1-1-2002

Altematives To S_w In The Bracketed Interval Of The Trimmed Mean

Jennifer Bunner
Wunderman Analytics, jeni_bunner@det.wunderman.com

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 176-181

1538 - 9472/02/$30.00

Graduate Student Research

Alternatives To S_w In The Bracketed Interval Of The Trimmed Mean

Jennifer Bunner
Wunderman Analytics, Detroit

Shlomo Sawilowsky
Educational Evaluation and Research
College of Education
Wayne State University

The aim of this Monte Carlo study is to examine alternatives to estimated variability in building bracketed intervals about the trimmed mean.

Keywords: Trimmed mean, Bracketed interval, Winsorized sample standard deviation, Robust tests

Introduction

The prevalence of nonnormally distributed data in applied studies has been documented (e.g., Micceri, 1989; Pearson & Please, 1975, Tan, 1982). Summary statistics, such as measures of central tendency, and parametric hypothesis tests, such as Student's t, are affected by nonnormal data, as many studies have also documented (e.g., Bradley, 1968, 1978; Blair, 1981; Blair & Higgins, 1980a, 1980b, 1985).

Nonnormality arises for a variety of reasons. In some cases, the underlying distribution of the variable is exponential (e.g., growth or decay), multimodal lumpy (e.g., Micceri, 1989), mass at zero with gap (e.g., Sawilowsky & Hillman, 1992), or some other non-Gaussian shape. In other cases, an essentially normal model can be adopted if perturbations, commonly called outliers, can be assumed to have contaminated the model. The latter case motivated the development of robust statistics.

Consider, for example, measures of central tendency for a single sample. The arithmetic mean, $\bar{x}$, is the most commonly used measure of the average. It is a sample statistic that is used as a point estimate of the population parameter μ. However, it has a finite sample breakdown point of only 1/n. This implies that even a single observation can vastly distort the obtained value of $\bar{x}$, and hence, it is not a robust measure.

In contradistinction, the median is much more robust. Its finite sample breakdown point is approximately ½. Thus, almost half of the values could be untoward perturbations, and yet the value of the median remains unaffected. Despite this robustness property, the median never emerged as a popular measure of central tendency. Three possible reasons can be offered as an explanation for this unpopularity: (1) the sampling distribution of the median is intractable (requiring reliance on asymptotic variances or some other approach), making the construction of hypothesis tests difficult, (2) the sample median is usually not a very good estimate of the population median, and (3) the value of the median is actually determined based on only one number for N = odd (e.g., the point on the scale below which half of the observations fall), or within the upper and lower real limits of a single value for N = even, essentially ignoring the information contained in all of the other scores.

A well known alternative to dealing with nonnormally distributed data, where an essentially Gaussian structure can be assumed to exist underlying the data, is the trimmed mean ($\bar{x}_t$). The trimmed mean is a compromise between the mean (i.e., trim = zero) and the median (i.e., trim approximately equal to but less than 50%).

The 2×10% trimmed mean means that 10% of the observations are deleted from both sides of the data set. As an illustration, the 2×10% trim is calculated on the data below by (1) ordering the data from low to high, (2) deleting the .1 × 10 = 1 observation on the left and the .1 × 10 = 1 observation on the right, and (3) computing the mean on the remaining 8 scores. This is illustrated in Table 1.

A question that naturally arises in working with $\bar{x}_t$ is how to form a bracketed interval around it. In other words, how well does $\bar{x}_t$ estimate the population mean μ? For example, consider a 95% bracketed interval. From a frequentist's perspective, the purpose is to determine if one can be 95% sure that μ is contained within the interval built around the sample trimmed mean. A Bayesian's perspective would view this differently, and determine if many such intervals were formed, would 95% of those intervals contain the population mean.

