Journal of Modern Applied Statistical Methods

Volume 1 | Issue 2 Article 37

11-1-2002

Within Groups Multiple Comparisons Based On Robust Measures Of Location

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

H. J. Keselman
University of Manitoba, kesel@ms.umanitoba.ca

Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 281-287

1538 – 9472/02/$30.00

Within Groups Multiple Comparisons Based On Robust Measures Of Location

Rand R. Wilcox
Dept of Psychology
University of Southern California

H. J. Keselman
Dept of Psychology
University of Manitoba

Consider the problem of performing all pair-wise comparisons among J dependent groups based on measures of location associated with the marginal distributions. It is well known that the standard error of the sample mean can be large relative to other estimators when outliers are common. Two general strategies for addressing this problem are to trim a fixed proportion of observations or empirically check for outliers and remove (or down-weight) any that are found. However, simply applying conventional methods for means to the data that remain results in using the wrong standard error. Methods that address this problem have been proposed, but among the situations considered in published studies, no method has been found that gives good control over the probability of a Type I error when sample sizes are small (less than or equal to thirty); the actual probability of a Type I error can drop well below the nominal level. The paper suggests using a slight generalization of a percentile bootstrap method to address this problem.

Key words: M-estimators, trimming, bootstrap.

Introduction

Outliers (unusually small or large values) can inflate the standard error of the sample mean which in turn can result in relatively poor power, and outliers can distort the sample mean resulting in a misleading representation of the typical response (e.g., Rosenberger & Gasko, 1983; Staudte & Sheather, 1990; Wilcox, 2001). When dealing with measures of location, two general strategies have been proposed for dealing with this problem.

Rand R. Wilcox is a Professor of Psychology, a fellow of the Royal Statistical Society and the American Psychological Society, has published over 170 journal articles, and has recently written his fifth book on statistics. E-mail him at rwilcox@usc.edu. H. J. Keselman is a Professor of Psychology, a fellow of the American Psychological Association and the American Psychological Society, and has published over 100 journal articles and book chapters related to the analysis of repeated measurements, multiple comparison procedures, and robust estimation and testing. E-mail him at kesel@ms.umanitoba.ca.

The first is to simply trim a fixed proportion of the extreme values. In terms of maintaining a relatively low standard error under normality yet deal with situations where outliers are rather common, a 20% trimmed mean is often recommended (which is formally defined in the next section of this paper). The other strategy is to empirically check for outliers and remove (or downweight) any that are found. Various textbooks recommend some variation of the latter strategy and often refer to this as data cleaning.

If outliers are removed and the values are not erroneous (merely unusually large or small), applying standard methods for means to the remaining data results in using the wrong standard error, which in turn means poor control over the probability of a Type I error and inaccurate confidence intervals. Effective methods for dealing with this problem were derived for a range of situations, but when comparing measures of location associated with the marginal distributions of dependent groups, practical problems remain. Methods that avoid Type I error probabilities well above the nominal level are available, but when empirically checking and discarding outliers, the actual probability of a Type I error can drop well below the nominal level.

For J dependent groups, let θ_j be some measure of location associated with the jth marginal distribution. More formally, this paper is concerned with all pairwise comparisons where for every j < k, the goal is to test

$$H_0 : \theta_j = \theta_k. \qquad (1)$$

Of particular interest is controlling the family-wise error rate (FWE), meaning the probability of at least one Type I error. When the sample size is small and the goal is to have FWE equal to .05, extant simulation results indicate that it is possible to ensure FWE will not exceed .05 by a substantial amount using 20% trimmed means in conjunction with a generalization of the bootstrap method (Wilcox, 1997b). A concern, however, is that the actual FWE can drop well below the nominal level suggesting that the method might have relatively low power.

Wilcox (1997b) also found that when using an estimator that in effect discards outliers (called a one-step M-estimator with Huber's Ψ), poor control over FWE is obtained with sample sizes less than or equal to thirty. Currently, no method has been found that performs reasonably well in simulations when using this particular M-estimator and the sample size is small. So a practical issue remains: Is it possible to find a method that, in simulations, not only avoids FWE rates larger than the nominal level, it ensures that FWE will not be substantially below the nominal level when extreme values are discarded. This paper describes such a method which is based on a slight generalization of the percentile bootstrap.

