The Utility of Robust Means in Statistics

Fara Goodwyn

Texas A&M University

Paper presented at the annual meeting of the Southwest Educational Research Association, New Orleans, February 1, 2012.

Abstract

Location estimates calculated from heuristic data were examined using traditional and robust statistical methods. The current paper demonstrates the impact outliers have on the sample mean and proposes robust methods to control for outliers in sample data. Traditional methods fail because they rely on the statistical assumptions of normality and homoscedasticity that are often not met with real data. Robust means are superior due to their ability to maintain power and control for Type I errors. Two robust location estimates, L-estimators (e.g., the trimmed mean and the Winsorized mean) and M-estimators, are reviewed.

Keywords: L-estimators, M-estimators, outliers, robust statistics

The Utility of Robust Means in Statistics

Statistical analysis has traditionally relied on the assumptions of normality and homoscedasticity (i.e., homogeneity of variance). These assumptions are rarely met with real data, because real data contain “weird” people (i.e., outliers; Thompson, 2006; Wilcox, 2010). When weird people cause even small violations of normality or homoscedasticity, it can greatly impact the rate of Type I and Type II errors (Erceg-Hurn & Mirosevich, 2008; Wilcox, 1998, 2010). Robust statistical methods (i.e., “the capacity of a statistic to be less influenced by outlying scores”; Thompson, 2006, p. 47) overcome the inherent weakness of traditional methods by maintaining power and controlling for Type I errors when statistical assumptions are not met in the data (Erceg-Hurn & Mirosevich, 2008; Rousseeuw, 1991). Although not an exhaustive list, modern methods have been developed for calculating robust measures of (a) location, (b) variance, (c) test statistics, and (d) effect sizes (Erceg-Hurn & Mirosevich, 2008; Wilcox, 2005, 2010). The current paper focuses on outliers in measures of location (e.g. mean, median), the need for robust measures of location, and discusses specific robust means including L- and M-estimators.

Outliers in Measures of Location

Measures of location or central tendency look for a single value to represent the typical individual in a data set (Wilcox, 2001). Well known measures of location or central tendency in research include the mean and the median. The sample mean is simply the mathematical average of sample data expressed as

$$M_X = \sum X_i / n \qquad (1)$$

As the formula demonstrates, each score, X_i, is summed prior to dividing by the number of scores, n, in the sample (Thompson, 2006).

The median is obtained by ordering the data by numerical value. When the number of scores is odd, the median is the middle score. When the number of scores is even, a mathematical average is taken of the two middle scores (Wilcox, 2001). Although beyond the scope of this paper, interested readers are directed to Thompson (2006) for the advanced statistical method to calculate the median:

$$\mathrm{M}dn_{\mathrm{x}} = \mathrm{P}_{50} = \mathrm{L} + [((q * n) - cum.f) / f] \quad (2)$$

Outliers are not necessarily bad. Thompson (2006) defined an outlier as “a participant whose data are distinctly atypical and thus would unduly influence characterizations of the data” (p. 43). Dixon (1950) described outliers as “dubious in the eyes of the researcher” (p. 488). These dubious outliers occur for many reasons. Data may be inaccurately recorded during the data collection process due to human or equipment error. Participants may intentionally misreport data, particularly with sensitive topics such as time spent studying, frequency of church attendance, and sexual history. Outliers also result from sampling error caused by pulling a sample from a population other than the population the sample is intended to generalize. Finally, outliers may occur in data due to the legitimate weirdness of individuals (Osborne & Overbay, 2004).

Heuristic data in Table 1 illustrate that an individual can be an outlier in one area but not another (e.g., John is an outlier for variable X1 but not for variables X2 and X3, Levi is an outlier for X2 but not for variables X1 and X3, George is an outlier for variable X3 but not for variables X1 and X2). Although identifying outliers in the heuristic data presented in Table 1 is easily achieved upon visual inspection alone, visual inspection is not a practical method for outlier detection in large data sets.

