Journal of Modern Applied Statistical Methods

Volume 6 | Issue 2 Article 2

11-1-2007

Optimal Trimming and Outlier Elimination

Philip H. Ramsey
Queens College of CUNY, Flushing, Philip.Ramsey@qc.cuny.edu

Patricia P. Ramsey
Fordham University

Journal of Modern Applied Statistical Methods
November, 2007, Vol. 6, No. 2, 355-360

1538 – 9472/07/$95.00

Invited Articles

Optimal Trimming and Outlier Elimination

Philip H. Ramsey
Queens College of CUNY, Flushing

Patricia P. Ramsey
Fordham University

Five data sets with known true values are used to determine the optimal number of pairs that should be trimmed in order to produce the minimum relative error. The optimal trimming in the five data sets is found to be 1%, 5%, 7%, 10% and 28%. The 28% rate is shown to be an outlier among the five data sets. Results of four data sets are used to establish cutoff values for outlier detection in two robust methods of outlier detection.

Key words: Median absolute deviation, Box-and-whisker plot, MAD statistic.

Introduction

Outliers have been considered a serious problem for the application of many statistical procedures, especially when assuming an underlying normal distribution. Barnett and Lewis (1978) provided a detailed treatment of outliers and a number of procedures for outlier detection. Barnett and Lewis state, "We shall define an outlier in a set of data to be an observation (or set of observations) which appears to be inconsistent with the remainder of that set of data" (p. 4). Similar definitions have been provided by others (Everitt, 2002; Marriott, 1990).

The presence of outliers has been shown to seriously bias traditional statistical procedures (Wilcox, 2001). Symmetric trimming of a data set by removing a specified percentage of data points from each tail of a distribution is a simple method of removing outliers. A 10% trim would remove the top and bottom 10% of the data. In general, 100α% trimming of a sample of size N would remove $[100\alpha N]$ from the top and bottom of the N ordered observations where [ ] implies the greatest lower integer.

Trimming the data biases the standard deviation of a data set but that problem can be overcome (Wilcox, 2001). However, the number

Philip H. Ramsey is Professor of Psychology. E-mail him at Philip.Ramsey@qc.cuny.edu. This research was supported in part by a PSC-CUNY grant. Patricia P. Ramsey is Professor of Management Systems. Email her at ramseyphd@fordham.edu.

of pairs trimmed (i.e. the value of α) must be determined. Wilcox as argued for $\alpha = .20$. Some researchers may find eliminating 40% of the data to be excessive. Some others may even resist any trimming unless outlier detection can be objectively confirmed. Trimming has been found to be beneficial in testing differences in means (Kowalchuk, Keselman, Wilcox, & Algina, 2006; Lix & Keselman, 1998).

Methodology

One of the simplest methods for evaluating an observation as a possible outlier would be to divide the deviation from the mean by the standard deviation. The problem is that an outlier biases the standard deviation upward thus reducing the ratio and making the observation appear less extreme. This "masking" effect is particularly strong when more than one outlier is present (Barnett & Lewis, 1978; Wilcox, 2001).

If a set of N observations, $X_1, \ldots, X_N$, is placed in order by size, the set can be identified by the order statistics, $X_{(1)}, \ldots, X_{(N)}$. If N is odd, the median, M, becomes the middle value, $X_{\left(\frac{N+1}{2}\right)}$.

If N is even, $\underline{M}$ becomes the midpoint of the middle two values, $\left\{ X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N+2}{2}\right)} \right\} / 2$. The median of the absolute deviations from the median (MAD), can be taken as a measure of variability. In particular, MAD/.6745 can be taken as an estimate of the population standard deviation, σ, in a normal distribution. Dividing an observation's absolute deviation from M by MAD/.6745, defines the MAD statistic which can be taken as an estimated value in a standard normal deviate (Wilcox, 2001. p. 36). Wilcox suggests that a ratio exceeding 2.0 identifies the observation as an outlier. The use of the MAD statistic removes the problem of masking. However, the criterion value, 2.0, may be too small, identifying too many observations as outliers. For example, if one is drawing random samples from a perfectly, normally distributed population then the probability of a standard normal deviate exceeding 2.0 is .0455. A sample of size, N = 100, could be expected to have four or five observations identified as outliers (i.e. 4 or 5 false positives).

Another approach using the median, M, can be raced back to Tukey's (1977) box-and-whisker plots. For N even, the ordered values, $X_{(i)}$, are divided into the top and bottom half. The median of the bottom half is Q_1, the first quartile of the original data. The median of the top half is Q_3, the third quartile of the original data. For N odd, the ordered values, $X_{(i)}$, are again divided into the top and bottom halves but the middle value (i.e. M) is included in both the top and bottom half. The values of Q_1 and Q_3 are again taken as the medians of the respective subgroups.

