Journal of Data Science **9**(2011), 1-13

Comparing Two Dependent Groups: Dealing with Missing Values

Rand R. Wilcox
University of Southern California

Abstract: The paper considers the problem of comparing measures of location associated with two dependent groups when values are missing at random, with an emphasis on robust measures of location. It is known that simply imputing missing values can be unsatisfactory when testing hypotheses about means, so the goal here is to compare several alternative strategies that use all of the available data. Included are results on comparing means and a 20% trimmed mean. Yet another method is based on the usual median but differs from the other methods in a manner that is made obvious. (It is somewhat related to the formulation of the Wilcoxon-Mann-Whitney test for independent groups.) The strategies are compared in terms of Type I error probabilities and power.

Key words: Bootstrap methods, medians, trimmed means.

1. Introduction

When comparing dependent groups, a commonly encountered concern is missing values. As is fairly evident, simply excluding missing data, known as complete case analysis, might result in inefficient estimation, which in turn might result in a substantial reduction in power when testing hypotheses (e.g., Liang, Wang, Robins, and Carroll, 2004). There is a massive literature on methods for handling missing observations, much of which is summarized in several books (e.g., Allison, 2001; Little and Rubin, 2002; McKnight, McKnight, Sidani and Figueredo, 2007; Molenberghs and Kenward, 2007; Daniels and Hogan, 2008; Schafer, 1997). When dealing with means, common approaches include maximum likelihood (Ibrahim, Chen and Lipsitz, 1999; Ibrahim, Lipsitz and Horton, 2004), weighted adjustment (Cochran, 1977), single imputation (Rao and Sitter, 1995) and multiple imputation (Rubin, 1987; Little and Rubin, 2002). Ludbrook (2008) suggests using a permutation test when comparing groups based on means. A positive feature of a permutation test is that when testing the hypothesis of identical distributions, the exact probability of a Type I error can be determined. But as a method for comparing means or medians, it can be unsatisfactory even when there are no missing values (e.g., Boik, 1987; Romano, 1990).

When the goal is to compare the marginal means and data are imputed, a simple strategy is to then compute a confidence interval using a normal or t approximation in the usual manner. It is known, however, that this approach can be unsatisfactory, as noted for example by Liang, Su and Zou (2008) as well as Wang and Rao (2002). Moreover, even when there are no missing values, concerns about relatively low power arise when sampling from a heavy-tailed distribution (e.g., Wilcox, 2005). And when the goal is to compare robust measures of location, it appears that no imputation strategy has been proposed and studied.

Liang *et al.* (2008) suggested an approach to missing values using an empirical likelihood method, based on single imputation, which is readily adapted to the problem of computing a confidence interval for μ_D the population mean associated with $D = X - Y$, where X and Y are the dependent random variables being compared. A concern, however, is that when dealing with heavy-tailed distributions, large sample sizes might be needed to get a reasonably accurate confidence interval for μ_D even when there are no missing values. For example, suppose D has the contaminated normal distribution

$$H(x) = (1 - \epsilon)\Phi(x) + \epsilon\Phi(x/10), \tag{1.1}$$

where $\Phi(x)$ is the standard normal distribution. Based on simulations with 5000 replications, if the sample size is $n = 100$, $\epsilon = .1$, and the goal is to compute a .95 confidence interval for the mean, the actual probability coverage is estimated to be .084. Using the Bartlett-correction studied by DiCiccio, Hall and Romano (1991), the actual probability coverage is .078. Here it was found that even with a few missing values, the probability coverage deteriorates, and so this approach was abandoned. Moreover, when sampling from a skewed heavy-tailed distribution, practical concerns are exacerbated. And even if accurate probability coverage could be attained, concerns about relatively low power, when using any method based on means, remains for reasons summarized, for example, in Wilcox (2005).

As is well known, heavy-tailed distributions, roughly meaning distributions for which outliers are likely to occur, appear to be quite common based on modern outlier detection techniques, as predicted by Tukey (1960). Moreover, many new and improved methods have been developed in an attempt to deal with this problem, which include methods based on robust measures of location. There are many results on comparing robust measures of location (e.g. Wilcox, 2005), but when comparing robust measures of location associated with two dependent groups, it appears that there are no results on how to handle missing values beyond the complete case strategy.