Jennifer Bunner is a doctoral student in Educational Evaluation and Research, Wayne State University. She is a marketing analyst with Wunderman Analytics, 550 Towne Center Drive, Suite 300, Dearborn, MI, 48126, e-mail: jeni_bunner@det.wunderman.com. Shlomo S. Sawilowsky, Wayne Statue University Distinguished Faculty Fellow, is Professor and Chair, Educational Evaluation and Research, College of Education, Wayne State University. His interests are in nonparametric, robust, exact, permutation, and other computer-intensive methods, especially via Fortran.

Table 1. Computing Trimmed Means.

Original	85	92	87	93	99	86	88	90	73	91
Ordered	73	85	86	87	88	90	91	92	93	99
Trimmed		85	86	87	88	90	91	92	93	

$$\bar{X}_t = \frac{85+86+87+88+90+91+92+93}{8} = 89$$

Many modern textbooks (e.g, Wilcox, 1996) address this question and give a formula similar to the following:

$$C.I._{1-\alpha}(\mu_t) = \bar{X}_t \pm \left(t_{1-\alpha} \times \frac{1}{1-2\gamma} \times \frac{s_w}{\sqrt{n}} \right) \qquad 1.$$

where γ is the amount of trimming and s_w is the sample winsorized standard deviation.

Assume $\alpha = 0.05$ and the amount to trim g = .1. The right side of (1) contains four expressions. The first term, $\bar{x}_t$, is the sample trimmed mean. With regard to the second expression, $t_{1-\alpha}$, Student's t is two tailed, and degrees of freedom after trimming is $\nu = n - 2g - 1$, where g is the percent to trim on one side. In the example above, n = 10, g = .1 × 10 = 1, and thus, $\nu = 10 - (2 \times 1) - 1 = 7$. Therefore, $t_{1-\alpha} = t_{.975} = 2.365$.

The third expression is a multiplier that is used to adjust the standard error (which is the fourth term) based on the amount of trimming. If there has been no trimming, this term reduces to 1, leaving the full expression of the standard error. As the amount of trimming increases, the denominator decreases, and this multiplier increases.

The final expression, the standard error, is in fact the focus of the current paper. The s_w term is a robust estimate of the population variance, which is unbiased after being divided by the square root of the sample size. The sample winsorzed standard deviation is obtained by "winsorizing the data", which is accomplished by recoding extreme values closer to the median.

For the current data, a 2 × 10% winsorization is accomplished by recoding the two most extreme values back (i.e, the 73 is recoded to an 85, and the 99 is recoded to a 93). Winsorization is a method of treating outliers without taking the harsh measure of deleting extreme values, but rather, recoding outliers to values that are toward the ends of the distribution but are more likely to be valid than perturbations. The value of S_w for the example data is calculated as follows in Table 2. (See bottom of page.)

The standard deviation of the winsorized values is 3.2. For comparison, the standard deviation of the original scores is 6.8.

An examination of the three right-most expressions that constitute the bracketed interval of the trimmed mean indicates that Formula (1), although widely circulated, certainly has no rigorous mathematical basis of support. There does not appear to be any justification for using the cdf of the t distribution, unless an underlying Gaussian data structure is strictly assumed. Moreover, modifications to ν (e.g., Satterithwaite) are just as likely to ensure the sampling distribution of $\bar{x}_t$ is Student's t as is the use of the multiplier in the third term. However, for the purposes of this paper, attention is turned to the use of s_w.

Wilcox (1996) and other textbooks that rely on some form of Formula (1) cite Tukey and McLaughlin (1963), which is the primary source for support of s_w. This paper is highly recommended to graduate students because it reads more like a fireside chat than a technical statistical paper. In this paper, Tukey and McLaughlin search for a robust measure of dispersion for the numerator of the fourth expression in Formula (1), recognizing that use of the sample standard deviation, which has the nonrobust arithmetic mean as its statistical engine, would be self-defeating in the presence of outliers.

The primary condition they sought to satisfy is that the average value of the denominator squared and the variance of the numerator are matched, or "in constant proportion over as broad a spectrum" (p. 337) of distributions as possible. Examination of the sample standard deviation of the trimmed mean based on this primary condition was shown to be unsatisfactory. Inspection of the results indicated that more consideration needed to be given to outliers than simply deleting them; hence, the winsorization approach was adopted.

