International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:2, No:6, 2008

Trimmed Mean as an Adaptive Robust Estimator of a Location Parameter for Weibull Distribution

Carolina B. Baguio

***Abstract*—**One of the purposes of the robust method of estimation is to reduce the influence of outliers in the data, on the estimates. The outliers arise from gross errors or contamination from distributions with long tails. The trimmed mean is a robust estimate. This means that it is not sensitive to violation of distributional assumptions of the data. It is called an adaptive estimate when the trimming proportion is determined from the data rather than being fixed a "priori'.

The main objective of this study is to find out the robustness properties of the adaptive trimmed means in terms of efficiency, high breakdown point and influence function. Specifically, it seeks to find out the magnitude of the trimming proportion of the adaptive trimmed mean which will yield efficient and robust estimates of the parameter for data which follow a modified Weibull distribution with parameter $\lambda = 1/2$, where the trimming proportion is determined by a ratio of two trimmed means defined as the tail length. Secondly, the asymptotic properties of the tail length and the trimmed means are also investigated. Finally, a comparison is made on the efficiency of the adaptive trimmed means in terms of the standard deviation for the trimming proportions and when these were fixed a "priori".

The asymptotic tail lengths defined as the ratio of two trimmed means and the asymptotic variances were computed by using the formulas derived. While the values of the standard deviations for the derived tail lengths for data of size 40 simulated from a Weibull distribution were computed for 100 iterations using a computer program written in Pascal language.

The findings of the study revealed that the tail lengths of the Weibull distribution increase in magnitudes as the trimming proportions increase, the measure of the tail length and the adaptive trimmed mean are asymptotically independent as the number of observations n becomes very large or approaching infinity, the tail length is asymptotically distributed as the ratio of two independent normal random variables, and the asymptotic variances decrease as the trimming proportions increase. The simulation study revealed empirically that the standard error of the adaptive trimmed mean using the ratio of tail lengths is relatively smaller for different values of trimming proportions than its counterpart when the trimming proportions were fixed a 'priori'.

***Keywords*—**Adaptive robust estimate, asymptotic efficiency, breakdown point, influence function, L-estimates, location parameter, tail length, Weibull distribution.

I. Introduction

IT has been generally realized that outliers in the data, which do not appear to come from the normal distribution but may have arisen from Weibull distribution with long (heavy) tails have unusually large influence on the estimates. Robust methods of estimation have been developed to reduce the influence of outliers in the data, on the estimates. Statistics which are represented as a linear combination of order statistics, called L-estimates make a proper class of robust estimates for estimating a location parameter. The trimmed mean denoted by $T_{\alpha,\beta}$ is a special class of L-estimates.

Adaptive estimators pertain to those which are derived from the distribution of the data. For convenience in the choice of trimming proportions for trimmed means the user or researcher simply consider the proportion of outlying data on the left and right tails which caused the inflation of the variance which are usually termed as trimming proportions when fixed a "priori". In many studies it was shown theoretically and empirically that choosing the trimming proportions based from the distribution of the data proved to be more efficient than being fixed. This approach is presently called the Exploratory Data Analysis (EDA) wherein the structure of the data can be discerned through the graphs of the distribution in order to rule out the presence of outliers as well as to get a visual perception on the length of the tails of the distribution and the symmetry. In the case of asymmetric distributions the side of the distributions having longer tails must be trimmed with sufficiently large proportion. The use of the ratio of the length of tails in the selection of the trimming proportions is not yet popular even up to these days and still has to be explored for both the symmetric and asymmetric distributions with long tails.

II. Objectives of the Study

The main objective of this study is to find out the robustness properties of the adaptive trimmed means in terms of efficiency, high breakdown point and influence function. Specifically, it seeks to find out the magnitude of the trimming proportion of the adaptive trimmed mean which will yield efficient and robust estimates of the parameter for data which follows a modified Weibull distribution with parameter $\lambda = 1/2$, where the trimming proportion is determined by a ratio of two trimmed means defined as the tail length. Secondly, the asymptotic properties of the tail length and the trimmed means are also investigated. Finally, a comparison is made on the efficiency adaptive trimmed means in terms of the standard deviation for the trimming proportions and when these were fixed a "priori".