Here, two related goals are of interest. The first is to test

$$H_0 : \theta_1 = \theta_2, \tag{1.2}$$

where θ_j is some measure of location associated with the jth marginal distribution ($j = 1$, 2). The second goal is to compare the groups by testing the hypothesis that the median of the distribution of $D = X - Y$, say θ_D, is zero, where X and Y are two random variables that are possibly dependent having some unknown bivariate distribution. Of course, from basic principles, $E(D) = E(X) - E(Y)$. However, although there are exceptions, under general conditions the median of D is not equal to the median of X minus the median of Y. The motivation for considering inferences about the distribution of D is that missing values can be addressed in a relatively simple fashion, as will become evident.

When comparing measures of location associated with the marginal distributions, the focus in this paper is on comparing means and 20% trimmed means, assuming that data are missing at random in the sense defined by Rubin (1976). That is, the reason why a data point is missing is not related to its actual value. Although methods based on means have known practical concerns even when no values are missing, results on comparing means are included anyway as a benchmark.

One reason for focusing on 20% trimmed means, versus other amounts of trimming, stems from efficiency considerations (Rosenberger and Gasko, 1983). The median represents the maximum amount of trimming, it can have relatively high efficiency when sampling from a very heavy-tailed distribution where a large proportion of outliers are expected, but its efficiency under normality or other relatively light-tailed distributions can be unsatisfactory. One strategy for dealing with this issue is to trim less. A 20% trimmed mean has nearly the same efficiency as the mean under normality but it can have substantially higher efficiency when sampling from a heavy-tailed distribution. It is not being suggested that 20% trimming is always optimal, it is not, but it is a reasonably good choice for general use. (Another approach is to use a one-step M-estimator based on Huber's Ψ, but this is not pursued here. In terms of maximizing power when dealing with data from actual studies, there is some evidence that a 20% trimmed mean is usually preferable; see Wu, 2002.)

The above discussion might seem to suggest that when making inferences about the distribution associated with D, a 20% trimmed mean should be used rather than $\hat{\theta}_D$, an estimate of the median of D . But now, in terms of efficiency, the median generally performs better than a 20% trimmed mean (Wilcox, 2006). Even under normality, efficiency compares well with the usual sample mean. For example, under bivariate normality with Pearson's correlation $\rho = 0$, the squared standard error of the mean, divided by the squared standard error of $\hat{\theta}_D$, is .94

with a sample size of $n = 10$. With $\rho = .5$ and .8, this ratio is .85 and .84, respectively. With $n = 50$, these ratios, again for $\rho = 0$, .5 and .8, are .96, .91 and .87, respectively. And it is already known that hypotheses about θ_D can be tested using a standard percentile bootstrap method (Wilcox, 2006). That is, in simulations covering a wide range of distributions, reasonably accurate probability coverage is obtained when there are no missing values. The method is easily extended to the case of missing values, but the impact of missing values is unknown.

2. Description of the Methods to be Compared

This section describes the details of the methods to be compared. All of the methods are based on relatively simple extensions of extant techniques and are based in part on bootstrap methods.

We begin with a brief review of the trimmed mean. Momentarily consider a single random sample: $X_1, \ldots, X_n$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the values written in ascending order and let $g = [\gamma n]$, $0 \leq \gamma < .5$, where $[\gamma n]$ is the greatest integer less than or equal to γn. Then the γ-trimmed mean is

$$\bar{X}_t = \frac{1}{n-2g} \sum_{i=g+1}^{n-g} X_{(i)}.$$

For reasons already explained, the focus is on $\gamma = .2$, the 20% trimmed mean.

Now consider the case of two dependent variables. It is assumed than n pairs of observations are randomly sampled where both values are available, which is denoted by $(X_1, Y_1), \ldots, (X_n, Y_n)$. The corresponding (marginal) γ-trimmed means are denoted by $\bar{X}_t$ and $\bar{Y}_t$. For the first marginal distribution, an additional n_1 observations are sampled for which the corresponding Y value is not observed. These observations are denoted by $X_{n+1}, \ldots, X_{n+n_1}$ and the trimmed mean of these n_1 observations is denoted by $\tilde{X}_t$. Similarly, n_2 observations are sampled for which the corresponding value for the first marginal distribution is not observed and the trimmed mean is denoted by $\tilde{Y}_t$.

