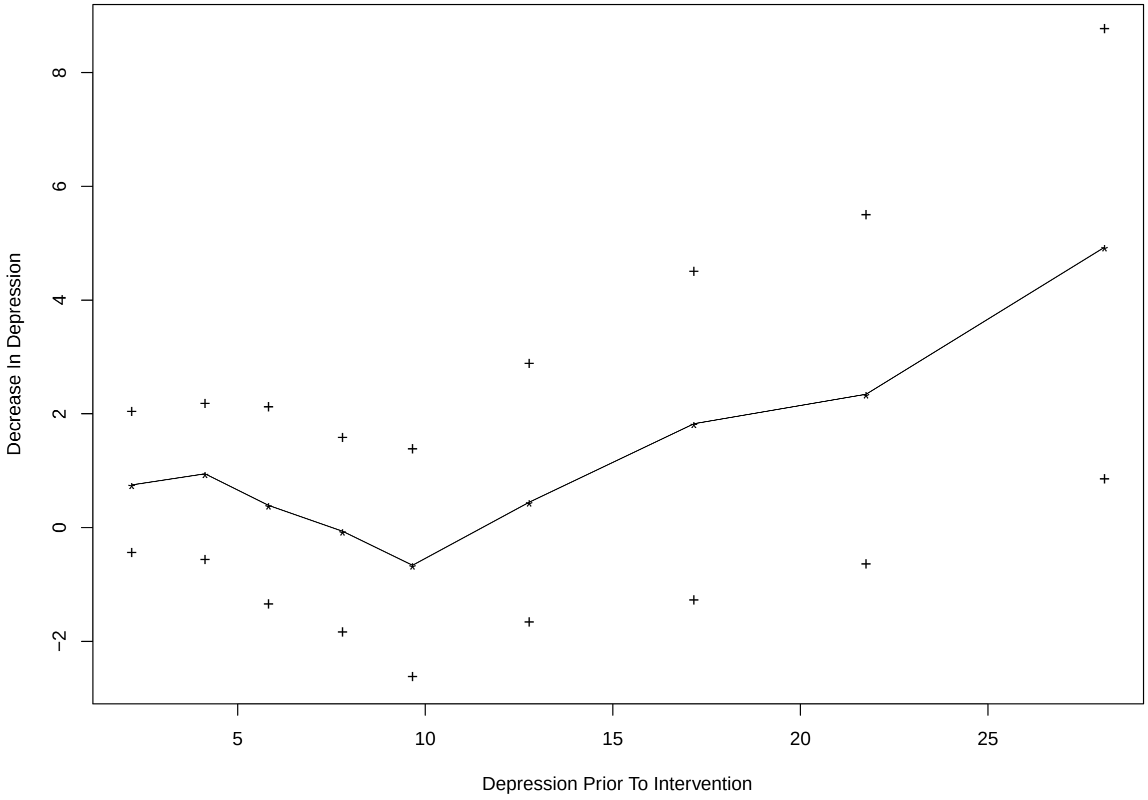

Figure 1: Decrease in depression is indicated by the y-axis. The x-axis indicates the level of depression prior to intervention

Table 6: Method D applied to the experimental group and a control group using data from the intervention study

	Exper. Group		Control Group	
q	$\hat{\Delta}$	p.value	$\hat{\Delta}$	p.value
0.05	5.448	0.024	-2.669	0.200
0.10	2.680	0.036	-3.402	0.114
0.15	2.441	0.052	-2.310	0.210
0.20	2.445	0.010	-1.041	0.494
0.25	2.385	0.014	0.001	0.986
0.30	1.961	0.022	0.364	0.750
0.35	1.638	0.020	0.099	0.948
0.40	1.629	0.052	-0.019	0.960

ance" is unclear. In the experiment, participants suffering from depression were randomly assigned to view several ads for antidepressant medications, or to a control group that saw no ads. The participants then completed a questionnaire. One of the questionnaire items asked participants to rate how likely they think it is that depression is caused by an imbalance of chemicals in the brain. Participants answered this question using a scale that ranged from 1 (very unlikely) to 9 (very likely). Three months after the initial phase of the study, a subset of 492 participants who had been exposed to antidepressant ads responded again to the questionnaire. The median of the difference scores was 0 suggesting that the typical participant's ratings were very stable between time 1 and time 2. However, applying Method D to the .05(.05).40 quantiles of the difference scores distribution, the p values for every quantile are less than .001. The estimate of $\theta_q + \theta_{1-q}$ is approximately 1 for all of the quantiles except for q=.40, where it drops to .07. These findings suggest that scores near the center of the difference score distribution are close to zero, which is consistent with the median of difference scores analysis. However, scores in the upper tail of the distribution tend to be about twice as large as at the corresponding quantile in the lower tail of the distribution. This suggests that among participants whose ratings changed over time, the ratings made at time 1 tended to be higher than those made 3 months later. So while the "typical" participant in the study may have stable scores, there is a subset of participants for whom the effect produced by the ads seems to have been lost over time.