Table 1

Impact of Outliers on Mean and Median

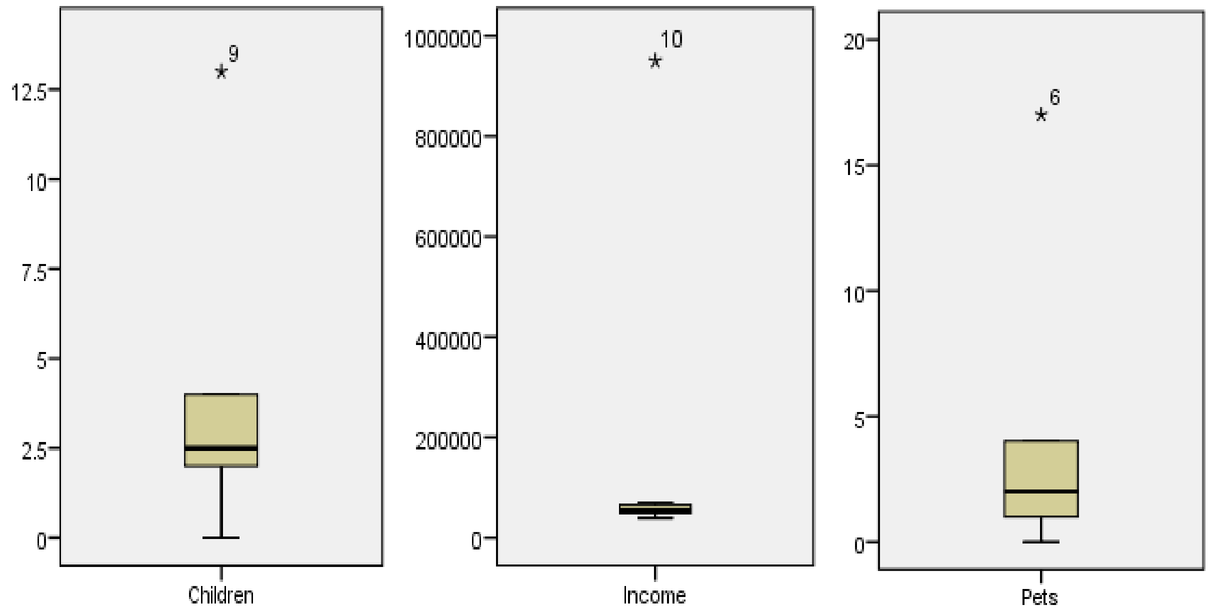
Score	Individual	Number of Children X1	Income X2	Number of Pets X3
1	Shane	2	$39,000	1
2	Ron	4	$45,000	2
3	Steven	2	$48,000	3
4	Mick	1	$50,000	2
5	Chris	3	$51,000	4
6	George	4	$59,000	17
7	Zach	3	$61,000	2
8	Paul	2	$65,000	0
9	John	13	$70,000	1
10	Levi	0	$950,000	4
	Mean	3.4	$143,800	3.6
	Median	2.00	$55,000	2.00

Figure 1 demonstrates one method for identifying outliers, the box plot (also known as the box-and-whiskers plot). The horizontal lines in the boxes represent each variables respective medians. The stars in the box plots identify outlying scores. There is one outlier for each variable. The outliers are (a) score nine, 13, for children, (b) score ten, $950,000, for income, and (c) score six, 17, for pets.

Computer intensive methods for identifying outliers are easily computed using statistical software such as SPSS® (IBM, 2010). For more information on outlier detection including the box plot, bootstrap, and jackknife procedures see Thompson (2006).

Sample means and medians are compared including and excluding outlying scores for heuristic purposes. The mean for income calculated without Levi’s $950,000 income is

Figure 1. *Box Plot Calculated with SPSS®*



$54,222.22. In contrast, the median for income calculated without Levi is $51,000. Levi's income alone alters the mean by a substantial amount (e.g., $143,800 - $54,222.22 = $89,577.78 or a 165.20% increase) and has a much smaller impact on the median (e.g., $55,000 - $51,000 = $4,000 or a 7.27% reduction). The heuristic data demonstrate that the median is a more robust measure of location than the mean.

Need for Robust Means in Statistics

One option for dealing with outliers is to omit them from analysis entirely. Rejecting outlying scores results in means that are much closer to their respective medians. Although it may produce desirable results, rejecting outliers without reason is not a robust method for calculating the mean.

