Research Article

New Technique to Estimate the Asymmetric Trimming Mean

A. M. H. Alkhazaleh and A. M. Razali

Faculty of Science and Technology, National University of Malaysia, 43600 UKM Bangi, Selangor D.E., Malaysia

Correspondence should be addressed to A. M. H. Alkhazaleh, ahmed_2005kh@yahoo.com

A trimming mean eliminates the extreme observations by removing observations from each end of the ordered sample. In this paper, we adopted the Hogg's and Brys's tail weight measures. In addition, a new algorithm was proposed as a linear estimator based on the quartile; we used a quartile to divide the data into three and four groups. Then two new estimators were proposed. These classes of linear estimators were examined via simulation method over a variety of asymmetric distributions. Sample sizes 50, 100, 150, and 200 were generated using R program. The results of 50 were tabulated, since we have similar results for the other sizes. These results were tabulated for 7 asymmetric distributions with total trimmed proportions 0.10 and 0.20 on both sides, respectively. The results for these estimators were ordered based on their relative efficiency.

1. Introduction

Trimming mean is a statistical measure of central tendency much like the mean and median. It involves the calculation of the mean after discarding given parts of a probability distribution or a sample at the beginning and the end of the whole data, and typically discarding an equal amount of both [1] . For most statistical applications, 5 to 25 percent of the ends are discarded. The trimming mean is a useful estimator because it is less sensitive to outliers than the mean but will still give a reasonable estimate of central tendency or mean for many statistical models. In this regard it is referred to as a robust estimator. Under normality, the best possible amount of trimming is zero. But under very small departures from normality, the mean is no longer optimal and can present rather poorly [2]. The best known study of the robust estimates of location was Princeton Robust Study [3]. They recommended α-trimmed mean to desire the trimmed proportion to minimize the estimated variance. Numerous papers on robust estimators of location have emerged since the Princeton Robust Study,

and the conclusions are as varied. Most of these studies also centered on symmetric and contaminated symmetric distributions. In addition, these estimators classically utilized one or more additional statistics, such as ancillary or selector statistics, to adjust the value of the estimator to adapt the sample distribution [4]. However, in practical applications, there is no guarantee that the observed sources are symmetric. Particularly, when the number of samples is small, the interference of outliers significantly increases. Therefore, proper asymmetric truncations need to be made. In this paper, asymmetric trimmed algorithms are suggested to estimate the parameters of asymmetric trimming mean which used the classification scheme, at the first step by tail length and skewness, and at the second step a ratio involving left and right truncations based on the quartile values. Additionally, the total truncations are assumed to be 0.10 and 0.20 of the whole data.

2. Robust Location Measures

Many location estimators can be presented in the same technique by ordering the values of the sample as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ and applying the weight function w_i [5]

$$\mu = \sum_{i=1}^{n} w_i x_{(i)}, \tag{2.1}$$

where w_i is specified to reduce the influence of certain observations in the form of weighting and x_i represents the ordered data. As for the sample mean, the value is $w_i = 1/n$. To make the comparison between different estimators easier, we will present Hogg's and Brys's Tail Weight Measures and new estimators based on the quartile in the next sections.

2.1. Hogg's and Brys's Tail Weight Measures

Hogg and Lenth [6] proposed and defined the tail measure, T, and the peakedness measure, P, while Schmid and Trede [7] have studied these estimators. In addition, Brys et al. [8] proposed and discussed the LQW_p tail weight measure; these estimators are defined as follows, respectively,

$$T = \frac{X_{0.975} - X_{0.025}}{X_{0.875} - X_{0.125}}, \tag{2.2}$$

$$P = \frac{X_{0.875} - X_{0.125}}{X_{0.75} - X_{0.25}}, \tag{2.3}$$

$$LQW_p = -\frac{X_{(1-p)/2} + X_{p/2} - 2X_{0.25}}{X_{(1-p)/2} - X_{p/2}}, \tag{2.4}$$

where $X_{(\cdot)}$ are the corresponding distribution percentiles, and the percentiles $p = 0.125$ and $p = 0.25$ should be used to preserve the robustness property of this measure according to Brys et al. [8]. The procedures of Reed and Stark [9] are adopted to define sets of adaptive linear estimators. We use these estimators to compute asymmetric trimming mean. The general proposal for their approach is presented as follows.