However, there was no theoretical dependency requiring Tukey and McLaughlin's selection of the winsorized procedure as a robust measure of dispersion.

Table 2. Computing Winsorized Means.

Original	85	92	87	93	99	86	88	90	73	91
Ordered	73	85	86	87	88	90	91	92	93	99
Winsorized	85	85	86	87	88	90	91	92	93	93

Indeed, Lax (1985) identified over 150 different robust measures of dispersion, and the list is certainly longer than that. How might some other robust measure of variability perform in creating a bracketed interval for the trimmed mean?

Purpose of the Study

Given that the choice of S_w was based on trial and error, and no theoretical underpinning, the purpose of this study is to examine the properties of bracketed intervals formed by using some alternative measures of dispersion.

Methodology

Measures of Dispersion

Three common measures of variability are considered: Mean Deviation (S_{md}); Median Deviation (S_{mdd}); and MAD, the median absolute deviation (S_{mad}). (Note that only the S_{mad} is considered a robust measure, as the other two procedures eventually incorporate an arithmetic mean.) We also present results for a new measure of dispersion described below that is noted as S_{bs}. All four measures are compared with S_w.

S_{md}

The mean deviation is defined as

$$S_{md} = \frac{\sum|X-\mu|}{N-1} \qquad 2.$$

The mean is subtracted from each score, the absolute value is taken, the results are summed, and then divided by N. For example, consider the original scores above, where the mean is 88.4 and median is 89. The results are taken from Table 3. $\sum|X - \mu| = 46$, and thus, $S_{md} = 5.11$.

S_{mdd}

This statistic is similar to the mean deviation, except it is based on the absolute value of the average of the *median* subtracted from each score, instead of the mean. The formula is

$$S_{mdd} = \frac{\sum|X-\text{Median}|}{N-1} \qquad 3.$$

Because $\sum|X - \text{Median}| = 46$, $S_{mdd} = 5.11$.

Note that coincidentally, this value is the same for S_{md}. Also, because the final value is obtained via the arithmetic mean for both S_{md} and S_{mdd}, the resulting statistics suffer from the lack of robustness ascribed to the arithmetic mean.

S_{mad}

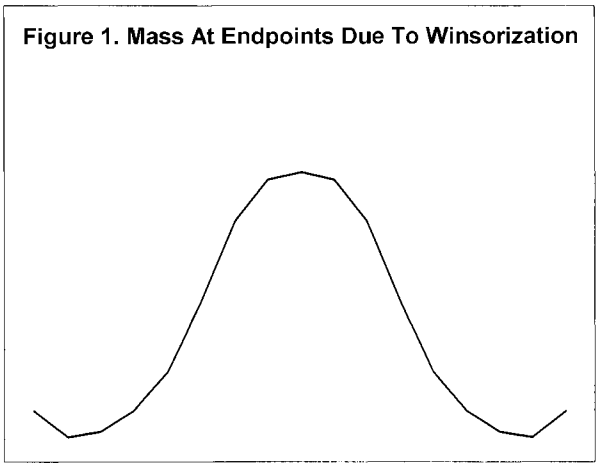
S_{mad} is similar to S_{mdd}, but with the important difference that instead of taking the *mean* of the absolute value of the deviations from the median, the *median* of the absolute value of the deviations from the median is taken, and thus, S_{mad} is a robust statistic. The median of the values in the 3rd column in Table 1 is 3. Thus, $S_{mad} = 3$.

S_{bs}

The idea behind the Bunner-Sawilowsky approach is to take into consideration the resulting histogram due to winsorizing, and attempt to smooth the end points. For example, if in a larger data set the winsorization method requires the recoding of the highest and lowest 10 values, then the endpoints of the distribution will have a mass at both recoding points, as noted in the Figure 1.

Table 3. Computing Recorded Scores.