B. Baguio is with MSU-IIT, Iligan City 9200, Philippines (e-mail: cbbaguio@yahoo.com).

III. Theory and Concepts

A. Weibull Distribution

The Weibull distribution for 4 parameter values of λ is shown in Fig. 1 below. It is apparent from this figure that the density functions for the four parameter values are long tailed. This implies that the variances of the distribution are relatively large and surely will affect the Ordinary Least Squares (OLS) estimates. In order to address this situation, the estimate to be considered is the adaptive trimmed mean. However, the problem that is outstanding is what is the magnitude of the trimming proportion on the tails of the distribution which can result to efficient estimates? The probability distribution of the Weibull distribution is:

$$(k/\lambda)(x/\lambda)^{(k-1)}e^{-(x/\lambda)^k} \qquad (1)$$

where k >0 is the shape parameter and the λ >0 is the scale parameter and x is nonnegative real number. The cumulative distribution function is $1 - e^{-(x/\lambda)^k}$

The mean is $\lambda\Gamma\left(1+\frac{1}{k}\right)$.

and the variance is $\lambda^2\Gamma\left(1+\frac{2}{k}\right) - \mu^2$.

The modified Weibull distribution used in this study has the probability density

$$F = (2\Gamma(1+1/\lambda))^{-1} e^{-|x|^{\lambda}}$$

where λ is the shape parameter. The variance of this distribution is

$$\text{Var}(F) = \frac{\Gamma(3/\lambda)}{\Gamma(1/\lambda)}$$

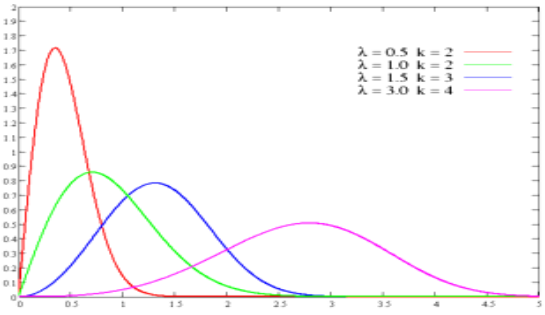

Fig. 1 Probability Density Function of Weibull Distribution (1) for parameters λ = .50,1.0,1.5, and 3.0 and k=2,2,3,4 respectively

B. Properties of Trimmed Mean ($T_{\alpha,\beta}$)

A number of nice properties of $T_{\alpha,\beta}$, have been cited in the textbook [7]. Refer to this book for short in the following. The properties are given as follows: (1) $T_{\alpha,\beta}$ is robust to outliers up to $100\alpha\%$ on the left side and $100(1-\beta)\%$ on the right side, (ii) the asymptotic efficiency relative to the untrimmed mean is $\geq (1-\alpha-\beta)^2$ and, (iii). It is simple to compute and its standard error may be estimated from the $(\alpha+\beta)-$ winsorized sample. Moreover, $T_{\alpha,\beta}$, is consistent and asymptotically normal when the underlying distribution F is continuous at the unique quantiles $\xi_\alpha = F^{-1}(\alpha)$ *and* $\xi_{1-\beta} = F^{-1}(1-\beta)$, [8].

Generally, a larger proportion of data points should be trimmed when F has a longer tail and a smaller proportion, otherwise. Therefore, the choice of the trimming proportions would depend upon a priori knowledge of the tail-length of F. This information may not be available. Therefore, there is a need to determine the trimming proportions from the sample itself. The trimmed mean is called adaptive when the trimming proportions are determined from the data. [4] considered the kurtosis as a measure of the tail-length, and use the sample kurtosis to determine the trimming proportions. His estimate of the center of a symmetric distribution is given by a combination of several trimmed means, associated with different trimming proportions. Subsequently, [5] proposed a choice of the trimming proportions, based on the ratio R of two L-estimates. In a recent paper, [1] have reviewed Hogg's method and presented certain theoretical and empirical results on the application of R as a measure of tail-length, to determine the trimming proportions of an adaptive trimmed mean. Here, we derive the robustness and asymptotic properties of the estimates $T_{\alpha,\beta}$, and R.