2.1 Method 1

The first method for testing (1.2) stems from a simple variation and generalization of the approach used by Lin and Stivers (1974) to derive a (non-bootstrap) method for handling missing values when the goal is to compare the marginal means. Let $h_j = [\gamma n_j]$ $(j = 1, 2)$, and let $\lambda_j = h/(h + h_j)$. Then an estimate of the difference between the marginal trimmed means, $\mu_{tD} = \mu_{t1} - \mu_{t2}$, is

$$\hat{\mu}_{tD} = \lambda_1 \bar{X}_{t1} - \lambda_2 \bar{X}_{t2} + (1 - \lambda_1)\tilde{X}_{t1} - (1 - \lambda_2)\tilde{X}_{t2},$$

a linear combination of three independent random variables.

An expression for the squared standard error of $\hat{\mu}_{tD}$ follows almost immediately from results summarized in Wilcox (2005). To briefly review, the γ-Winsorized variance associated with X is

$$\sigma_{wx}^2 = \int_{x_\gamma}^{x_{1-\gamma}} (x - \mu_w)^2 dF(x) + \gamma[(x_\gamma - \mu_w)^2 + (x_{1-\gamma} - \mu_w)^2],$$

where x_γ is the γ quantile. Using the notion of Winsorized expected values (e.g., Wilcox, p. 39), or the influence function of the trimmed mean, the squared standard of $\lambda_1 \bar{X} - \lambda_2 \bar{Y}$ is

$$\sigma_0^2 = \frac{1}{(1-2\gamma)^2 n}(\lambda_1^2 \sigma_{wx}^2 + \lambda_2 \sigma_{wy}^2 - 2\lambda_1\lambda_2\sigma_{wxy}), \tag{2.1}$$

where σ_{wxy} is the population Winsorized covariance between X and Y. The squared standard error of $(1-\lambda_1)\tilde{X}$ is

$$\sigma_1^2 = \frac{(1-\lambda_1)^2 \sigma_{wx}^2}{(1-2\gamma)^2(n+n_1)} \tag{2.2}$$

and the squared standard error of $(1-\lambda_2)\tilde{Y}$ is

$$\sigma_2^2 = \frac{(1-\lambda_2)^2 \sigma_{wy}^2}{(1-2\gamma)^2(n+n_2)}. \tag{2.3}$$

So the squared standard error of $\hat{\mu}_{tD}$ is

$$\tau^2 = \sigma_0^2 + \sigma_1^2 + \sigma_2^2.$$

For convenience, let $N_1 = n + n_1$ and $g_1 = [\gamma N_1]$. The Winsorized values corresponding to $X_1, \ldots, X_{N_1}$ are

$$W_{xi} = \begin{cases} X_{(g_1+1)} & \text{if } X_i \le X_{(g_1+1)} \\ X_i & \text{if } X_{(g_1+1)} < X_i < X_{(N_1-g_1)} \\ X_{(N_1-g_1)} & \text{if } X_i \ge X_{(N_1-g_1)}. \end{cases}$$

The (sample) Winsorized mean is

$$\bar{W}_x = \frac{1}{N_1}\sum_{i=1}^{N_1} W_{xi},$$

an estimate of the Winsorized variance, σ_{wx}^2, is

$$s_{wx}^2 = \frac{1}{N_1 - 1}\sum (W_{xi} - \bar{X}_x)^2,$$

and an estimate of σ^2_{wy} is obtained in a similar fashion. The Winsorized covariance between X and Y is estimated with

$$s_{wxy} = \frac{1}{n-1}\sum_{i-1}^{n}(W_{xi} - \tilde{W}_x)(W_{yi} - \tilde{W}_y),$$

where

$$\tilde{X}_w = \frac{1}{n}\sum_{i=1}^{n} W_{xi}$$

and $\tilde{Y}_w$ is defined in a similar manner. (Perhaps there is some practical advantage to using $\bar{W}_x$ rather than $\tilde{W}_x$ when computing the sample Winsorized covariance, but this has not been considered.)