Description of the Robust Estimators

The focus is on three measures of location. The first is a 20% trimmed mean. Generally, trimmed means simply remove a fixed proportion of the extreme observations. By fixed proportion is meant that the amount of trimming is not determined empirically by, for example, checking to see what proportion of the observations are outliers. The median and mean are trimmed means that represent the two extremes of the maximum amount and least amount of trimming, respectively. The choice of 20% trimming provides reasonably good efficiency under normality and it maintains relatively high efficiency in situations where the sample mean performs poorly (Rosenberger & Gasko, 1983; Wilcox, 1997a), so we focus on it here. The 20% trimmed mean removes the smallest 20% of the observations, as well as the largest 20%, and averages the values that remain. If $X_1,\ldots,X_n$ is a random sample, let $X_{(1)} \leq \ldots \leq X_{(n)}$ be the observations written in ascending order and let g be equal to .2n rounded down to the nearest integer. Then a 20% trimmed mean is

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

However, 20% trimmed means in particular, and trimmed means in general, suffer from at least two practical concerns. First, the amount of trimming is assumed to be fixed in advance. If the amount of trimming is set at 20%, efficiency is reasonably good versus the mean under normality, but when sampling from a sufficiently heavy-tailed distribution, efficiency can be poor versus using more trimming or switching to some robust M-estimator of location. A second general concern is that typically trimmed means assume symmetric trimming. That is, the same proportion of observations are trimmed from both tails of an empirical distribution. When sampling from an approximately symmetric distribution, symmetric trimming seems reasonable, but asymmetric trimming might be more appropriate as the degree of skewness increases. Well known theoretical results indicate how to estimate the standard error of a trimmed mean when asymmetric trimming is used (e.g., Huber, 1981), but now unsatisfactory probability coverage can result when sample sizes are small (e.g., Wilcox, 1997a). Also, if the amount of trimming is empirically determined, and the standard error is estimated by conditioning on this amount of trimming, even poorer control over probability coverage can result.

The second measure of location is a particular robust M-estimator. Generally, robust M estimators are more flexible than trimmed means in the sense that they empirically determine whether a value is unusually large or small and then such values are down weighted in some manner. The particular M-estimator of interest here is the one-step M-estimator based on Huber's Ψ:

$$\frac{1.28(MADN)(i_2 - i_1) + \sum_{i=i_1+1}^{n-i_2} X_{(i)}}{n - i_1 - i_2}, \quad (2)$$

where M is the usual median, MAD is the median of the values X_1-M ,…, X_n-M , MADN=MAD/.6745, i_1 is the number of observations X_i such that (X_i - M) < -K(MADN), i_2 is the number of observations X_i such that (X_i - M) > K(MADN), and K is some constant usually chosen to achieve good properties under normality. (See, for example, Staudte and Sheather, 1990.) This estimator empirically determines whether an observation is an outlier, trims it, averages the values that remain, but with asymmetric trimming an adjustment is made based on a measure of scale, MAD. The adjustment based on MAD is a consequence of how the population value of the one-step M-estimator is defined. It is the value θ satisfying

$$E\left[\Psi\left(\frac{X-\theta}{MADN}\right)\right] = 0, \quad (3)$$

where Ψ(x) = max[-K; min(K; x)]. Equation (3) can be solved with the Newton-Raphson method and a single iteration of this technique yields (with K = 1.28) equation (2). The choice K = 1.28 provides good efficiency under normality and its finite sample breakdown point is .5, the highest possible value. (The finite sample breakdown point of an estimator is the smallest proportion of observations, which when altered, can drive the value of an estimator to plus or minus infinity.) However, when performing all pair-wise comparisons among J dependent groups based on this one-step M-estimator, none of the techniques examined by Wilcox (1997b) performed well in simulations. Moreover, situations arise where even the most successful method can have Type I error probabilities well below the nominal level.

The third measure of location considered here is a so-called modified one-step M-estimator (MOM). The MOM estimator belongs to the class of skipped estimators originally proposed by Tukey and studied by Andrews, Bickel, Hampel, Huber, Rogers and Tukey (1972). The idea is simple: Check for outliers, discard any that are found, and then average the values that remain. The class of skipped estimators studied by Andrews et al. is based on a boxplot outlier detection rule which has a finite sample breakdown point of only .25. Here an outlier detection rule based on M and MADN is used instead resulting in a location estimator having a finite sample breakdown point of .5 as well. (Huber, 1993, argues that at a minimum, an estimator should have a finite sample breakdown point of at least .1.)

