1-2001

High Breakdown Analogs of the Trimmed Mean

David J. Olive
Southern Illinois University Carbondale, dolive@math.siu.edu

Published in *Statistics & Probability Letters*, 51, 87-92.

High Breakdown Analogs of the Trimmed Mean

David J. Olive*

Southern Illinois University

April 11, 2004

Abstract

Two high breakdown estimators that are asymptotically equivalent to a sequence of trimmed means are introduced. They are easy to compute and their asymptotic variance is easier to estimate than the asymptotic variance of standard high breakdown estimators.

KEY WORDS: MAD; M-estimators; Outliers.

*David J. Olive is Assistant Professor, Department of Mathematics, Mailcode 4408, Southern Illinois University, Carbondale, IL 62901-4408, USA.

1 INTRODUCTION

Consider the location model

$$X_i = \mu + e_i, \quad i = 1, \ldots, n \tag{1}$$

where $X_1, \ldots, X_n$ are independent and identically distributed (iid) with cumulative distribution function (cdf) F, median MED(X), mean $E(X)$, median absolute deviation MAD(X), and variance $V(X)$ if they exist. This model is often summarized by obtaining point estimates and confidence intervals for a location parameter. The natural choice for the location parameter is μ if the errors are symmetric about 0 but when the errors are asymmetric, there are many other reasonable choices.

The classical point and interval estimators use the sample mean $\bar{x}$ and standard deviation S. If a graph of the data indicates that the classical assumptions are violated, then an alternative estimator should be considered. Robust estimators can be obtained by giving zero weight to some cases and applying classical methods to the remaining data. Bickel (1965) and Stigler (1973) consider trimmed means while Davies and Gather (1993), Hampel (1985), Kim (1992), and Simonoff (1987) consider metrically trimmed means. Shorack (1974) and Shorack and Wellner (1986, section 19.3) derive the asymptotic theory for a large class of robust procedures for the iid location model. Special cases include trimmed, Winsorized, metrically trimmed, and Huber type skipped means. Also see papers in Hahn, Mason, and Weiner (1991).

One of the most popular robust methods is the $(\alpha, 1-\beta)$ trimmed mean

$$T_n = T_n(l_n, u_n) = \frac{1}{u_n - l_n} \sum_{l_n+1}^{u_n} X_{(i)} \tag{2}$$

where

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$$

are the order statistics,

$$l_n = [n\alpha],$$

$$u_n = [n\beta]$$

and [.] is the "greatest integer function" (eg [7.7] = 7). Note that the proportion of cases trimmed and the proportion of cases covered is fixed. If $\alpha = 1 - \beta$, we will call the estimator the α trimmed mean. Hence the 10% trimmed mean is the (0.1, 0.9) trimmed mean. The Winsorized mean

$$W_n = W_n(l_n, u_n) = \frac{1}{n}[l_n X_{(l_n+1)} + \sum_{i=l_n+1}^{u_n} X_{(i)} + (n - u_n)X_{(u_n)}]. \quad (3)$$

These estimators have a breakdown point of $\min(\alpha, 1 - \beta)$.

A randomly trimmed mean is

$$R_n = R_n(L_n, U_n) = \frac{1}{U_n - L_n} \sum_{i=L_n+1}^{U_n} X_{(i)} \quad (4)$$

where $L_n < U_n$ are integer valued random variables. For example, the metrically trimmed mean M_n discards data outside of the interval

$$[\text{MED}(n) - k_1\text{MAD}(n), \text{MED}(n) + k_2\text{MAD}(n)]$$

where MED(n) is the sample median, MAD(n) is the sample median absolute deviation, $k_1 \geq 1$, and $k_2 \geq 1$. The amount of trimming will depend on the distribution of the data. For example, if $k_1 = k_2 = 5.2$ and the data is normal (Gaussian), about 1% of the data will be trimmed while if the data is Cauchy, about 24% of the data will be trimmed. Hence

the upper and lower trimming points estimate lower and upper population percentiles $L(F)$ and $U(F)$ and change with the distribution F.

A high breakdown analog of the trimmed mean $R_n^*(L_n, U_n)$ takes L_n to be the maximum of the number of observations which fall to the left of MED$(n) - k$ MAD(n) and the number of observations which fall to the right of MED$(n) + k$ MAD(n) where $k > 1$ is fixed in advance. Let $U_n = n - L_n$. (Take R_n^* to be the sample median if $U_n \leq L_n$.) That is, first metrically trim, then symmetrically trim by increasing the smaller trimming proportion to equal the larger trimming proportion.