The sample Winsorized variances yield estimates of σ^2_0, σ^2_1 and σ^2_2 via equations (2.1), (2.2) and (2.3), say $\hat{\sigma}^2_0$, $\hat{\sigma}^2_1$ and $\hat{\sigma}^2_2$, in which case an estimate of the squared standard error of $\hat{\mu}_{tD}$ is

$$\hat{\tau}^2 = \hat{\sigma}^2_0 + \hat{\sigma}^2_1 + \hat{\sigma}^2_2.$$

So a reasonable test statistic for testing (1.2) is

$$T = \frac{\hat{\mu}_{tD}}{\hat{\tau}}. \tag{2.4}$$

There remains the problem of approximating the null distribution of T and here a basic bootstrap-t method is used (e.g., Efron and Tibshirani, 1993). Based on results derived by Hall (1988a, 1988b), a symmetric two-sided confidence interval is used as opposed to an equal-tailed confidence interval. The method begins by randomly sampling with replacement $N = n + n_1 + n_2$ pairs of observations from $(X_1, Y_2), \ldots, (X_N, Y_N)$ yielding $(X^*_1, Y^*_2), \ldots, (X^*_N, Y^*_N)$. Based on this bootstrap sample, estimate μ_{tD} and τ, label the results $\hat{\mu}^*_{tD}$ and $\hat{\tau}^*$, respectively, and let

$$T^* = \frac{|\hat{\mu}^*_{tD} - \hat{\mu}_{tD}|}{\hat{\tau}^*}.$$

Repeat this process B times and put the resulting T^* values in ascending order yielding $T^*_{(1)} \le \cdots \le T^*_{(B)}$. Then an approximate $1-\alpha$ confidence interval for μ_{tD} is

$$(\bar{X}_{t1} - \bar{X}_{t2}) \pm T^*_{(c)}\hat{\tau}$$

where $c = (1-\alpha)B$ rounded to the nearest integer.

As for the choice for B the initial strategy was to use $B = 100$ with the goal of reducing execution time, but this was found to be unsatisfactory in terms of controlling the probability of a Type I error. Increasing B to 300, good control over the Type I error probability was achieved for a range of situations, but exceptions occur, as will be seen. For these latter situations, simulations were repeated with $B = 600$. A criticism of not using a larger value for B is that this might lead to some loss of power. Racine and MacKinnon (2007) discuss this issue at length and proposed a method for reducing this problem. (Also see Jöckel, 1986). Davidson and MacKinnon (2000) proposed a pretest procedure for choosing B.

2.2 Method 2

It is known that when comparing population means, a bootstrap t method is generally preferable to a percentile bootstrap method (e.g., Hall and Wilson, 1991; Wilcox, 2005). However, for various robust location estimators, it is known that the reverse is true (Wilcox, 2005). This suggests using a percentile bootstrap method when comparing the marginal 20% trimmed means, which deals with missing values (at random) in a straightforward and simple manner.

As with method 1, generate a bootstrap sample and let $\tilde{D}_t^* = \tilde{X}_t^* - \tilde{Y}_t^*$, where $\tilde{X}_t^*$ is the trimmed mean based on all of the X_i^* values not missing and $\tilde{Y}_t^*$ is computed in a similar manner. Repeat this B times, put the resulting yielding $\tilde{D}_t^*$ values in ascending order, and label the results $D_{t(1)}^* \leq \cdots \leq D_{t(B)}^*$. Then an approximate $1 - \alpha$ confidence interval for μ_{tD} is

$$(\tilde{D}_{t(\ell+1)}^*, \tilde{D}_{t(u)}^*),$$

where $\ell = \alpha B/2$, rounded to the nearest integer, and $u = B - \ell$. To compute a p-value, estimate $p = P(\hat{\mu}_{tD}^* > 0)$ with $\hat{p}$, the proportion of $\tilde{D}_t^*$ values greater than 0. Then a (generalized) p-value is

$$P = 2\text{min}(\hat{p}, 1 - \hat{p})$$

(Liu and Singh, 1997).

2.3 Method 3

Now focus on θ_D, the median of the distribution of $D = X - Y$. The goal is to test H_0: $\theta_D = 0$.

The method begins by forming all pairwise differences among the observed X and Y values. That is, compute $D_{ij} = X_i - Y_j$, $i = 1, \ldots, N_1$; $j = 1, \ldots, N_2$ resulting in $N_1 \times N_2$ D_{ij} values. Then an estimate of θ_D is obtained by computing the sample median of the D_{ij} values.