An apparent disadvantage of skipped estimators is that expressions for their standard errors are very complicated when sampling from an asymmetric distribution. One of the main points in this paper is that a variation of the percentile bootstrap method not only circumvents this problem, it provides good probability coverage in simulations where no effective method based on a robust M-estimator has been found.

The modified one-step M-estimator begins by declaring X_i an outlier if

$$\frac{.6745|X_i - M|}{MAD} > K,$$

where K is adjusted so that efficiency is good under normality. (Outlier detection rules based on the sample mean and variance are known to be unsatisfactory, e.g., Wilcox, 2001, pp. 34-35.) Then MOM is given by

$$\hat{\theta} = \sum_{i=i_1+1}^{n-i_2} \frac{X_{(i)}}{n - i_1 - i_2}, \quad (4)$$

where now i_1 (i_2) is the number of observations less (greater) than the median that are declared outliers. Here, K = 2.24 is used which is approximately equal to the square root of the .975 quantile of a chi-square distribution with one degree of freedom. This particular outlier detection rule is a special case of a general method suggested by Rousseeuw and van Zomeren (1990.) It is noted that this choice for *K* yields good efficiency under normality.

In particular, using simulations with 10,000 replications, we found that with *K* = 2.24, the standard error of the sample mean divided by

the standard error of $\hat{\theta}$ is approximately .9 for n = 20(5)100. For n = 10 and 15, this ratio is .88.

The Proposed Method for Pair-wise Comparisons

Here, $\hat{\theta}_j$ represents the estimate of the measure of location associated with jth marginal distribution. Let X_{ij}, $i = 1, \ldots, n$, $j = 1,\ldots, J$ represent a random sample of size n from some J-variate distribution. So for fixed j and when using a trimmed mean, $\hat{\theta}_j$ would be the 20% trimmed mean associated with $X_{1j},\ldots, X_{nj}$, ignoring the other data.

First consider a basic percentile bootstrap method for testing (1) which stems from Liu and Singh (1997) as well as Hall (1986) and is applied as follows. Obtain bootstrap samples by resampling with replacement n rows from the n by J matrix of X_{ij} values. Repeat this process B times and let $\hat{\theta}^*_{bj}$ be the bootstrap estimate of θ_j based on the bth bootstrap sample, $b = 1,\ldots, B$; $j = 1,\ldots, J$. (Here, θ_j represents the population value of any of the three estimators under consideration.) Let

$$p^*_{jk} = P(\hat{\theta}^*_j > \hat{\theta}^*_k)$$

based on a random bootstrap sample. Here this probability is estimated with $\hat{p}^*_{jk}$, the proportion of bootstrap samples having $\theta^*_{bj} > \theta^*_{bk}$. Then if H_0 is true, $\hat{p}^*_{jk}$ has, asymptotically, a uniform distribution, so reject if $\min(\hat{p}^*_{jk}, 1-\hat{p}^*_{jk}) \le \alpha/2$.

To control FWE, some type of sequentially rejective method can be used. Here consideration was given to the approach derived by Rom (1990) as well as Hochberg (1988) which are outlined below. A positive feature of the methods just outlined is that for all three measures of location, simulation estimates of the FWE were less than or equal to the nominal level for all of the situations described in our simulations. This is true when using the Rom or the Hochberg method. However, a negative feature when testing at the .05 level was that when using MOM or Huber's M-estimator, the estimated FWE was typically less than .05 by an unacceptable amount. In fact, estimates dropped below .01, particularly when the correlations among the variables are high.

An examination of the simulation results indicated why this problem arose. When $\hat{\theta}_j = \hat{\theta}_k$, it should be the case that $\hat{p}^*_{jk} = .5$. Near equality was found when the correlation between X_{ij} and X_{ik} is close to zero, but as the correlation increased, the difference between $E(\hat{p}^*_{jk})$ and .5 increased as well.

This observation suggests the following modification. Set

$$D_{ij} = X_{ij} - \hat{\theta}_j .$$

That is, shift the data so that the null hypothesis is true. Obtain a bootstrap sample of size n from the D_{ij} values and let $\hat{\theta}^*_{cj}$ be the resulting estimate of θ_j. Repeat this process B times and let $\hat{p}^*_{cjk}$ be the proportion of times $\hat{\theta}^*_{cj}$ is greater than $\hat{\theta}^*_{ck}$. Set

$$\hat{p}^*_{ajk} = \hat{p}^*_{jk} - \lambda(\hat{p}^*_{cjk} - .5),$$

where λ is a constant to be determined. Then for fixed j and k, reject $H_0 : \theta_j = \theta_k$ if $\hat{p}^*_{ajk}$ is sufficiently large or small.