(i) Let ξ be the value for the total amount of trimming from the sample.

(ii) Then, the proportion to be trimmed from the lower end (ξ_L) of the sample can be determined by the proportion $\xi_L = \xi[UE_X/(UE_X + LE_X)]$, where UE_X and LE_X are the numerator and denominator portions of the previous defined T, P, and LQW_p equations (2.2), (2.3), and (2.4), respectively.

(iii) The upper trimming proportion can be defined by $\xi_2 = \xi - \xi_1$. Based on this general system, the hinge estimators, which are trimmed, these hinges used to identify the values of lower and upper observations that should be trimmed; it can be defined as the following:

$$HT\xi_1 = \xi\left[\frac{UW_T}{(UW_T + LW_T)}\right], \tag{2.5}$$

$$HP\xi_1 = \xi\left[\frac{UW_P}{(UW_P + LW_P)}\right], \tag{2.6}$$

$$HLQW_p\xi_1 = \xi\left[\frac{UW_{LQW_p}}{\left(UW_{LQW_p} + LW_{LQW_p}\right)}\right]. \tag{2.7}$$

Then, we computed the greatest integer number of ξ_2 and ξ_1. After that, we assume that $\xi_1 = r$ and $\xi_2 = s$ and the (r,s)-trimmed mean $T_{(r,s)}$ is defined as

$$T_{(r,s)} = \frac{1}{n-r-s}\left[\sum_{i=r+1}^{n-s} X_i\right], \tag{2.8}$$

where X_i is the ith ordered observation.

2.2. New Estimators Based on the Quartile

A quartile represents any of the three values, which divides the sorted data set into four equal groups. This enables each group to represent one fourth of the whole sample. Now, we define the new estimator based on the quartile to determine the asymmetric trimming mean as follows: assume that $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are ordered data, and let Q_1, Q_2, and, Q_3 be the quartiles. Then, two estimators are defined based on these values. The first estimator is computed when the data is divided into three groups, while the second estimator is computed by dividing the data into four groups. The general proposal for their approaches is defined as follows. Let $x_{(1)}$ and $x_{(n)}$ be the first and last values of the ordered data, respectively. Then, we assume that the interval $(x_{(1)}, Q_1)$ is the first group, the interval (Q_1, Q_2) is the second group, the interval (Q_2, Q_3) is the third group, and the interval $(Q_3, x_{(n)})$ is the fourth group which is the last group. This permits the new estimators to be defined.

2.2.1. Three-Group Estimator

The three-groups ($3GQ$) estimator method is defined as follows.

(i) $L3G$ is the mean of the observations for the first interval $(x_{(1)}, Q_1)$, and it can be defined as

$$L3G = \sum_{i=1}^{q_1} x_i, \tag{2.9}$$

where q_1 signifies the number of observations in the first group.

(ii) $M3G$ is the mean of the observations in the interval (Q_1, Q_3), and it can be defined as

$$M3G = \sum_{i=q_1+1}^{q_2+q_3} x_i, \tag{2.10}$$

where q_2 and q_3 are the number of observations in the second and third groups, respectively.

(iii) $U3G$ is the mean of the observations in the third group $(Q_3, x_{(n)})$, and it can be defined as

$$U3G = \sum_{i=q_1+q_2+q_3+1}^{n} x_i, \tag{2.11}$$

where n is the number of observations in the whole data set.

Then, the hinge estimator for the three-group method is defined as follows:

$$H3G = \frac{(U3G - M3G)}{(M3G - L3G)}. \tag{2.12}$$

In order to determine a proportion to be trimmed from the lower end of the sample (ξ_L), then we use the following proportion:

$$\xi_L = \xi \left[\frac{UH3G}{(UH3G + LH3G)} \right], \tag{2.13}$$

where $UH3G$ and $LH3G$ are the numerator and denominator of the estimator $H3G$, respectively, ξ is the total trimming proportion to be trimmed from the sample, and the upper trimming proportion can be defined by $\xi_U = \xi - \xi_L$. Let $\xi_L = r$ and $\xi_U = s$, then, the asymmetric trimming mean is calculated as defined by (2.8).