If we don't reject or partially reject outliers, the mean fails to represent the typical respondent. The mean income of $143,800 is not a typical income earned by any of the individuals in Table 1. Regardless of whether outliers result from error or from the weirdness of individuals, outliers often cause the sample mean to differ substantially from the population mean. When Type I and Type II errors are not controlled, valuable research findings are often dismissed when the results would have held up using robust statistical methods (Wilcox, 1998). Statistical measures need to remain stable with or without outliers included and provide reliable information under various statistical assumptions. Researchers should not assume sample data is normally distributed or meets the homogeneity of variance assumption.

Robust methods "accommodate them (outliers) in a wider inferential scheme" (Barnett, 1978, p. 47). Location estimates derived by robust methods better represent the typical individual by minimizing the effects of extreme scores (Rousseeuw, 1991; Wilcox, 2001). Many strategies exist for minimizing the effects of outliers. Outliers may be eliminated from a data set by (a) excluding a predetermined percentage of scores, (b) substituting a predetermined percentage of scores for other values, (c) altering the weights of scores, or (d) using rejection tests. Each of these methods controls for the effects of outliers without bias.

Robust Means

L-estimators. The trimmed mean and Winsorized mean are the two most common L-estimators used in robust statistics (Wilcox, 2005). L-estimators require ordering data prior to calculating a weighted mean (Wilcox, 2003). When calculating a weighted mean, the weights typically add to 1 (Wilcox, 2001). Table 2 illustrates calculating a sample mean and a weighted mean using variable X2 from Table 1.

Weighted Means. The sample mean is typically calculated by adding individual scores together and dividing by the total number of values (e.g., 1,438,000 divided by 10) as expressed by Equation 1. An alternative method for sample mean calculation is to multiply each score by a weight. The numerator of the weight is always one. The denominator equals the number of values in the data set (e.g., 1/10). Because there are 10 values in Table 2, each score is multiplied by a weight of 1/10 or .10 which results in the sample mean of $143,000.

Table 2

Mean Calculation with Weights

Individual	Income X2	Weights for Sample Mean	Weights for Weighted Mean
Shane	$39,000	.10	.20
Ron	$45,000	.10	.10
Steven	$48,000	.10	.10
Mick	$50,000	.10	.10
Chris	$51,000	.10	.10
George	$59,000	.10	.10
Zach	$61,000	.10	.10
Paul	$65,000	.10	.10
John	$70,000	.10	.05
Levi	$950,000	.10	.05
Mean		$143,800	$96,700

The weights for weighted mean calculation in Table 2 were selected for heuristic purposes only. The weighted mean resulted in a mean smaller than the original sample mean due to putting less emphasis on extreme values (i.e., .05 weight versus .10 weight on the outlying $950,000 value).

Trimmed Means. Trimmed means are calculated by removing a predetermined percentage of scores from both tails of a distribution prior to evaluating the data (Wilcox, 1997). Table 3 contains a 20% trimmed mean for the heuristic data originally presented in Table 1 on the variable X2. The top and bottom 20% of scores were given a weight of zero. The remaining values in the heuristic data set have a weight of 1/6 applied to calculate the trimmed mean. The denominator of the weight, six, was determined by the number of remaining values after trimming 20% of scores from each tail of the sampling distribution.

Table 3

Comparison of Mean to Trimmed and Winsorized Means for X2, Income

Individual	Original Values Income X2	Trimmed Scores X2	Winsorized Scores X2
Chris	$39,000	---	$48,000
George	$45,000	---	$48,000
John	$48,000	$48,000	$48,000
Levi	$50,000	$50,000	$50,000
Mick	$51,000	$51,000	$51,000
Paul	$59,000	$59,000	$59,000
Ron	$61,000	$61,000	$61,000
Shane	$65,000	$65,000	$65,000
Steven	$70,000	---	$65,000
Zach	$950,000	---	$65,000
Mean	$143,800	$55,667[a]	$56,000[b]
Median	$55,000	$55,000	$55,000

[a] The mean trimmed 20% on each side of the distribution.
[b] 20% Winsorized mean

The median and mean can be described in reference to trimming. The median is a trimmed mean, trimmed at 50%. At the opposite extreme, the mean is a location measure with 0% of the scores trimmed prior to calculation (Wilcox, 2001). A 20% trimmed sample mean is recommended by Wilcox (2003b) as a compromise between the 0% trimming of the mean and the 50% trimming of the median. As previously discussed, the original mean of $143,800 is not representative of any of the individuals in the table. The trimmed mean, $55,667, is much more successful at deriving a single score that represents the typical income of the individuals in this data set. The trimmed mean lies between the original sample mean of $143,800 and the sample median of $55,000.