Original Scores	$\|X-\mu\|$	$\|X-\text{Median}\|$	2×20% Bunner-Sawilowsky Recoded Scores
73	15.4	16	86
85	3.4	4	87
86	2.4	3	86
87	1.4	2	87
88	.4	1	88
90	1.6	1	90
91	2.6	2	91
92	3.6	3	92
93	4.6	4	91
99	10.6	10	92



Suppose the winsorization was 2×10%, meaning for the original data set both the lowest and highest score would be recoded back one score when N = 10. In this case, the S_{bs} procedure is identical to the winsorization. However, suppose that a 2×20% recoding was desired, where two scores were to be recoded on each end of the distribution. In the original data set, the winsorized procedure would recode the 73 and the 85 to 86s, yielding a mass of three 86s; and the 99 and 93 would be recoded to 92s, yielding a mass of three 92s.

The Bunner-Sawilowsky approach smooths this mass by recoding the 73 to 86, and the 85 to the next value score, which is an 87. Similarly, the 99 is recoded to a 92, but the 93 is recoded to a 91. The standard deviation of the recoded scores is 2.45.

To summarize, the values for the example in descending order are S = 6.8, S_{md} = 5.1, S_{mdd} = 5.1, S_w = 3.2, S_{mad} = 3, and S_{bs} = 2.45.

Bracketed Intervals

Two criteria were evaluated with regard to the performance of the various measures of dispersion being substituted into Formula (1) above. The first was the Type I error, where $\alpha = 0.05$. The second was the width of the resulting interval, which is simply the range (upper - lower).

Methodology

The study proceeded as follows: A *Minitab* Version 13.1 macro was written to randomly select variates from a standard unit Gaussian (i.e., de Moivreian) distribution N(0,1), with samples of sizes n=30. Next, four models of outliers were used. They were:

- one wild score on the left of the distribution (1WL)
- two wild scores on the left of the distribution (2WL)
- three wild scores on the left of the distribution (3WL)
- three wild scores on the left and one wild score on the right of the distribution (3WL-1WR)

The wild scores were created by taking the lowest score and subtracting 3.5; and where there were two wild scores to the left, also taking the second lowest score and subtracting 3.0; and where there were three wild scores to the left, also taking the third lowest score and subtracting 2.5; and where there were three wild scores to the left and one wild score to the right, also taking the highest score and adding 1.5. Because the μ and σ of the population are 0 and 1, respectively, this procedure takes the lowest score out an additional 3.5 standard deviations further from the mean, the second score is moved 3 standard deviations further from the mean, and so forth. The various measures of dispersion were computed, the resulting bracketed interval of the trimmed mean was calculated, the interval was checked to see if the population parameter was found within it, and the width of the interval was determined. Each experiment was repeated 1,000 times.

Results

The results are compiled in Table 4 below. Note that the common alternatives for measures of dispersion, the Mean Deviation, Median Deviation, and MAD resulted in Type I errors that were greatly inflated, typically from 0.05 to about 0.248, almost five times nominal alpha. Even the use of the robust MAD statistic performed poorly. Although the width of the resulting intervals are typically about 45% narrower than bracketed intervals formed with the winsorized standard deviation, these procedures will no longer be considered due to their lack of ability in preserving Type I errors to nominal alpha.

The dispersion measure based on the Bunner-Sawilowsky approach resulted in robust Type I errors according to Bradley's (1968) liberal criteria, where $.5\alpha \leq$ Type I error $\leq 1.5\alpha$, or 0.025 - .075. These results were not within Bradley's conservative criteria, however, which is $.9\alpha \leq$ Type I error $\leq 1.1\alpha$, or 0.045 - 0.055. The advantage of the Bunner-Sawilowsky approach, however, is that the resulting bracketed intervals are approximately 5.13% more narrow than the intervals formed by using the winsorized standard deviation.

Conclusion

The initial motivation for trying to improve on the bracketed interval of the trimmed mean was the consideration of no theoretical connection of the winsorized standard deviation to the trimmed mean. Furthermore, winsorization is a process that by definition creates a mass at the recoding points, which is at the extreme points of the distribution.