Let F_n be the empirical distribution derived from a sample of n observations from the modified Weibull distribution with parameter λ = ½. Consider now the robustness of $T_{\alpha,\beta}$ with respect to the given measures of robustness such as the breakdown point and the Influence function developed by [6]. Denote the descriptive measure of the trimmed mean by $T_{\alpha,\beta}(\mathbf{F})$ and the sample estimate $T_{\alpha,\beta}(\mathbf{F_n})$ by $T_{\alpha,\beta}$ suppressing n for convenience. Assume that F is a continuous distribution with a location parameter θ which is to be estimated. If $\beta = (1-\alpha)$ then $T_{\alpha,\beta}(F)$ is a measure of location of F. [7]. By an outlier in a location parameter context, without being very specific, refers to an observation which is considerably larger in absolute value than the bulk of the sample values. Various specific definitions of an outlier can be given. For example, an observation may be designated as an outlier if it is more than two or three times the interquartile range from the median. [7] defined the breakdown point ε^* of T(F) as the minimum proportion ε of outlier contamination at x for which $T(F_{x,\varepsilon})$ is unbounded in x, where $F_{x,\varepsilon}$ is given by $(1-\varepsilon)F + \varepsilon\Delta_x$. The finite sample

breakdown point ε^* is the smallest proportion of the n observations in the sample which can render the estimator out of bound. It is easily seen that the breakdown point ε^* for the trimmed mean is equal to $\min(\alpha, 1-\beta)$. [2] gives the derivation of the influence function of $T_{\alpha,\beta}$ as shown (2) with ξ_α and ξ_β as the α *and* β quantiles of F. These quantiles are uniquely determined.

$$(\beta-\alpha)I_{T_{\alpha,\beta},F}(x)=\begin{cases}\xi_\alpha - W_{\alpha,\beta}, & x<\xi_\alpha\\ x-W_{\alpha,\beta,} & \xi_\alpha \le x \le \xi_\beta\\ \xi_\beta - W_{\alpha,\beta}, & x>\xi_\beta\end{cases} \quad (2)$$

where

$$W_{\alpha,\beta} = (\beta-\alpha)T_{\alpha,\beta}(F) + \alpha\xi_\alpha + (1-\beta)\xi_\beta \quad (3)$$

denotes the $(\beta-\alpha)$ – winsorized mean. Note that

$$\mathrm{E}\ I_{T_{\alpha,\beta},F}(x)=0 \quad (4)$$

C. Asymptotic Property of the Trimmed Mean

[2] derived the asymptotic property of the trimmed mean as follows

$$\begin{aligned}T_{\alpha,\beta} &= T_{\alpha,\beta}(F) + \int I_{T_{\alpha,\beta},F}(x)d(F_n - F(x) + R_n\\ &= T_{\alpha,\beta}(F) + \int I_{T_{\alpha,\beta},F}\, d\, F_n(x) + R_n\\ &= T_{\alpha,\beta}(F) + 1/n\sum_{i=1}^{n} I_{T_{\alpha,\beta}}F(x_i) + R_n\end{aligned} \quad (5)$$

where $\sqrt{n}\ R_n \xrightarrow{P} 0,$ as $n \to \infty$ and x_i denote the sample values. In fact, $R_n = O_p(1/n)$ under reasonable conditions. Therefore

$$\sqrt{n}\ (T_{\alpha,\beta} - T_{\alpha,\beta}(F) \approx n^{1/2}\sum_{i=1}^{n} I_{T_{\alpha,\beta},F}(x_i) \quad (6)$$

An application of the Central Limit Theorem shows that

$$\sqrt{n}\ (T_{\alpha,\beta} - T_{\alpha,\beta}(F)) \xrightarrow{d} N(0, V(F)) \quad (7)$$

where

$$\begin{aligned}V(F) &= \mathrm{var}\ I_{T_{\alpha,\beta},F}(x)\\ &= E(I_{T_{\alpha,\beta},F}(x))^2.\end{aligned} \quad (8)$$

Here x denotes a random observation from the distribution F.