2.2.2. Four-Group Estimator

The four-group estimator method $(4GQ)$ is defined as follows.

(i) Let $L4G$ be the mean of the observations in the first interval, and this mean is the same as the $L3G$ equation (2.9).

(ii) $M4G1$ is the mean of the observations in the interval (Q_1, Q_2), and it can be defined as

$$M4G1 = \sum_{i=q_1+1}^{q_2} x_i. \qquad (2.14)$$

(iii) $M4G2$ is the mean of the observations in the interval (Q_2, Q_3), and it can be defined as

$$M4G2 = \sum_{i=q_1+q_2+1}^{q_3} x_i. \qquad (2.15)$$

(iv) $U4G$ is the mean of the observations in the fourth group, and it is same as $U3G$.

Then, the hinge estimator can be described for three-group method as follows:

$$H4G = \frac{(U4G - L4G)}{(M4G2 - M4G1)}. \qquad (2.16)$$

In order to determine a proportion to be trimmed from the lower end of the sample (ξ_L), then we use the following proportion:

$$\xi_L = \xi \left[\frac{UH4G}{(UH4G + LH4G)} \right], \qquad (2.17)$$

where $UH4G$ and $LH4G$ are the numerator and denominator of estimator $H4G$, respectively, ξ is the total trimming proportion to be trimmed from the sample, and the upper (ξ_U) and lower (ξ_L) trimming proportions are defined as the above.

3. Simulation Methods

In the Princeton Robust Study [3], efficiency was introduced to provide a basis for comparing two estimators. The relative efficiency of two procedures is the ratio of their efficiencies, although, frequently this phrase is used where the comparison is made between two procedures. The natural logarithm of the relative efficiency of the estimator was introduced by [9]. In this paper, we computed the relative efficiency (RE) for the proposed new measure estimators depending on the quartile values method and the comparative Hogg's and Brys's estimators method, selecting the smallest variance from the proposed methods as base, and then each variance divides this base as the Relative Efficiency. Additionally, the values of the total trimming proportions to be trimmed from the sample are 0.10 and 0.20. Random numbers were generated for each of the seven asymmetric distributions using R program. These distributions were selected Beta (2,4), Gamma (3,2), Chi square (4), Burr (3,1), Pareto (3,1), Weibull (3,1), and Skewed-normal distributions; sample sizes of 50, 100, 150, and 200 were generated for each iteration (N) of 1000, 2500, and 5000. Since similar results were found for all iteration values, we tabulated the results of the $N = 2500$ iterations with fit of the mean

for each distribution with ±0.0005. Then, we tabulated the results of a simulation study when the sample size is 50 for these distributions. The estimators were ordered within each table based on the type of the distribution with minimum relative efficiency.

4. Results

Within these classes of estimators, we are concerned with the behavior of the new estimators $3G0.10$, $3G0.20$, $4G0.10$, and $4G0.20$. The $3G0.10$ and $3G0.20$ estimators are two of the top four estimators in all the distributions studied (Table 1), while there are mixed results for the estimators $4G0.10$, and $4G0.20$. Additionally, $T0.10$ is one of the top four estimators in five of the distributions studied (five of seven). In order to investigate the performance of the new estimators, there are four properties of estimators: biases, efficiency, mean square error, and consistency [10]. In this study, the relative mean square errors are computed, as the values of the MSE for different sample sizes cannot be analogous. However, an estimator can be chosen as the minimum of the $3G10$, $3G20$, $4G10$, and $4G20$, and the relative MSE of the estimators which can be calculated by the following equation:

$$\text{RMSE(Method)} = \frac{\text{MSE(Minimum proposed Estimator)}}{\text{MSE(Estimator)}}, \tag{4.1}$$

while Table 2 showed the relative mean square errors based on the seven asymmetric distributions. The methods $3G$, $4G$, and T were consistent among the top three estimators.