For convenience, set

$$\hat{p}^*_{mjk} = \min(\hat{p}^*_{ajk}, 1-\hat{p}^*_{ajk})$$

and assume the goal is to have FWE equal to α. One approach to controlling FWE is to proceed along the lines in Hochberg (1988). Writing the $C = (J^2 - J)/2$ values $\hat{p}_{mjk}$ as $p_{m1},\ldots,p_{mC}$, put these C values in ascending order yielding $\hat{p}_{m(1)} \le \ldots \le \hat{p}_{m(C)}$. For any i = C, C-1, … , 1, if $\hat{p}_{m(i)} \le \alpha / 2(C - i + 1)$, reject the corresponding hypothesis as well as all hypotheses having smaller $\hat{p}_{m(i)}$ values.

Rom's (1990) method is applied in the same manner as Hochberg's technique, only $\alpha / 2(C - i + 1)$ is replaced by a value tabled by Rom. Situations were found where Rom's method was a bit less satisfactory in avoiding FWE above the nominal level, so it is not considered further. Yet another approach was derived by Benjamini

and Hochberg (2000), but it is known that this method does not control FWE, so it is not considered here.

There remains the problem of choosing λ. The strategy was to determine an appropriate value under normality with all correlations equal to zero and all marginal distributions having a common variance. The reason for considering all correlations equal to zero was that when using a trimmed mean, MOM, or an M-estimator with Huber's Ψ, this was found to maximize the probability of at least one Type I error among all the situations considered in the next section. For n = 11 and 20, it was found that λ = .1 gave good results when using MOM or the M-estimator considered here when used in conjunction with Hochberg's method, and as n increases, the term $\lambda(\hat{p}^*_{cjk} - .5)$ becomes negligible. Using λ = 0 results in FWE typically being less than the nominal level, but often it was far below the nominal level. As for 20% trimmed means, λ = 0 performed well (no correction is needed) when using Hochberg.

Results

The small-sample properties of the methods just described were studied for J = 4 with simulations where observations were generated from a multivariate normal distribution via the IMSL (1987) subroutine RNMVN. Nonnormal distributions were generated using the g-and-h distribution (Hoaglin, 1985). That is, first generate Z_{ij} from a multivariate normal distribution and set

$$X_{ij} = \frac{\exp(gZ_{ij}) - 1}{g} \exp(hZ_{ij}^2/2).$$

For g = 0 this last expression is taken to be

$$X_{ij} = Z_{ij} \exp(hZ_{ij}^2/2).$$

The case g = h = 0 corresponds to a normal distribution. Setting g = 0 yields a symmetric distribution, and as g increases, skewness increases as well. Heavy-tailedness increases with h. The values for g and h were taken to be (g, h) = (0, 0), (0, .5), (.5, 0) and (.5, .5). Table 1 contains skewness (κ_1) and kurtosis (κ_2) values for the four g-and-h distributions used in the simulations.

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

g	h	κ_1	κ_2	$\hat{\kappa}_1$	$\hat{\kappa}_2$
0.0	0.0	0.00	3.00	0.00	3.0
0.0	0.5	0.00	—	0.00	11,896.2
0.5	0.0	1.75	8.9	1.81	9.7
0.5	0.5	—	—	120.10	18,393.6

When h > 1/k, $E(X-\mu)^k$ k is not defined and the corresponding entry in Table 1 is left blank. A possible criticism of simulations performed on a computer is that observations are generated from a finite interval, so the moments are finite even when in theory they are not, in which case observations are not being generated from a distribution having the theoretical skewness and kurtosis values listed in Table 1. In fact, as h gets large, there is an increasing difference between the theoretical and actual values for skewness and kurtosis. Accordingly, Table 1 also lists the estimated skewness $(\hat{\kappa}_1)$ and kurtosis $(\hat{\kappa}_2)$ values based on 100,000 observations generated from the distribution. Simulations were also run where the marginal distributions were lognormal or exponential.

Simulations were run where the marginal distributions had equal and unequal variances. When working with skewed distributions, the marginal distributions were first shifted so that they have a θ value of zero, and for the unequal variance case the ith observation in the jth group was multiplied by σ_j , $(\sigma_1, \sigma_2, \sigma_3, \sigma_4) = (1, 3, 4, 5)$. That is, for skewed distributions, before multiplying the X_{ij} by σ_j, the observations were shifted by subtracting the population value of θ so that when multiplying by σ_j, the null hypothesis remains true.