From (2), the equation in (9) was derived.

$$(\beta-\alpha)^2 V(F) = \alpha(\xi_\alpha - W_{\alpha,\beta})^2 + (1-\beta)(\xi_\beta - W_{\alpha,\beta})^2 + \int_{\xi_\alpha}^{\xi_\beta}(x-w_{\alpha,\beta})^2 dF(x). \quad (9)$$

In comparison, the asymptotic variance of the ε^{th} quantile is given by

$$\mathrm{var}\ (\varsigma_{\varepsilon,F}(x)) = \varepsilon(1-\varepsilon)/(f(\xi_\varepsilon))^2. \quad (10)$$

D. Relative Efficiency

Since the trimmed mean is a special member of the class of L-estimator, it is interesting to compare its asymptotic variance with the asymptotic variance of any other member of the class such as the untrimmed mean. The descriptive measure of an L-estimate, measuring location, is given by

$$T(F) = \int_0^1 F^{-1}(t)\, d\, k(t) \quad (11)$$

where k is a probability distribution on (0,1).

The trimmed mean $T_{\alpha,1-\alpha}(F)$ is obtained by taking k uniform on $(\alpha, 1-\alpha)$. Let T_1 and T_2 be two L-estimators determined by k_1 and k_2, and let f_1 and f_2 denote the densities of K_1 and K_2, respectively. Suppose that

$$0 \le (f_2(t)/f_1(t)) \le A \quad (12)$$

where $A > 1$. [3] have shown that (12) implies

$$V(T_2, F) \le A^2 V(T_1, F).$$

Here, $V(T_1, F)$ and $V(T_2, F)$ denote the asymptotic variances of T_1 and T_2, respectively. It follows from the above result that the asymptotic relative efficiency of $T_{\alpha,1-\alpha}$ relative to the untrimmed mean is bounded below by $(1-2\alpha)^2$.

E. Estimate of the Variance of $T_{\alpha,\beta}$

Let $r = [n\alpha]$ *and* $s = [n(1-\beta)]$ where [x] denotes the integer part of x, and let $X_{(I)}$ denote the I^{th} order statistic from the sample. The winsorized sample is given by $y_1, \ldots, y_n$, where

$$y_i = \begin{cases}x_{(r+1)}, & r \le i\\ x_{(i)} & r < i \le s\\ x_{(s)} & s < i\end{cases} \quad (13)$$

the winsorized sample variance, denoted by S_w^2, is given by

$$S_w^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \qquad (14)$$

where $\bar{y}$ denotes the winsorized sample mean. From (9) and the asymptotic relation (7) an estimate of the variance of $T_{\alpha,\beta}$, was derived given by

$$\hat{v}\,ar\ T_{\alpha,\beta} = S_w^2/(n(\beta-\alpha)^2). \qquad (15)$$

E. Tail-length

A sample from a distribution with longer tail is likely to contain a larger number of outliers compared to a sample from a distribution with shorter tail. Therefore, for estimating a location parameter, using a trimmed mean, we should trim the sample more if the underlying distribution has longer tail, for the sake of robustness. In order to determine the trimming proportions we need to have a measure of the tail-length which we should be able to estimate with some precision. A measure of tail-length which has been referred to in the introduction is given, as follows.

It is assumed throughout the following that the modified Weibull distribution F with parameter λ = ½ is symmetric about a location parameter θ and that the trimmed mean is symmetrized, that is, the trimming proportions are equal ($\beta = 1-\alpha$). In this case the descriptive measure of the trimmed mean is denoted by

$$T_\alpha(F) = \frac{1}{1-2\alpha}\int_{\xi_\alpha}^{\xi_{1-\alpha}} x\,dF(x) \qquad (16)$$

and the sample estimate of $T_\alpha(F)$ is given by

$$T_\alpha(F_n) = \frac{1}{n-2m}\sum_{i=(m+1)}^{n-m} X_{(i)} \qquad (17)$$

where $m = [n\alpha]$ and $x_{(i)}$ denotes the i[th] ordered value in the sample.