An even simpler estimator is the two stage trimmed mean $T_{2,n}^*$. In the first stage, find L_n as defined for R_n^*. Then round $100L_n/n$ up to the nearest integer, say J_n. Then $T_{2,n}^*$ is the J_n% trimmed mean. Again let $T_{2,n}^* = MED(n)$ if $J_n \geq 50$. For example, suppose that there are $n = 205$ cases and M_n trims the smallest 15 cases and the largest 20 cases. Then $L_n = 20$ and $J_n = 10$. Thus R_n^* is the 9.7561% trimmed mean while $T_{2,n}^*$ is the 10% trimmed mean.

The following section reviews the asymptotic theory of the trimmed mean and shows that R_n^* is asymptotically equivalent to the trimmed mean when the errors are symmetric. The theory of the two stage mean does not require symmetric errors.

2 ASYMPTOTICS

Truncated and Winsorized random variables are important because they simplify the asymptotic theory of location estimators. Let X be a random variable with cdf F and

let $\alpha = F(a) < F(b) = \beta$. The truncated random variable $X_T(a,b) = X_T$ has cdf

$$F_{X_T}(x|a,b) = G(x) = \frac{F(x) - F(a-)}{F(b) - F(a-)} \tag{5}$$

for $a \leq x \leq b$. Also G is 0 for $x < a$ and G is 1 for $x > b$. Below we will assume that F is continuous at a and b.

The mean and variance of X_T are

$$\mu_T = \mu_T(a,b) = \int_{-\infty}^{\infty} x dG(x) = \frac{\int_a^b x dF(x)}{\beta - \alpha}$$

and

$$\sigma_T^2 = \sigma_T^2(a,b) = \int_{-\infty}^{\infty} (x - \mu_T)^2 dG(x) = \frac{\int_a^b x^2 dF(x)}{\beta - \alpha} - \mu_T^2.$$

See Cramer (1946, p. 247).

Another type of truncated random variable is the Winsorized random variable

$$X_W = X_W(a,b) = \begin{cases} a, & X \leq a \\ X, & X \leq b \\ b, & X \geq b. \end{cases}$$

If the cdf of $X_W(a,b) = X_W$ is F_W, then

$$F_W(x) = \begin{cases} 0, & X < a \\ F(a), & X = a \\ F(x), & a < X < b \\ 1, & X \geq b. \end{cases}$$

Since X_W is a mixture distribution with a point mass at a and at b, the mean and variance of X_W are

$$\mu_W = \mu_W(a,b) = \alpha a + (1 - \beta)b + \int_a^b x dF(x)$$

and

$$\sigma_W^2 = \sigma_W^2(a,b) = \alpha a^2 + (1-\beta)b^2 + \int_a^b x^2 dF(x) - \mu_W^2.$$

Wilcox (1997, p. 141-181) replaces ordinary population means by truncated population means and uses trimmed means to create analogs of one, two, and three way anova, multiple comparisons, random comparisons, and split plot designs.

Shorack and Wellner (1986, section 19.3) develops the theory of randomly trimmed (and Winsorized) means and uses empirical process theory in the derivations. A key concept in empirical process theory is the quantile function

$$Q(t) = \inf\{x : F(x) \geq t\}. \tag{6}$$

Note that $Q(t)$ is the left continuous inverse of F and if F is strictly increasing and continuous, then F has an inverse F^{-1} and $F^{-1}(t) = Q(t)$. The following conditions on the cdf are used.

Regularity Conditions. R1) Let $X_1, \ldots, X_n$ be iid with cdf F, and let L_n and U_n be integer valued random variables such that $0 \leq L_n < U_n \leq n$.

R2) Let $a = Q(\alpha)$ and $b = Q(\beta)$.

R3) Suppose Q is continuous at α and β and that

R4)

$$\frac{L_n}{n} = \alpha + O_P(n^{-1/2}),$$

and R5)

$$\frac{U_n}{n} = \beta + O_P(n^{-1/2}).$$

Thus $\sqrt{n}((U_n/n) - \beta)$ is "tight" or bounded in probability. Note that R2) and R3) imply that $F(a) = \alpha$ and $F(b) = \beta$.

Under these conditions with $L_n = l_n$ and $U_n = u_n$,

$$\sqrt{n}(T_n - \mu_T(a,b)) \rightarrow N[0, \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}]. \tag{7}$$

The asymptotic variance can be consistently estimated with the scaled sample Winsorized variance

$$V_W(n) = \frac{(1/n)[l_n X_{(l_n+1)}^2 + \sum_{i=l_n+1}^{u_n} X_{(i)}^2 + (n-u_n)X_{(u_n)}^2] - [W_n(l_n,u_n)]^2}{[(u_n - l_n)/n]^2}. \tag{8}$$

This result is a special case of the following two lemmas. We will say

$$X_n \stackrel{a}{=} Y_n$$

if $X_n - Y_n \stackrel{P}{\rightarrow} 0$ as $n \rightarrow \infty$. Note that the trimmed mean $T_n = R_n(l_n, u_n)$, and

$$\frac{l_n}{n} - \alpha = o_P(n^{-1/2}), \text{ and } \frac{u_n}{n} - \beta = o_P(n^{-1/2}).$$

Hence $\sqrt{n}((l_n/n) - \alpha)$ converges to zero in probability.