Five patterns of correlations were used. Four of the five correlation matrices have a common correlation, ρ, with ρ = 0, .1, .5 and .8. The fifth correlation matrix had $\rho_{12} = .8$, $\rho_{13} = .5$, $\rho_{14} = .2$, $\rho_{23} = .5$, $\rho_{24} = .2$ and $\rho_{34} = .2$. The largest and smallest estimates of FWE consistently occurred with the first and latter two correlation matrices, so for brevity, only the results for the first and fifth matrices are reported. These two correlation matrices are labeled C1 and C2, respectively.

Table 2 contains the estimated probability of at least one Type I error when using the multiple comparison procedure described in the previous section. The results are based on 2,000 replications. As is evident, reasonably good control over the probability of a Type I error is achieved. The main difficulty is that when using MOM, there are two instances where the estimate drops below .02.

Table 2: Estimated Type I error probabilities for g-and-h distributions, $n = 11$, $J = 4$

g	h	Correlation	σ	Method MOM	M-EST	$\bar{X}_t$
0.0	0.0	C1	(1,1,1,1)	.030	.020	.036
		C2		.067	.053	.051
0.0	0.5	C1	(1,1,1,1)	.040	.037	.044
		C2		.014	.015	.027
0.5	0.0	C1	(1,1,1,1)	.056	.059	.048
		C2		.025	.035	.033
0.5	0.5	C1	(1,1,1,1)	.041	.044	.042
		C2		.016	.023	.026
0.0	0.0	C1	(1,3,4,5)	.053	.051	.036
		C2		.057	.049	.034
0.0	0.5	C1	(1,3,4,5)	.041	.044	.038
		C2		.037	.048	.028
0.5	0.0	C1	(1,3,4,5)	.050	.058	.042
		C2		.056	.046	.034
0.5	0.5	C1	(1,3,4,5)	.041	.051	.041
		C2		.041	.048	.033

Conclusion

The main point is that currently, no method for comparing robust measures of location associated with the marginal distributions is very satisfactory in simulations with small sample sizes. The results reported here illustrate that by using a slight generalization of the percentile bootstrap method, good control over the probability of a Type I error can be achieved in a wide range of situations when outliers are removed.

As for trimmed means, a basic (unmodified) percentile bootstrap method performs well. The three estimators used in Table 2 are designed to have reasonably good efficiency under normality, they have high efficiency when sampling from a heavy-tailed distribution where the sample mean performs poorly, so comparing groups as described would seem to have practical value. The M-estimator and modified M-estimator seem particularly attractive, and now it appears that a viable method for performing all pair-wise comparisons, based on the measures of location associated with the marginal distributions, is available when sample sizes are small.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances*. Princeton University Press, Princeton, NJ.

Benjamini, Y. & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics*, *25*, 60-83.

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

Hall, P. (1986). On the bootstrap and confidence intervals. *Annals of Statistics*, *14*, 1431-1452.

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

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

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Huber, P. (1993). Projection pursuit and robustness. In S. Morgenthaler, E. Ronchetti & W. Stahel (Eds.) *New directions in statistical data analysis and robustness*. Boston: Birkh¨auser Verlag.

IMSL (1987). *Library I*, *vol. II*. Houston: International Mathematical and Statistical Libraries.

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

Rom, D. M. (1990). A sequentially rejective test procedure based on a modified Bonferroni inequality. *Biometrika*, *77*, 663-666.

Rosenberger, J. L., & Gasko, M. (1983). In D. C. Hoaglin, F. Mosteller and J. W. Tukey (Eds.) *Understanding robust and exploratory data analysis*. New York: Wiley.

Rousseeuw, P. J., & van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points (with discussion). *Journal of the American Statistical Association*, *85*, 633-639.

Staudte, R. G., & Sheather, S. J. (1990). *Robust estimation and testing*. New York: Wiley.

Wilcox, R. R. (1997a). *Introduction to robust estimation and hypothesis testing*. San Diego, CA: Academic Press.

Wilcox, R. R. (1997b). Pairwise comparisons using trimmed means or M-estimators when working with dependent groups. *Biometrical Journal*, *39*, 677-688.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. New York: Springer.