5. Conclusion

Most of the previous studies on the problem of the intervention of the mean estimation data sets ignored the type of the distribution of the datasets. But statistics such as skewness and tail length, both describe the distribution characteristics. In this paper, we proposed new measure estimators by dividing the whole datasets into groups with respect to the quartile values. The boundaries of the groups were derived from quartile, in order to determine the proportions of trimming on both sides of the datasets. Therefore, these boundaries were divided with regard to the type of the probability density function with respect to the value of the quartile (first, second, and third). The proposed method was tested via a simulation study over the adaptive estimators which were proposed by Brys et al. [8] for seven asymmetric distributions. The relative efficiency (RE) is defined as the ratio of the variance of the minimum estimators for the proposed methods (three and four with proportions 0.10 and 0.20) to the variance that was derived from other hinge estimators proposed by Brys et al. [8] with the same values of proportions. The variances for proposed methods are lower than the variance derived from the Brys et al. [8] method. All the relative efficiencies for these estimators are less than or equal one. Similarly, most of the values of the RMSE for all the asymmetric distribution are also less than one except for the estimator P0.20 under Burr distribution. The methods $3G$, $4G$, and T were consistent among the top three estimators for the distributions in this study.

Table 1: Linear estimators.

Estimator	Estimate	Variance	RE	Ln(RE)	Bias
Beta distributions (2,4)					
3G0.20	0.330061	0.034774	1.000000	0	0.003266
3G0.10	0.326579	0.034809	1.001000	0.001	0.006748
4G0.10	0.342047	0.034839	1.001880	0.00188	−0.008720
T0.10	0.342047	0.034839	1.001880	0.00188	−0.008720
P0.10	0.342047	0.034839	1.001880	0.00188	−0.008720
T0.20	0.343893	0.034875	1.002900	0.0029	−0.010566
P0.20	0.343893	0.034875	1.002900	0.0029	−0.010566
LQW$_{0.125}$0.10	0.313242	0.035167	1.011290	0.01123	0.020085
LQW$_{0.25}$0.10	0.313242	0.035167	1.011290	0.01123	0.020085
4G0.20	0.357454	0.035345	1.016430	0.0163	−0.024127
LQW$_{0.125}$0.20	0.303437	0.035657	1.025380	0.02506	0.029890
LQW$_{0.25}$0.20	0.291850	0.036483	1.049160	0.04799	0.041477
Gamma distribution (3,2)					
T0.20	6.036901	7.783501	0.999751	−0.0002	−0.036792
3G0.10	5.942697	7.785443	1.000000	0	0.057412
T0.10	5.942697	7.785443	1.000000	0	0.057412
LQW$_{0.125}$0.10	5.942697	7.785443	1.000000	0	0.057412
3G0.20	5.845643	7.806007	1.002641	0.00264	0.154466
4G0.10	6.168469	7.810492	1.003217	0.00321	−0.168360
P0.10	6.168469	7.810492	1.003217	0.00321	−0.168360
LQW$_{0.125}$0.20	5.845643	7.806007	1.002641	0.00264	0.154466
LQW$_{0.25}$0.10	6.168469	7.810492	1.003217	0.00321	−0.168360
P0.20	6.228680	7.834392	1.006287	0.00627	−0.228571
LQW$_{0.25}$0.20	6.228680	7.834392	1.006287	0.00627	−0.228571
4G0.20	6.439132	7.974888	1.024333	0.02404	−0.439023
Chi-square distribution (4)					
3G0.10	3.997780	9.828042	1.000000	0	0.002420
T0.10	3.997780	9.828042	1.000000	0	0.002420
P0.10	3.997780	9.828042	1.000000	0	0.002420
3G0.20	4.089484	9.836008	1.000811	0.00081	−0.089284
P0.20	3.881409	9.842147	1.001435	0.00143	0.118790
T0.20	3.881409	9.842147	1.001435	0.00143	0.118790
4G0.10	4.301156	9.918610	1.009215	0.00917	−0.300956
4G0.20	4.348989	9.949690	1.012378	0.0123	−0.348789
LQW$_{0.125}$0.10	3.345020	10.257297	1.043677	0.04275	0.655180
LQW$_{0.25}$0.10	3.345020	10.257297	1.043677	0.04275	0.655180
LQW$_{0.125}$0.20	2.791903	11.288018	1.148552	0.1385	1.208297
LQW$_{0.25}$0.20	2.791903	11.288018	1.148552	0.1385	1.208297
Burr distribution (3,1)					
3G0.20	0.494108	0.278845	1.000000	0	0.006350
P0.20	0.494108	0.278845	1.000000	0	0.006350
3G0.10	0.492520	0.278868	1.000081	0.000081	0.007938
T0.10	0.492520	0.278868	1.000081	0.000081	0.007938
P0.10	0.492520	0.278868	1.000081	0.000081	0.007938
T0.20	0.462763	0.280226	1.004951	0.004939	0.037694