The idea of making inferences about θ_D is not new and in fact is a fundamental component of well-known methods for comparing independent groups. More precisely, consider the Wilcoxon-Mann-Whitney test or any of its modern extensions (e.g., Brunner, Domhof and Langer, 2002; Cliff, 1996; Wilcox, 2005). Let $p = P(X < Y)$ be the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second group. As is well known, these methods provide a test of the hypothesis

$$H_0 : p = .5$$

and can be viewed as a method for testing the hypothesis that the distribution of D has a median of 0. (But under general conditions it does not test the hypothesis that the median of X is equal to the median of Y.) This follows almost immediately from developments in Cliff (1996) and this perspective is discussed at some length in Wilcox (2005).

Again, a basic percentile bootstrap method is used to make inferences about θ_D. Generate a bootstrap sample as done in Method 2 and let $\hat{\theta}^*_D$ be the resulting estimate of θ_D. Repeat this process B times yielding $\hat{\theta}^*_{Db}$, $b = 1, \ldots, B$. Next, put these B values in ascending order yielding $\hat{\theta}^*_{D(1)} \leq \cdots \leq \hat{\theta}_{D(B)}$ and let ℓ and u be defined as before. Then a $1 - \alpha$ confidence interval for θ_D is

$$(\hat{\theta}^*_{D(\ell+1)}, \hat{\theta}^*_{D(u)}).$$

3. Simulation Results

Simulations were used to check the small-sample properties of the methods described in the previous section. Observations were generated from a bivariate distribution having marginal g-and-h distributions with correlation $\rho = 0$ or .5.

To elaborate, let Z be a standard normal random variable. 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 (Hoaglin, 1985). The standard normal distribution corresponds to $g = h = 0$. The case $g = 0$ corresponds to a symmetric distribution, and as g increases, skewness increases as well. The parameter h determines heavy tailedness. As h increases, heavy tailedness increases. The six marginal distributions considered here are $(g, h) = (0, 0)$, (0, .5), (.5, 0), (5., .5), (1, 0), and (1, 0). So bivariate normal distributions are included and correspond to $g = h = 0$. Table 1 summarizes the skewness (κ_1) and kurtosis (κ_2) for the g-and-h distributions used in the simulations. When $h > 1/k$, $E(X - \mu)^k$ is not defined and the corresponding entry in Table 1 is left blank.

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

g	h	κ_1	κ_2
0.0	0.0	0.00	3.00
0.0	0.5	0.00	—
0.5	0.0	1.75	8.9
0.5	0.5	—	—
1.0	0.0	6.19	113.94
1.0	0.5	—	—

There remains the issue of generating data having a specified correlation. 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. Let V be an $n \times 2$ matrix of data where the independent marginal distributions have one of the g-and-h distributions previously described. Then the matrix product XU produces an $n \times 2$ matrix of data that has population correlation matrix R.

The sample size used here was $N = 30$ with $n_1 = n_2 = 5$ as well as $(n_1, n_2) = (10, 0)$. Table 2 reports the estimated type I error probabilities based on 1000 replications and $B = 300$, where the number of replications was chosen to keep execution time at a reasonable level while simultaneously providing a reasonably accurate estimate of the actual Type I error probability. If, for example, the actual level is .05, then the standard error of the estimated level is $\sqrt{.05(.95)/1000} = .0069$.

Table 2. Estimated type I error probabilities, $\alpha = .05$, $n = 30$, $n_1 = n_2 = 5$

g	h	ρ	M1 (μ)	M1 (μ_t)	M2	M3
0.0	0.0	0.0	.032	.039	.049	.056
0.5	0.0	0.5	.037	.037	.062	.063
0.0	0.5	0.0	.024	.023	.051	.065
0.0	0.5	0.5	.026	.023	.055	.059
0.5	0.5	0.0	.019	.021	.059	.065
0.5	0.5	0.5	.019	.018	.047	.064
1.0	0.0	0.0	.022	.026	.057	.065
1.0	0.0	0.5	.006	.010	.038	.064
1.0	0.5	0.0	.009	.014	.051	.065
1.0	0.5	0.5	.008	.009	.034	.057

Tables 2 and 3 report the estimated Type I error probabilities, where M1 (μ) indicates method 1 using means, M1 (μ_t) is method M1 using 20% trimmed means, and M2 and M3 are methods 2 and 3, respectively. Note that both versions of method M1 avoid estimated Type I error probabilities above the nominal level, but both yield estimates less than .025 for a variety of situations involving a heavy-tailed distribution. This is of particular concern when using means because there are at least two reasons why comparing means can result in relatively low power: the actual level of the test can be substantially less than the nominal level and the standard error of the mean can be relatively high when sampling from a heavy-tailed distribution. Both methods M2 and M3 are more satisfactory, with M2 usually giving the best results. With very few exceptions, the estimated Type I error probability using M2 is closer to the nominal level compared to using M3.