Let $0 < \gamma < \delta < \alpha < 1/2.$ The tail-length of F is given by

$$R_{\gamma,\delta}(F) = (T_{1-\gamma,1}(F) - T_{0,\gamma}(F)) / (T_{1-\delta,1-\gamma}(F) - T_{\gamma,\delta}(F)). \qquad (18)$$

notice that the numerator and denominator of the right side of (18) are each invariant with respect to translation. Therefore, $R_{\gamma,\delta}(F)$ is invariant with respect to both translation and scale transformation. Clearly, $R_{\gamma,\delta}(F) \geq 1$. It is also clear that a longer value of the tail-length will induce a larger value of $R_{\gamma,\delta}$. **Therefore, $R_{\gamma,\delta}(F)$ is a suitable measure of the tail-length of F.**

The sample estimate of the tail-length is given by

$$R_{\gamma,\delta}(F_n) = (T_{1-\gamma,1}(F_n) - T_{0,\gamma}(F_n))/(T_{1-\delta,1-\gamma}(F_n) - T_{\gamma,\delta}(F_n)). \qquad (19)$$

$$= A_n/B_n, \text{ say.}$$

The asymptotic property of $R_{\gamma,\delta}(F_n)$ is derived from an application of the asymptotic relation (6). Using this and the symmetry of F, a simple algebraic computation shows that the trimmed mean $T_\alpha(F_n)$ and A_n and B_n are pair-wise uncorrelated when n is large. Moreover they are jointly normally distributed, asymptotically. Thus we have that

Theorem 1: The modified Weibull distribution F with parameter λ = ½ is symmetrically distributed then it follows that $R_{\gamma,\delta}(F_n)$ and $T_\alpha(F_n)$ are asymptotically independent, as $n \to \infty$. Moreover, $R_{\gamma,\delta}(F_n)$ is asymptotically distributed as a ratio of two independent normal random variables, given by A_n/B_n.

Remarks:

The statistical independence of the trimmed mean $T_\alpha(F_n)$ and the tail-length $R_{\gamma,\delta}(F_n)$, given by Theorem 1, is a useful result. If the trimming proportions α of the trimmed mean is based on the tail-length $R_{\gamma,\delta}(F_n)$ the result implies that the asymptotic distribution of the adaptive trimmed mean is the same as if α was fixed a priori.

IV. Methods and Materials

By using the modified form of the standard Weibull distribution with the formula for the density function as:

$$(2\Gamma(1+1/\lambda))^{-1}e^{-|x|^\lambda}$$

The tail lengths $R_{\gamma,\delta}(F)$ for certain values of γ and δ and the corresponding respective asymptotic variances $V_\alpha(F)$ for the Weibull distribution were computed for λ = ½ using the formulas shown in the previous sections. The results are shown in Table I.

Table I shows that the tail lengths increase as the (γ,δ) increase while the variances of the trimmed means decrease when α increase. This reveals the desirability of the tail length in the light of high breakdown point and efficiency as robustness properties.

In Table II, the selection rule is presented given the tail lengths which can be summarized as follow:

Selection Rule Says that

- if the tail lengths are x = $R_{.1,.2}(F)$ < 1.5 or y = $R_{.15,.3}(F)$ < 2.0 then use $\alpha = .05$.
- if the tail lengths are $1.5 \le x < 2.0$ or $2.0 \le y < 2.5$ then use $\alpha = .10$.
- if the tail lengths are $2.0 \le x < 2.5$ or $2.5 \le y < 3.0$ then use $\alpha = .15$.
- if the tail lengths are $2.5 \le x < 3.0$ or $3.0 \le y < 3.5$ then use $\alpha = .20$.
- if the tail lengths are $3.0 \le x$ or $3.5 \le y$ then use $\alpha = .30$.

V. Empirical Results

From the modified Weibull distribution, a random sample of n = 40 observations was generated. From these samples, the adaptive trimmed means T_A, $T_{A'}$, and the trimmed means for two fixed values T_α, for $\alpha = .05$ and .10 were computed. With m = 100 iterations, the m values of T_A, $T_{A'}$, $T_{.05}$, $T_{.10'}$ were obtained from which the standard deviations were also computed such as $S(T_A)$, $S(T_{A'})$, $S(T_{.05})$, and $S(T_{.10})$ respectively as shown in Table III. It is apparent from this table that the standard deviations are relatively smaller for the adaptive trimmed means T_A and $T_{A'}$ than those trimmed means wherein the trimming proportions were fixed a "priori" at .05 and .10. This indicates the relative efficiency of the adaptive procedure which substantiates the findings of [1].