Table 1: Continued.

Estimator	Estimate	Variance	RE	Ln(RE)	Bias
4G0.10	0.541519	0.280491	1.005902	0.005885	−0.041061
4G0.20	0.546247	0.280902	1.007375	0.007348	−0.045790
LQW$_{0.125}$0.10	0.388761	0.291281	1.044597	0.043631	0.111697
LQW$_{0.25}$0.10	0.388761	0.291281	1.044597	0.043631	0.111697
LQW$_{0.125}$0.20	0.314296	0.313462	1.124141	0.117019	0.186162
LQW$_{0.25}$0.20	0.314296	0.313462	1.124141	0.117019	0.186162
Pareto distribution (3,1)					
3G0.20	1.513229	5.210438	1.000000	0	−0.013318
4G0.20	1.513229	5.210438	1.000000	0	−0.013318
T0.20	1.513229	5.210438	1.000000	0	−0.013318
3G0.10	1.622767	5.225354	1.002863	0.00286	−0.122856
4G0.10	1.622767	5.225354	1.002863	0.00286	−0.122856
T0.10	1.622767	5.225354	1.002863	0.00286	−0.122856
P0.10	1.370864	5.226914	1.003162	0.00316	0.129048
p0.20	1.099740	5.370398	1.030700	0.03024	0.400172
LQW$_{0.125}$0.10	0.919826	5.546760	1.064548	0.06255	0.580086
LQW$_{0.25}$0.10	0.919826	5.546760	1.064548	0.06255	0.580086
LQW$_{0.125}$0.20	0.728119	5.805925	1.114287	0.10821	0.771793
LQW$_{0.25}$0.20	0.775311	5.735306	1.100734	0.09598	0.724600
Weibull distribution (1,3)					
3G0.10	2.999831	9.738460	1.000000	0	0.000277
P0.10	2.999831	9.738460	1.000000	0	0.000277
3G0.20	3.074395	9.743978	1.000567	0.00057	−0.074287
P0.20	3.074395	9.743978	1.000567	0.00057	−0.074287
4G0.10	3.253236	9.802534	1.006579	0.00656	−0.253128
T0.10	2.765150	9.793665	1.005669	0.00565	0.234958
4G0.20	3.328679	9.846419	1.011086	0.01103	−0.328572
T0.20	2.618965	9.883730	1.014917	0.01481	0.381143
LQW$_{0.125}$0.10	2.346914	10.165121	1.043812	0.04288	0.653193
LQW$_{0.25}$0.10	2.346914	10.165121	1.043812	0.04288	0.653193
LQW$_{0.125}$0.20	1.692180	11.449135	1.175662	0.16183	1.307928
LQW$_{0.25}$0.20	1.692180	11.449135	1.175662	0.16183	1.307928
Skewed-normal distribution					
3G0.20	2.593322	0.214400	1.000000	0	0.003977
T0.20	2.593322	0.214400	1.000000	0	0.003977
LQW$_{0.125}$0.20	2.593322	0.214400	1.000000	0	0.003977
3G0.10	2.586861	0.214494	1.000434	0.00043	0.010438
T0.10	2.586861	0.214494	1.000434	0.00043	0.010438
LQW$_{0.125}$0.10	2.586861	0.214494	1.000434	0.00043	0.010438
P0.10	2.623642	0.215079	1.003163	0.00316	−0.026343
p0.20	2.627732	0.215311	1.004246	0.00424	−0.030432
LQW$_{0.25}$0.10	2.552086	0.216429	1.009461	0.00942	0.045214
4G0.10	2.661124	0.218458	1.018926	0.01875	−0.063825
4G0.20	2.697760	0.224477	1.046999	0.04593	−0.100461
LQW$_{0.25}$0.20	2.495980	0.224650	1.047806	0.0467	0.101319