Table 3. Estimated type I error probabilities, $\alpha = .05$, $n = 30$, $n_1 = 10$, $n_2 = 0$

g	h	ρ	M1 (μ)	M1 (μ_t)	M2	M3
0.0	0.0	0.0	.051	.043	.055	.058
0.0	0.0	0.5	.046	.046	.053	.050
0.0	0.5	0.0	.029	.032	.050	.057
0.0	0.5	0.5	.025	.030	.046	.054
0.5	0.0	0.0	.044	.037	.058	.058
0.5	0.0	0.5	.044	.037	.048	.048
0.5	0.5	0.0	.025	.025	.052	.058
0.5	0.5	0.5	.021	.025	.043	.047
1.0	0.0	0.0	.035	.029	.057	.058
1.0	0.0	0.5	.014	.012	.034	.048
1.0	0.5	0.0	.016	.022	.051	.058
1.0	0.5	0.5	.010	.007	.027	.049

For the heavy-tailed distributions, where and M1 (μ) and M1 (μ_t) have estimated Type I error probabilities well below the nominal level, the simulations were repeated with $B = 600$, but very similar results were obtained.

A few additional simulations were run where the two marginal distributions differ in shape. As expected, based on results summarized in Wilcox (2005), the method for comparing means can be unsatisfactory in terms of Type I errors and probability coverage when the marginal distributions differ in skewness, but the percentile bootstrap method with a 20% trimmed mean performed reasonably well. For example, if the first distribution is standard normal and the second is lognormal, shifted so that the measures of locations being compared are equal, M1 (μ) has an estimated Type I error probability of .082 with $n_1 = n_2 = 5$ and $\rho = 0$. For M1 (μ_t) it is .030, and for M2 and M3 the estimates are .065 and .074. Very similar results are obtained when $\rho = .5$. Again M2 is best with only a slight decrease in control over the probability of a Type I error.

Altering the marginal variances can make matters worse when comparing means. For example, if the standard normal is replaced by a normal distribution with standard deviation .25, even with no missing data, the estimated Type I error is .108 using M1 (μ) and only .062 using M3.

3.1 Comments on power

At some level, power comparisons are meaningless because three different measures of location are being used. Under general conditions, for example, the difference between the marginal means is not equal to the difference between the marginal 20% trimmed means. So situations can be constructed where M1 (μ) can have more power than M1 (μ_t) and the reverse is true as well. And an added complication is that the levels of the various methods can differ. Nevertheless, some comments about power under a common shift in location might help provide a useful perspective.

First consider bivariate normality where the marginal distributions have population means 0 and .5. With $\rho = 0$, $N = 30$, and $n_1 = n_2 = 5$, power for methods M1 (μ), M1 (μ_t), M2 and M3 was estimated to be .346, .274, .389 and .414, respectively. For $\rho = .5$ the estimates were .560, .439, .512, and .561. So although the mean has optimal efficiency, the level of method M1 (μ) is less than the Type I error probability associated with methods M2 and M3 making it possible for methods M2 and M3 to have power approximately the same or a bit higher than M1 (μ).

Now consider the situation where the marginal distributions are g-and-h distributions with $g = 0$ and $h = .5$. Shifting the second marginal distribution by .8, power (with $\rho = 0$) for the four methods is now .148, .307, .488 and .520. Now the sample mean has the worst efficiency, which combined with an actual level of only .024 when testing at the .05 level, results in relatively poor power.

4. Concluding Remarks

In summary, both versions of method M1 were found to be less satisfactory relative to methods M2 and M3. However, if there is a specific interest in comparing the marginal 20% trimmed means, rather than making inferences about the 20% trimmed mean of $X - Y$, M1 appears to avoid Type I error probabilities greater than the nominal level. But when using M1 with the goal of comparing the marginal means, the Type I error probability can be well above the nominal level. All indications are that M2 is a bit more satisfactory than M3 in terms of Type I errors, with both methods performing reasonably well among the situations considered. In terms of power, M3 is a bit more satisfactory than M2. But as stressed, M2 and M3 are based on different measures of location, so when dealing with data from actual studies, M2 might have higher power in some situations.