The interquartile range, IR, is $Q_3 - Q_1$. Any observation X_i exceeding $Q_3 + \underline{m}IR$ (with $\underline{m}$ usually taken to be 1.5), is identified as an outlier. Likewise, any observation X_i less than $Q_1 - \underline{m}IR$, is identified as an outlier. In sampling from a normal distribution, the probability of obtaining a single observation outside this interval (with $\underline{m} = 1.5$) would be .0070. In a sample of size, N = 100, one should expect only about one such observation identified as an outlier (i.e. one false positive). The multiplier, $\underline{m}$, could be increased to reduce the number of false positives but how high should it be and what balance should be set between false positives and false negatives?

Some authors have presented illustrative data sets when defining outliers. Everitt (2002, p. 274) identified the value 198 as an outlier in the data set {125. 128, 130, 131, 198}. For that data set, $\underline{M} = 130$ and MAD = 2. The MAD statistic for the observation, 198, would be 22.5 and well above the 2.0 cutoff value. If Everitt's data set were to be taken as a defining criterion for an outlier then the MAD statistic would need to exceed 22.5. The values $Q_3 = 131$ and IR = 3 would require an IR multiplier of m = 22.4 to match the value 198. It is unlikely that Everitt or any other author intended to use a data set to define a cutoff point for an outlier but Everitt is using a much more extreme example than has been recommended for outlier detection.

Results

Stigler (1977) reported 24 data sets that may be of use in the present investigation. Most of the

data sets were subsets of larger sets. Each data set contained observations of 18th and 19th century investigations of physical phenomena for which nearly exact values are now known. Such data sets make it possible to compare statistical estimates to 'true values' in real data. Data Sets 1 to 8 all estimated the parallax of the sun with a 'true value' of 8.798. The 158 values were combined and designated Data Set 25 for the present investigation. Data Set 17 included 23 observations from Michelson's 1882 data estimating the velocity of light with a 'true value' of 710.5. Data Set 23 included 66 observations from Newcomb's measurements of the passage of light with a 'true value' of 33.02. Data Set 19 included 29 observations from Cavendish's 1798 determinations of the density of the earth with a 'true value' of 5.517. Data Set 24 included 100 observations from Michelson's 1879 estimation of the velocity of light in air with a 'true value' of 734.5. These five data sets include all of the data reported by Stigler.

Stigler (1977) reported trimming at 10%, 15%, and 25%. Stigler included eight other robust estimators for a total of 11. For each data set the 11 estimators were used to estimate the true value. The mean absolute deviation of the 11 estimators from the true value was designated s_j for data set j. For a given data set j, each of the eleven estimators had a relative error computed as the deviation of the estimated value and true value then divided by s_j. These relative errors were one criterion used to compare the 11 estimators.

The five data sets selected for the present investigation were used to evaluate various degrees of trimming. The present approach is to remove one observation from each end of the ordered data set and calculate the relative error just as was done by Stigler. Additional pairs were removed until the minimum relative error was determined. The minimum relative error satisfied two objectives. First, it established an ideal degree of trimming for each data set. Second, it provided an estimator of an outlier detection criterion. That is, if outliers are responsible for poor estimation then the point at which estimation is best might be taken as the point at which an outlier or multiple outliers have been eliminated.

Table 1 presents all 23 observations for Data Set 17 and the analysis needed for outlier detection. The largest observation, 1051, produces a MAD statistic of 4.061 as the most extreme of the 23 observations. Table 2 presents the relative errors (REs) for the mean, trimmed means eliminating one to five pairs of observations, and the median. The minimum RE is .8418 and occurs with a single pair of means removed or 5% trimming.

From Table 1 the largest and smallest observations, 1051 and 573, are considered to be potential outliers. Their elimination produces the minimum RE. The criterion for MAD statistic must be less than 4.061 in order to ensure that this most extreme pair is rejected. However, if the criterion is less than 2.874 then a second pair of means would be trimmed. The midpoint, 3.468, of 4.062 and 2.874 could be taken as the best estimate for outlier detection for Data Set 17 to reject one and only one pair of means.

The interquartile range, IR, in Table 1 is IR = 803 – 703.5 = 99.5. The maximum IR multiplier, $\underline{m}$, to ensure that either $Q_3 + \underline{m}IR$ or $Q_1 - \underline{m}IR$ will lead to the rejection of the most extreme pair, 1051 and 573, is 2.5. Similarly, the minimum value of m to prevent the detection of a second pair of means is 1.261. The midpoint is $\underline{m} = (1.261 + 2.5)/2 = 1.88$.

Applying the same analysis as was applied to Data Set 17 to the other four data sets produces the results summarized in Table 3. Averages are calculated for four data sets (17, 19, 23 & 25). Data Set 24 is separated and appears to be a possible outlier among the five data sets. The averages of the four relevant data sets are shown and the midpoints of maximum and minimum averages are presented as well. The value of 3.5 for MAD statistic cutoffs is well above the 2.0 value suggested by Wilcox. The 2.0 value for $\underline{m}$, the IR multiplier, is well above the original value of 1.5.