Table 2: Relative mean square error.

Estimator	Skewed normal	Beta (2,4)	Gamma (3,2)	Chisq (4)	Burr (3,1)	Pareto (3,1)	Weibull (1,3)
3G10	0.9991	0.997996	1	1	1	0.9943	1
3G20	1	1	0.9947	0.9983	1.0001	1	0.9988
4G10	0.9635	0.9962	0.9936	0.9819	0.9884	0.9943	0.98701
4G20	0.9140	0.9682	0.9536	0.9758	0.9856	1	0.9783
T10	0.9991	0.9962	1	1	1	0.9943	0.988
T20	1	0.9942	1.0004	0.9971	0.9903	1	0.9710
P10	0.9937	0.9962	0.9936	0.9196	1	0.9937	1
P20	0.9915	0.9942	0.9875	0.7709	1.0001	0.9421	0.9988
LQW$_{0.125}$10	0.9991	0.9779	1	0.9196	0.9182	0.8856	0.9194
LQW$_{0.125}$20	1	0.9516	0.9947	0.7709	0.8012	0.8139	0.7400
LQW$_{0.25}$10	0.9814	0.9779	0.9936	1	0.9182	0.8856	0.91943
LQW$_{0.25}$20	0.9127	0.9105	0.9875	0.9971	0.8012	0.8323	0.7400

Acknowledgment

References

[1] R. R. Wilcox, "Trimmed means," *Encyclopedia of Statistics in Behavioral Science*, vol. 4, pp. 2066–2067, 2005.

[2] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, *Robust Statistics*, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York, NY, USA, 1986.

[3] D. F. Andrews, P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey, *Robust Estimates of Location: Survey and Advances*, Princeton University Press, Princeton, NJ, USA, 1972.

[4] J. F. Reed and D. B. Stark, "Robustness estimators of location: a classification of linear and adaptive estimators," *Journal of Applied Statistics*, vol. 21, pp. 95–124, 1994.

[5] Z. Leonowicz, J. Karvanen, and S. L. Shishkin, "Trimmed estimators for robust averaging of event-related potentials," *Journal of Neuroscience Methods*, vol. 142, no. 1, pp. 17–26, 2005.

[6] R. V. Hogg and R. V. Lenth, "A review of some adaptive statistical techniques," *Communications in Statistics. A*, vol. 13, no. 13, pp. 1551–1579, 1984.

[7] F. Schmid and M. Trede, "Simple tests for peakedness, fat tails and leptokurtosis based on quantiles," *Computational Statistics & Data Analysis*, vol. 43, no. 1, pp. 1–12, 2003.

[8] G. Brys, M. Hubert, and A. Struyf, "Robust measures of tail weight," *Computational Statistics & Data Analysis*, vol. 50, no. 3, pp. 733–759, 2006.

[9] J. F. Reed III and D. B. Stark, "Hinge estimators of location: robust to asymmetry," *Computer Methods and Programs in Biomedicine*, vol. 49, no. 1, pp. 11–17, 1996.

[10] R. S. Pindyck and D. L. Rubinfeld, *Econometric Models and Economic Forecasts*, McGraw-Hill, Singapore, 4th edition, 1998.