COMPARISON OF ESTIMATORS OF THE MEAN WHICH ADJUST FOR LARGE OBSERVATIONS

Lawrence R. Ernst, U.S. Bureau of the Census

1. INTRODUCTION

A question that arises frequently in survey sampling is what to do with very large observations which occur in the sample. In many situations alternative estimators of the mean are used which reduce the effect of these large values and which often have a lower mean square error than the ordinary sample mean.

In sections 3 and 4 of this paper we compare seven types of such estimators. Four of them adjust for sample values greater than or equal to some predetermined cutoff value t. The first of these, $\bar{X}_t^{(1)}$, results from substituting t itself for each of the large values. The second estimator, $\bar{X}_{t,W}^{(2)}$, is obtained by giving all of these large values a reduced weight W. $\bar{X}_t^{(3)}$ is simply the mean of all values below the cutoff. $\bar{X}_t^{(4)}$ is obtained by first substituting a new sample value below the cutoff for each large value, and then taking the ordinary sample mean of the modified sample.

The final three estimators adjust for the r largest sample values, where r is a predetermined positive integer. $\bar{X}_r^{(5)}$, which is known as the r-th Winsorized mean, results from substituting the (r + 1)-st largest value for each of the r largest values. $\bar{X}_r^{(6)}$, the r-th trimmed mean, is the ordinary sample mean of the values remaining after the r largest are discarded. Finally, $\bar{X}_{r,W}^{(7)}$ is obtained by giving all of the r largest values a reduced weight W.

Several of these estimators were studied by Searls (1963) who compared the efficiency of each of them with that of the ordinary sample mean. One result that he obtained was that under quite general conditions there always exists a value τ which minimizes $MSE(\bar{X}_t^{(1)})$ and that $\bar{X}_\tau^{(1)}$ is more efficient than the ordinary sample mean.

In sections 3 and 4 we show that in a sense $\bar{X}_t^{(1)}$ is the best among the seven estimators by proving that $\bar{X}_\tau^{(1)}$ is, for the optimal τ, at least as efficient as any of the other six estimators for any choice of t, W, and r.

In section 5 we illustrate the results of sections 3 and 4 using the exponential distribution.

2. NOTATION AND TERMINOLOGY

The underlying population distribution X will be assumed continuous with finite mean and variance; its probability density function (pdf) will be positive on some subinterval of $[0, \infty)$ with left endpoint a. Let $\mu = E(X)$, while μ, σ_t^2 are, for $t \geq 0$, respectively the mean and variance of X truncated on the right at t.

We assume simple random sampling with replacement with sample size n. Let $X_1, X_2, \ldots, X_n$ denote the unordered variates and let m_t denote the number of them with values at least t. Let $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ denote the ordered variates. Furthermore, in order to define $\bar{X}_t^{(4)}$ sampling will be continued untill n sample observations below t are obtained with $X_{1,t}, X_{2,t}, \ldots, X_{n,t}$ taken to be the unordered variates corresponding to the n sample observations below t in this extended sample. Also in the case of $\bar{X}_t^{(3)}$ with $m_t = n$, sampling is continued until one observation below t, namely $X_{1,t}$, is obtained.

For any function f of X let $\overline{f(X)} = \dfrac{\sum_{i=1}^{n} f(X_i)}{n}$.

We next define the estimators $\bar{X}_t^{(1)}$, $\bar{X}_{t,W}^{(2)}$, $\bar{X}_t^{(3)}$, $\bar{X}_t^{(4)}$, $\bar{X}_r^{(5)}$, $\bar{X}_r^{(6)}$, $\bar{X}_{r,W}^{(7)}$ for $t \geq 0$ ($t > a$ in the case of $\bar{X}_t^{(3)}$ and $\bar{X}_t^{(4)}$), $W \in [0, 1)$, $r \in \{1, 2, \ldots, n-1\}$. We let

$$\bar{X}_t^{(1)} = \frac{\sum_{i=1}^{n-m_t} X_{(i)} + m_t t}{n} = \overline{f_t^{(1)}(X)},$$

where

$$f_t^{(1)}(x) = \begin{cases} x & \text{if } x < t \\ t & \text{if } x \geq t \end{cases};$$

$$\bar{X}_{t,W}^{(2)} = \frac{\sum_{i=1}^{n-m_t} X_{(i)} + W \sum_{i=n-m_t+1}^{n} X_{(i)}}{n} = \overline{f_t^{(2)}(X)},$$

where

$$f_{t,W}^{(2)}(x) = \begin{cases} x & \text{if } x < t \\ Wx & \text{if } x \geq t \end{cases};$$

$$\bar{X}_t^{(3)} = \begin{cases} \dfrac{\sum_{i=1}^{n-m_t} X_{(i)}}{n-m_t} & \text{if } m_t < n \\ X_{1,t} & \text{if } m_t = n \end{cases};$$

(We note that the definition of $\bar{X}_t^{(3)} = X_{1,t}$ in the case when $m_t = n$, is given to define $\bar{X}_t^{(3)}$ in what would otherwise be an undefined situation.)

$$\bar{X}_t^{(4)} = \frac{\sum_{i=1}^{n} X_{i,t}}{n};$$

$$\bar{X}_r^{(5)} = \frac{\sum_{i=1}^{n-r} X_{(i)} + rX_{(n-r)}}{n};$$

$$\bar{X}_r^{(6)} = \frac{\sum_{i=1}^{n-r} X_{(i)}}{n-r} ;$$

$$\bar{X}_{r,W}^{(7)} = \frac{\sum_{i=1}^{n-r} X_{(i)} + W \sum_{i=n-r+1}^{n} X_{(i)}}{n} .$$

Finally, we note that in the proof of theorems 4.1-4.3 certain expressions will, in special cases, be of the form $\sum_{i=j}^{k} a_i$ with $j > k$. In such situations we define $\sum_{i=j}^{k} a_i = 0$.