The optimal trimming percentages of the five data sets are 1, 5, 7, 10, and 28. The MAD statistic for the value 28 is 4.72. That exceeds the original 2.0 criterion as well as the 3.5 criterion derived from the other four data sets. The cutoff point for the IR multiplier, $\underline{m} = 2.0$, would be $Q_3 + 2.0IR = 10 + 2.0(5) = 20.0$. The 28% trimming of Data Set 24 is well above this 20.0 cutoff.

Table 1. Analysis of Data Set 17, Michelson's 1882 Data Estimating the Velocity of Light with a 'true value' of 710.5

	X	Order Sequence	D = |X-M|	Ordered D Values	D(.6745/MAD)
	1051	1	277	277	4.061
	883	2	109	201	1.598
	851	3	77	196	1.129
	820	4	46	175	0.674
	816	5	42	163	0.616
$Q_3 = 803$	809	6	35	109	0.513
	797	7	23	92	0.337
	796	8	22	78	0.323
	796	9	22	77	0.323
	781	10	7	63	0.103
	778	11	4	51	0.059
M = 774	774	12	0	**46**	0.000
	772	11	2	42	0.029
	748	10	26	35	0.381
	748	9	26	26	0.381
	723	8	51	26	0.748
	711	7	63	23	0.924
$Q_1 = 703.5$	696	6	78	22	1.144
	682	5	92	22	1.349
	611	4	163	7	2.390
	599	3	175	4	2.566
	578	2	196	2	2.874
	573	1	201	0	2.947

MAD = 46

Table 2. Trimmed Means and Relative Errors (REs) for Data Set 17 with $s_j = 48$ with RE Calculated for the Mean and Up to Five Pairs of Values Trimmed. Optimal trimming occurs at 5% with RE = 0.8418.

	Value	RE	Trimming
Mean =	756.217	.9524	0%
Mean – 1 =	750.905	**0.8418**	5%
Mean – 2 =	753.053	0.8865	10%
Mean – 3 =	756.353	0.9553	15%
Mean – 4 =	761.800	1.0688	20%
Mean – 5 =	763.769	1.1098	25%
Median =	774	1.3229	

Table 3. Maximum and Minimum Values Needed for the Optimal Trimming

	DS25	DS 17	DS23	DS19	Ave.	DS24
Opt. Trim	1%	5%	7%	10%	5.75%	28%
MAD-MAX	7.05	4.062	2.474	1.64	3.805	0.5395
Midpoint	6.695	3.468	2.2485	1.58	3.5	0.5395
MAD-MIN	6.34	2.874	2.023	1.52	3.189	0.5395
IR-MAX	5.479	2.5	1.143	0.646	2.439	-0.0556
Midpoint	4.6175	1.8805	0.9285	0.549	2.0	-0.11
IR-MIN	3.756	1.261	0.714	0.452	1.547	-0.1667

The cutoff from the original, $\underline{m} = 1.5$, would be $Q_3 + 1.5IR = 10 + 1.5(5) = 17.5$. Of course, 28exceeds this more conservative value of 17.5.

Conclusion

In sampling from a standard normal distribution the probability of exceeding a value of 3.5 is approximately .0005. Even in a sample of size, N = 1000, a single, false-positive indication of an outlier would not be expected. Again sampling from a standard normal distribution the probability of identifying an outlier with the $\underline{m} = 2.0$ multiplier for IR would be approximately 0.0008. In that case a sample of size, N = 1000, might be expected to produce one, false-positive observation.

As a final point, note that Data Set 24 does suggest that trimming even in excess of 20% may sometimes be justified. However, to the extent that present results are applicable, trimming by no more than 10% is more likely to be optimal.

References

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

Everitt, B. S. (2002), *The Cambridge dictionary of statistics* (2nd ed.), Cambridge, UK: Cambridge University Press.

Kowalchuk, R. K., Keselman, H. J., Wilcox, R. R., & Algina, J. (2006). Multiple comparison procedures, trimmed means and transformed statistics, *Journal of Modern Applied Statistical Methods*, *5*, 44-65.

Lix, L. M. & Keselman, H. J. (1998). To trim or not to trim: Tests of mean equality under heteroscadasticity and non normality, *Educational and Psychological Measurement*, *58*, 409-429 (Errata -58,853).

Marriott, F. H. C. (1990), *A dictionary of statistical terms* (5th ed.). New York: Longman Scientific & Technical.

Stigler, S. M. (1977), "Do robust estimators work with real data?" *The Annals of Statistics*, *5*, 1055-1098.

Tukey, J. W. (1977), *Exploratory data analysis*, Reading, MA: Addison-Wesley.

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