VI. Conclusion

1. The tail length of the Weibull distribution increases in magnitudes as the trimming proportion increases.
2. The measure of the tail length and the adaptive trimmed mean are asymptotically independent as the number of observations n becomes very large or approaching infinity.
3. The tail length is asymptotically distributed as the ratio of two independent normal random variables.
4. The asymptotic variances decrease as the trimming proportions increase.
5. The standard error of the adaptive trimmed mean is relatively smaller than its counterpart wherein the trimming proportions were fixed a 'priori'.
6. The proposed adaptive rule may be used with advantage in the absence of any prior ground for an appropriate choice of the trimming proportions.

VII. Future Direction

The refinement and applicability of the selection rule on the trimming proportions corresponding to the tail lengths $R_{\gamma,\delta}(F)$ specifically for modified Weibull distribution are recommended. Furthermore, investigation on the applicability of the selection rule with some adjustments on the degree of skewness for assymetrical distribution are hereby recommended.

The extension of the procedure can be explored for quantile regression analysis on the choice of the ξ th quantile values that instead of being symmetrized, that is (ξ, 1-ξ) at two tails of the error distribution being fixed, adaptive trimming proportions based on tail lengths can be a potential choice.

References

[1] Alam, K. and Mitra, A. (1996). Adaptive Robust Estimate Based on Tail-length.Sankhya,Indian Journal of Statistics, Series B(58) 672-678
[2] Baguio, Carolina B. (1999). An Adaptive Robust Estimator of a Location Parameter. A Doctoral Dissertation.
[3] Bickel,P.J. and Lehmann,E.L. (1976). Descriptive Statistics for Nonparametric, Models III Ann. Stat. (4)1139-1158.
[4] Hogg, R.V. (1967). Some Observations in Robust Estimation. Jour. Amer. Statist. Assoc.(62) 1179=1186.
[5] Hogg,R.V.(1974). Adaptive Robust Procedures: A Partial Review and Some Suggestions for Future Applications and Theory. Jour. Amer. Statist. Assoc.(69) 909-923.
[6] Huber,P.J. (1964). Robust Estimation of a Location Parameter. Ann. Math. Stat. (35) 73-101.
[7] Staudte, R.G. and Sheather, S.J. (1990). Robust Estimation and Testing. Wiley Series in Probability and Statistics.
[8] Stigler, S. M. (1969). "Linear Functions of order Statistics," Ann. Math. Stat. 40. 770-788.

TABLE I

TAIL-LENGTH $R_{\gamma,\delta}(F)$ AND ASYMPTOTIC VARIANCE OF TRIMMED MEAN T_α

Trim size	Tail Length $R_{\gamma,\delta}(F)$				Asymptotic Variance $V_\alpha(F)$				
(γ,δ)	(.1,.2)	(.1,.25)	(.15,.30)	(.20,.30)	α = 0	α = .1	α=.15	α=.20	α=.25
Distribution									
Modified Weibull λ = ½	3.06	3.58	4.17	4.28	120.0	45.41	34.73	22.97	20.64

TABLE II

SELECTION RULE FOR THE VALUES OF TAIL LENGTHS AND α

Values of α	A. Select x = $R_{.1,.2}(F)$ if	A'. Select y = $R_{.15,.3}(F)$ if
.05	x < 1.5	y < 2.0
.10	$1.5 \le x < 2.0$	$2.0 \le y < 2.5$
.15	$2.0 \le x < 2.5$	$2.5 \le y < 3.0$
.20	$2.5 \le x < 3.0$	$3.0 \le y < 3.5$
.30	$3.0 \le x$	$3.5 \le y$

TABLE III

STANDARD ERRORS OF T_A, $T_{A'}$, T_α, $T_{\alpha'}$

Distribution	$S(T_A)$	$S(T_{A'})$	$S(T_\alpha)$ α = .05	$S(T_{\alpha'})$ α = .10
Modified Weibull λ = ½	.560	.553	.759	.632