A sample trimmed mean is not intended to estimate the population mean for the entire distribution. Instead, the sample trimmed mean estimates the population trimmed mean. When normal distribution assumptions are not met in the population, the population trimmed mean often better represents the typical individual than the population mean (Wilcox, 2003b). When normal distribution assumptions are met, the trimmed population mean and population mean have identical values (Erceg-Hurn & Mirosevich, 2008).

***Winsorized Means*.** Table 3 contains a 20% Winsorized distribution. A Winsorized distribution substitutes a predetermined percent of scores at both tails of the distribution for less extreme values (Wilcox, 2003a). Emphasis is put on the values in or near the center of the distribution of sample scores (Wilcox, 2005). The less extreme values that replace original data are the smallest and largest numbers not trimmed when computing the trimmed mean (Wilcox, 2003a). The Winsorized mean is the mathematical average (i.e., Equation 1) of the Winsorized distribution. The heuristic data resulted in a Winsorized mean similar to the trimmed mean. Like the trimmed mean, the Winsorized mean is between the values of the original sample mean and median.

M-estimators. M-estimators have more flexibility in the way outliers are controlled than L-estimators. Like the trimmed mean, M-estimators rely on trimming as the robust statistical procedure to control for or reject outliers in sample data. Unlike the trimmed mean, empirical methods, as opposed to predetermined percentages, are used to determine how much trimming occurs (Wilcox, 2005). This results in the M-estimator being better equipped to deal with large numbers of outliers since the amount trimmed is not bound to a predetermined percentage (i.e., an M-estimator could trim 30% of scores or more). M-estimators are also more flexible than L-estimators in that they do not require the amount of trimming on each tail of the distribution to be symmetrical. In fact, M-estimators may trim from one tail and not the other (Wilcox, 2005).

The downside to M-estimators is that they are not easily calculated by hand or easily demonstrated with heuristic data. Fortunately, M-estimators are easily calculated with software created by Wilcox (1998).

Discussion

When robust means are calculated, Type I errors are controlled and power is maintained. Controlling for Type I and Type II errors allows us to preserve important research findings. Outliers should not be omitted without reason. Outliers must be minimized without bias. The effects of outliers are minimized by excluding a predetermined percentage of scores when calculating a trimmed mean. The Winsorized mean substitutes a predetermined percentage of scores for other values. Both the trimmed mean and the Winsorized mean use altered weights to calculate robust means. Finally, empirical methods are used to reject outlying scores when calculating an M-estimator.

Robust statistical procedures produce sample means that are more representative of the typical respondent in a data set. Researchers should report statistics with robust measures. Due to the ease of calculation provided by statistical software programs, there is no reason not to report robust means in research.

References

Barnett, V., & Lewis, T. (1978). *Outliers in statistical data.* New York: John Wiley & Sons.

Dixon, W. J. (1950). Analysis of extreme values. *The Annals of Mathematical Statistics, 21,* 488-506.

Erceg-Hurn, D. M., & Mirosevich, V. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63,* 591-601.

IBM. (2010). IBM SPSS Statistics 19. Somers, NY: IBM Corporation.

Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research and Evaluation, 9*(6). Retrieved from http://PAREonline.net/getvn.asp?v=9&n=6.

Rousseeuw, P. (1991). Tutorial to robust statistics. *Journal of Chemometrics, 5,* 1-20.

Thompson, B. (2006). *Foundations of behavioral statistics: An insight-based approach.* New York: The Guilford Press.

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

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300-314.

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods.* New York: Springer.

Wilcox, R. R. (2003a). *Applying contemporary statistical techniques.* San Diego, CA: Academic Press.

Wilcox, R. R. (2003b). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8,* 254-274.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). San Diego, CA: Elsevier Academic Press.

Wilcox, R. R. (2010). *Fundamentals of modern statistical method* . New York: Springer.