In this particular case, method M is a less informative in the sense that it does not involve a direct comparison of how participants scores changed over time. Nevertheless, applying method M to the deciles we find that ratings at time 2 are typically a little lower than at time 1, which is consistent with the results obtained using method D. The p values were .025 or less for all deciles between .2 and .8. In contrast, if Lombard's (2005) method for comparing all quantiles is used, only the .563-.608 quantiles differ significantly. Moreover, the confidence intervals produced by method M are shorter than those produced by Lombard's method. For example, for $q = .10$, the .95 confidence interval produced by method M is $(-.001, .080)$, whereas using Lombard's method it is $(-3, 0)$. These findings underscore the fact that method M can be considerably more powerful than Lombard's procedure.

5 Concluding Remarks

In summary, when comparing the lower or upper quantiles of the marginal distribution of two dependent groups, method M was found to perform well in terms of both Type I errors and power, even when there are tied values. The only restriction is that the sample size must be reasonably large to ensure that the Type I error probability is reasonably close to the nominal level. When the goal is to compare the quartiles, $n = 20$ appears to suffice. When comparing the .9 quantiles, $n \geq 30$ is required. A similar result was obtained when using method D. Currently, when there are tied values, no other method has been found that performs reasonably well. Also, the power associated with methods M and D can be substantially higher than the power of Lombard's method when there is direct interest in the quartiles or quantiles even closer to zero or one.

As was illustrated, the choice between methods M and D is not academic. Presumably the choice between these two methods might be dictated in part by what is important in a given situation. That is, they answer different questions and so an issue is whether one method addresses a more meaningful question compared to the other.

An issue is how best to control the probability of one or more Type I errors when testing two or more hypotheses. As was done in the illustration, a simple approach is to use some improvement on the Bonferroni method such as the sequentially rejective technique derived by Hochberg (1988). However a possible concern is that as the number of hypotheses increases, the expectation is that at some point the actual probability of one or more Type I errors will be substantially less than the nominal level in which case Lombard's method might have more power. Suppose method D is used to test (3) for each $q = .05(.05).40$. Under normality, with $n = 40$ and $\rho = 0$, the probability of one or more Type I errors was estimated to be .040 with the nominal family wise error probability taken to be .05. So perhaps for a fairly wide range of situations, Hochberg's method performs reasonably well, but a more definitive study is needed.

Yet another issue, motivated by the results in Table 6, is dealing with a 2-by-2, between-by-within design. In Table 6, for the control group, (3) is not rejected for any q, the lowest p-value being .116 . In contrast, for the experimental group, using Hochberg's method, all eight hypotheses are significant at the .052 level. Let $\Delta_1 = \theta_{1-q} + \theta_q$ for the control group and let Δ_2 be the corresponding parameter for the experimental group. Of interest is testing H_0: $\Delta_1 = \Delta_2$. An obvious guess is that a simple extension of the bootstrap method used here would suffice, but this is in need of further study.

Finally, R functions Dqdif and difQpci can be used to apply method D and the R function Dqcomhd applies method M. These functions are available on the JAS web page and they are scheduled to be added to the R package WRS, which can be installed with the R command install.packages("WRS", repos="http://R-Forge.R-project.org") assuming that the most recent version of R has been installed.

REFERENCES

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

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

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 D.

Hoaglin, F. Mosteller & J. Tukey (Eds.) *Exploring Data Tables Trends and Shapes.* New York: Wiley, pp. 461–515.

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika, 75*, 800–802.

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 data depth and bootstrap. *Journal of the American Statistical Association, 92*, 266–277.

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

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

Price, R. M. & Bonett, D. G. (2001). Estimating the variance of the median. *Journal of Statistical Computation and Simulation, 68*, 295–305.

Radloff L. (1977). The CES-D scale: a self report depression scale for research in the general population. *Applied Psychological Measurement, 1*, 385-401.

Sfakianakis, M. E. & Verginis, D. G. (2006). A new family of nonparametric quantile estimators *Communications in StatisticsSimulation and Computation, 37*, 337345.

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

Wilcox, R. R. (2012). *Introduction to Robust Estimation and Hypothesis Testing* (3rd Edition). San Diego, CA: Academic Press.