Lemma 1. Shorack and Wellner (1986, p. 681). Assume that the regularity conditions hold. Then

$$S_n = \sqrt{n}[\frac{1}{n}\sum_{i=L_n+1}^{U_n} X_{(i)} - \int_{L_n/n}^{U_n/n} Q(t)dt] \stackrel{d}{\rightarrow} N[0, \sigma_W^2(a,b)].$$

Lemma 2. Shorack and Wellner (1986, p. 678-679). Assume that the regularity conditions hold. Then

$$\sqrt{n}(R_n - \mu_T) \stackrel{a}{=} \frac{1}{\beta-\alpha}[S_n + (\mu_T - a)\sqrt{n}(\frac{L_n}{n} - \alpha) + (b - \mu_T)\sqrt{n}(\frac{U_n}{n} - \beta)]. \tag{9}$$

A consequence of these two lemmas is that R_n and T_n will be asymptotically equivalent if

$$\frac{L_n}{n} - \alpha = o_P(n^{-1/2}) \text{ and } \frac{U_n}{n} - \beta = o_P(n^{-1/2}).$$

R_n^* is a very robust estimator that has simple asymptotic theory under symmetry. The key idea is that the choice $U_n = n - L_n$ causes the last two terms in lemma 2 to sum to zero. Shorack and Wellner (1986, p. 282-283) show that the regularity conditions R4) and R5) hold for the metrically trimmed mean provided that

$$\sqrt{n}(MED(n) - MED(X)) = O_P(1) \tag{10}$$

and

$$\sqrt{n}(MAD(n) - MAD(X)) = O_P(1). \tag{11}$$

This result is used to show that the metrically trimmed mean is asymptotically equivalent to a sum of two Gaussian random variables under symmetry. Assume R6) $P(U_n > L_n) \to 1$. That is, $R_n^* \neq MED(n)$, with arbitrarily high probability if n is large enough.

Corollary. Let F be symmetric. Assume regularity conditions R1), R2), R3), and R6) hold. Then

$$\sqrt{n}[R_n^* - \mu_T(a,b)] \to N(0, \frac{\sigma_W^2(a,b)}{(\beta - \alpha)^2}).$$

Proof. Let $a = MED(X) - kMAD(X)$ and let $b = MED(X) + kMAD(X)$. Then the result follows from Lemma 2 provided that R4) and R5) hold, that is if equations (10) and (11) hold, but the left hand sides of these equations are asymptotically normal. See Falk (1997) or Hall and Welsh (1985). QED

As stated in Shorack and Wellner (1986, p. 680), a natural estimator for the asymptotic variance is the scaled sample Winsorized variance

$$V_A(n) = \frac{(1/n)[L_n X_{(L_n+1)}^2 + \sum_{i=L_n+1}^{U_n} X_{(i)}^2 + (n - U_n)X_{(U_n)}^2] - [W_n(L_n, U_n)]^2}{[(U_n - L_n)/n]^2} \tag{12}$$

since

$$V_A(n) \xrightarrow{P} \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}$$

if the regularity condition R3) holds and if

$$\frac{L_n}{n} \xrightarrow{P} \alpha \text{ and } \frac{U_n}{n} \xrightarrow{P} \beta.$$

Thus if the errors are symmetric, any procedure that uses R_n^* and V_A is asymptotically equivalent to a procedure that uses T_n and V_W.

The following lemma gives the asymptotic theory for the two stage trimmed mean and is immediate.

Lemma 3. Assume that $MED(n)-kMAD(n) \to a$ and $MED(n)+kMAD(n) \to b$. Let $t = 100\max(F(a-), 1-F(b))$. Assume that $0 < t < 49$, and that t is not integer valued. Let $J \in \{1, ..., 49\}$ be the smallest integer greater than t. Then $T_{2,n}$ is asymptotically equivalent to the J % trimmed mean.