References

Allis, P. D. (2001). *Missing Data.* Sage.

Boik, R. J. (1987). The Fisher-Pitman permutation test: A non-robust alternative to the normal theory F test when variances are heterogeneous. *British Journal of Mathematical and Statistical Psychology* **40**, 26-42.

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

Cochran, W. G. (1977). *Sampling Techniques* (3rd Ed.). Wiley.

Daniels, M. J. and Hogan, J. W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis.* Chapman and Hall/CRC.

Davidson, R. and MacKinnon, J. G. (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews* **19**, 55-68.

Domhof, S., Brunner, E. and Osgood, D. (2002). Rank procedures for repeated measures with missing values. *Sociological Methods and Research* **30**, 367-393.

Efron, B. and Tibshirani, R. J. (1993) *An Introduction to the Bootstrap.* Chapman and Hall.

Hall, P. (1988a). On symmetric bootstrap confidence intervals. *Journal of the Royal Statistical Society, Series B* **50**, 35-45.

Hall, P. (1988b). Theoretical comparison of bootstrap confidence intervals. *Annals of Statistics*, **16**, 927-953.

Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics* **,** **47**, 757-762.

Hoaglin, D. C. (1985) Summarizing shape numerically: The g-and-h distributions. In *Exploring data tables, trends, and shapes.* (Editec by D. Hoaglin, F. Mosteller and J. Tukey, 461-515). Wiley.

Ibrahim, J. G., Chen, M. H. and Lipsitz, S. R. (1999). Monte Carlo EM for missing covariates in parametric regression models. *Biometrics* **55**, 591-596.

Ibrahim, J. G., Lipsitz, S. R. and Horton, N. (2001). Using auxiliary data for parameter estimation with nonignorable missing outcomes. *Applied Statistics* **50**, 361-373.

Jöckel, K.-H. (1986). Finite sample properties and asymptotic efficiency of Monte Carlo tests. *Annals of Statistics* **14**, 336-347.

Keselman, H. J., Lix, L. M. and Kowalchuk, R. K. (1998). Multiple comparison procedures for trimmed means. *Psychological Methods*, **3**, 123-141.

Liang, H., Su, H. and Zou, G. (2008). Confidence intervals for a common mean with missing data with applications in an AIDS study. *Computational Statistics and Data Analysis* **53**, 546-553.

Little, R. J. A. and Rubin, D. (2002). *Statistical Analysis with Missing Data*, 2nd Ed. Wiley.

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

Ludbrook, J. (2008). Outlying observations and missing values: how should they be handled? *Clinical and Experimental Pharmacology and Physiology* **35**, 670-678.

McKnight, P. E., McKnight, K. M., Sidani, S. and Figueredo, A. J. (2007). *Missing Data: A Gentle Introduction.* Guilford Press.

Racine, J. and MacKinnon, J. G. (2007). Simulation-based tests than can use any number of simulations. *Communications in Statistics-Simulation and Computation* **36**, 357-365.

Rogan, J. C., Keselman, H. J. and Mendoza, J. L. (1979). Analysis of repeated measurements. *British Journal of Mathematical and Statistical Psychology*, **32**, 269-286.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association* **85**, 686-692.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581-592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley.

Staudte, R. G. and Sheather, S. J. (1990). *Robust Estimation and Testing.* Wiley.

Tukey, J. W. (1960). A survey of sampling from contaminated normal distributions. In *Contributions to Probability and Statistics* (Edited by I. Olkin *et al.*). Stanford University Press.

Wang, Q. H. and Rao, J. N. K. (2002). Empirical likelihood-based inference in linear models with missing data. *Scandanavian Journal of Statistics*, **29**, 563-576.

Westfall, P. H. and Young, S. S. (1993). *Resampling Based Multiple Testing* Wiley.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Ed. Academic Press.

Wilcox, R. R. (2006). A note on inferences about the median of difference scores. *Educational and Psychological Measurement* **66**, 624-630.

Wu, P.-C. (2002). Central limit theorem and comparing means, trimmed means one-step M-estimators and modified one-step M-estimators under non-normality. Unpublished doctoral disseration, Dept. of Education, University of Southern California.

Received July 20, 2009; accepted August 26, 2009.

Rand R. Wilcox
Dept of Psychology
University of Southern California
Los Angeles, CA 90089, USA
rwilcox@usc.edu