Table 4. Width and Type I Error For Bracketed Interval of The Trimmed Mean For Various Alternatives of S_w, Gaussian Distribution With Perturbations; 1,000 repetitions, $\alpha = 0.05$.

	1WL		2WL	
Statistic	Type I Error	Width	Type I Error	Width
S_w	.052	.766	.046	.760
S_{bs}	.064	.727	.061	.724
S_{mad}	.222*	.471	.226*	.474
S_{md}	.267*	.421	.263*	.420
S_{mdd}	.267*	.415	.267*	.415
	3WL		3WL - 1 WR	
S_w	.050	.768	.054	.765
S_{bs}	.063	.732	.064	.730
S_{mad}	.225*	.471	.206*	.474
S_{md}	.265*	.423	.239*	.439
S_{mdd}	.269*	.416	.244*	.434

Note: nWL = Number of wild observations in the left tail. nWR = Number of wild observations in the right tail. S_w = winsor. S_{bs} = Bunner-Sawilowsky. S_{mad} = Median Absolute Deviation. S_{md} = Mean Deviation. S_{mdd} = Median Deviation. * = Exceeds Bradley's (1968) liberal definition of robustness with respect to Type I error.

The new recoding scheme (S_{bs}) examined in this paper ameliorated the mass at the recoded end points by smoothing out the tails of the distribution. The scheme investigated is equivalent to the usual winsorization when the number of values to be recoded is one. However, when additional points are identified as outliers, they are recoded to the next values closer to the median. If the four lowest values are noted as x_1, x_2, x_3, and x_4, and two values are to be recoded, then the usual winsorization procedure would recode both x_1 and x_2 to x_3. However, the S_{bs} procedure would recode the value of x_1 to x_3 and x_2 to x_4. This, in effect, helps to reduce the mass at the recoding points.

Morever, the example data yielded the smallest estimate of variance for the S_{bs} as compared with all other competitors investigated. This indicates its resistence to the presence of outliers. This property directly translated into producing bracketed intervals with widths smaller than that achieved by using the winsorized standard deviation in the bracketed interval of the trimmed mean formula.

An inspection of the results indicated that the S_{bs} produced intervals that were more than 5% narrower than the usual winsorization. However, further study of this recoding scheme, and similar alternatives, is necessary because the Type I error rates were slightly inflated (e.g., $\approx .06$).

References

Blair, R. C. (1981). A reaction to 'Consequences Of Failure To Meet Assumptions Underlying The Fixed Effects Analysis of Variance and Covariance'. *Review of Educational Research, 51*, 499-507.

Blair, R. C., & Higgins, J. J. (1980a). A comparison of the power of the t test and Wilcoxon statistics when samples are draw from a certain mixed normal distribution. *Evaluation Review, 4*, 645-656.

Blair, R. C., & Higgins, J. J. (1980b). A comparison of the power of the Wilcoxon's rank-sum statistic to that of Student's t statistic under various nonnormal distributions. *Journal of Educational Statistics, 5*, 309-355.

Blair, R. C., & Higgins, J. J. (1985). A comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under population shapes. *Psychological Bulletin, 97*, 119-128.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice Hall.

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

Lax, D. A. (1985). Robust estimators of scale: Finite sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association, 80*, 736-741.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Minitab. (2000). MINITAB v. 13.1. State College, PA: Minitab, Inc.

Pearson, E. S., & Please, N. W. (1975). Relation between the shape of population distribution and the robustness of four simple test statistics. *Biometrika, 62*, 223-241.

Sawilowsky, S. S., & Hillman, S. B. (1992). Power of the independent samples t test under a prevalent psychometric distribution. *Journal of Consulting and Clinical Psychology, 60*, 240-243.

Tan, W. Y. (1982). Sampling distributions and robustness of t, F, and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communications in Statistics, 11*, 2485-2511.

Tukey, J. W., & McLaughlin, D. H. (1963). Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsorization 1. *Sankhya, 25*, 331-352.

Wilcox, R. R. (1996). *Statistics for the social sciences*. San Diego, CA: Academic Press.