To find the asymptotic efficiency of these estimators, formulas for the asymptotic variance

$$AV = \frac{\sigma_W^2(a,b)}{(\beta-\alpha)^2}$$

are useful. Let $b = \mu + kMAD(X)$. Suppose that the error distribution is Gaussian. Let $\Phi(x)$ be cdf and let $\phi(x)$ be the density of the standard normal. Then

$$AV = (\frac{1-\frac{2z\phi(z)}{2\Phi(z)-1}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2})\sigma^2 \tag{13}$$

where $\alpha = \Phi(-z)$, and $z = k\Phi^{-1}(0.75)$. For the two stage estimator, round 100α up to the nearest integer J. Then use $\alpha_J = J/100$ and $z_J = -\Phi^{-1}(\alpha_J)$ in equation (13). Then

the asymptotic efficiency (AE) with respect to the mean is $AE = 1/AV$. If $k = 6$, then $AE(R_n^*, \bar{x}) \approx 1.0$. Since $J = 1$, $AE(T_{2,n}, \bar{x}) \approx 0.996$.

Assume that the errors are double exponential DE(0,1). Then $AV =$

$$\frac{\frac{2-(z^2+2z+2)e^{-z}}{1-e^{-z}}}{1-2\alpha} + \frac{2\alpha z^2}{(1-2\alpha)^2} \tag{14}$$

where $z = k\log(2)$ and $\alpha = 0.5\exp(-z)$. For the two stage estimator, compute α_J as above and let $z_J = -\log(2\alpha_J)$. Then the asymptotic efficiency (AE) with respect to the mean is $AE = 2/AV$. If $k = 6$, then $AE(R_n^*, \bar{x}) \approx 1.054$. Since $J = 1$, $AE(T_{2,n}, \bar{x}) \approx 1.065$.

The results from a small simulation are presented in table 1. For each sample size n, 500 samples were generated. The sample mean $\bar{x}$, sample median, 1% trimmed mean, R_n^*, and $T_{2,n}$ were computed. The latter two estimators were computed using the trimming parameter $k = 5$. Next the sample variance $S^2(T)$ of the 500 values $T_1, ..., T_{500}$ was computed where T is one of the five estimators. The value in the table is $nS^2(T)$. These numbers estimate the asymptotic variance, which is reported in the rows $n = \infty$. The simulations were performed for normal and double exponential data, and the simulated values are close to the theoretical values.

Table 1: Simulated Variance, 500 Runs, k = 5

F	n	$\bar{x}$	MED	1% TM	R_n^*	$T_{2,n}^*$
N(0,1)	10	1.116	1.454	1.116	1.166	1.166
N(0,1)	50	0.973	1.556	0.973	0.974	0.974
N(0,1)	100	1.040	1.625	1.048	1.044	1.044
N(0,1)	1000	1.006	1.558	1.008	1.008	1.010
N(0,1)	∞	1.000	1.571	1.004	1.000	1.004
DE(0,1)	10	1.919	1.403	1.919	1.646	1.646
DE(0,1)	50	2.003	1.400	2.003	1.777	1.777
DE(0,1)	100	1.894	0.979	1.766	1.595	1.595
DE(0,1)	1000	2.080	1.056	1.977	1.904	1.886
DE(0,1)	∞	2.000	1.000	1.878	1.834	1.804

3 References

Bickel, P.J., 1965. On some robust estimates of location, Ann. Math. Stat. 36, 847-858.

Cramer, H., 1946. Mathematical Methods of Statistics, Princeton University Press, Princeton, NJ.

Davies, L., and Gather, U., 1993. The identification of multiple outliers, J. Amer. Statist. Assoc. 88, 782-792.

Falk, M., 1997. Asymptotic independence of median and MAD, Statist. Prob. Lett. 34, 341-345.

Hahn, G.H., Mason, D.M., and Weiner, D.C. (ed.s), 1991. Sums, Trimmed Sums, and Extremes, Birkhauser, Boston.

Hall, P., and Welsh, A.H., 1985. Limit theorems for the median deviation, Ann. Inst. Statist. Math. Part A, 37, 27-36.

Hampel, F.R., 1985. The breakdown points of the mean combined with some rejection rules, Technom. 27, 95-107.

Kim, S., 1992. The metrically trimmed mean as a robust estimator of location, Ann. Statist. 20, 1534-1547.

Shorack, G.R., 1974. Random means, Ann. Statist. 1, 661-675.

Shorack, G.R., and Wellner, J.A., 1986. Empirical Processes With Applications to Statistics, Wiley, NY.

Simonoff, J.S., 1987. Outlier Detection and robust estimation of scale. J. Statist. Comp. Sim. 27, 79-92.

Stigler, S.M., 1973. The asymptotic distribution of the trimmed mean, Ann. Math. Statist. 1, 472-477.

Wilcox, R.R., 1997. Introduction to Robust Estimation and Testing, Academic Press, San Diego, CA.