3. COMPARISON OF $\bar{X}_t^{(1)}$, $\bar{X}_{t,W}^{(2)}$, $\bar{X}_t^{(3)}$, AND $\bar{X}_t^{(4)}$

We proceed to establish that $\bar{X}_\tau^{(1)}$ is for the optimal τ at least as efficient as $\bar{X}_{t,W}^{(2)}$, $\bar{X}_t^{(3)}$, and $\bar{X}_t^{(4)}$ for any t and W.

Lemma 3.1:

$$\{E(\bar{X}_t^{(1)}):\ t \in [0, \infty)\}$$

$$= \{E[f_t^{(1)}(X)]:\ t \in [0,\infty)\} \supset [0,\mu) .$$

Proof: The first relation is obvious, while the second follows upon noting that $f_t^{(1)}(X)$ is a nondecreasing continuous function of t, $E[f_0^{(1)}(X)] = 0$ and $\lim_{t\to\infty} E[f_t^{(1)}(X)] = \mu$.

Lemma 3.2: If g is a measurable function such that $0 \le g(x) \le x$ for all $x \ge 0$ and $E[g(X)] < \mu$, then there exists $\tau > 0$ for which $MSE(\bar{X}_\tau^{(1)}) \le MSE[g(X)]$.

Proof: By lemma 3.1 there exists $\tau \ge 0$ with $E[f_\tau^{(1)}(X)] = E[g(X)]$. We observe that to obtain our result for this τ is equivalent to showing that $E[f_\tau^{(1)}(X)]^2 \le E[g(X)]^2$, which we proceed to do.

For any set S define K_S, the indicator function of S by $K_S(x) = 1$ if $x \in S$, $K_S(x) = 0$ if $x \notin S$. Let

$$A = \{x:\ f_\tau^{(1)}(x) > g(x)\},$$

$$B = \{x:\ f_\tau^{(1)}(x) < g(x)\}.$$

(We note that we may assume that neither A nor B is empty, since otherwise $f_\tau^{(1)}(x) = g(x)$ almost everywhere and, consequently, $E[f_\tau^{(1)}(X)]^2 = E[g(X)]^2$.) Then

$$E[f_\tau^{(1)}(X)]^2 \le E[g(X)]^2$$

$$\leftrightarrow E[f_\tau^{(1)}(X)K_A]^2 + E[f_\tau^{(1)}(X)K_B]^2 \le E[g(X)K_A]^2 + E[g(X)K_B]^2$$

$$\leftrightarrow E[([f_\tau^{(1)}(X)]^2 - [g(X)]^2)K_A] \le E[([g(X)]^2 - [f_\tau^{(1)}(X)]^2)K_B] \qquad (3.1)$$

Furthermore, by the definitions of A and B we have

$$E[([f_\tau^{(1)}(X)]^2 - [g(X)]^2)K_A] = E([f_\tau^{(1)}(X) + g(X)][f_\tau^{(1)}(X) - g(X)]K_A)$$
$$\le 2\sup\{f_\tau^{(1)}(x):\ x \in A\}E([f_\tau^{(1)}(X) - g(X)]K_A), \qquad (3.2)$$

and

$$E[([g(X)]^2 - [f_\tau^{(1)}(X)]^2)K_B] = E([g(X) + f_\tau^{(1)}(X)][g(X) - f_\tau^{(1)}(X)]K_B)$$
$$\ge 2\inf\{f_\tau^{(1)}(x):\ x \in B\}E([g(X) - f_\tau^{(1)}(X)]K_B) . \qquad (3.3)$$

Also, since $f_\tau^{(1)}(x) = x \ge g(x)$ if $x < \tau$, and $f_\tau^{(1)}(x) = \tau$ if $x \ge \tau$, it follows that

$$\sup\{f_\tau^{(1)}(x):\ x \in A\} \le \tau = \inf\{f_\tau^{(1)}(x):\ x \in B\} \qquad (3.4)$$

Finally, we note that

$$E([f_\tau^{(1)}(X) - g(X)]K_A) = E([g(X) - f_\tau^{(1)}(X)]K_B) \tag{3.5}$$

since $E[f_\tau^{(1)}(X)] = E[g(X)]$, and then combine (3.1) - (3.5) to complete the proof.

Theorem 3.1: For any t, W there exists τ for which

$$MSE(\bar{X}_\tau^{(1)}) \leq MSE(\bar{X}_{t,W}^{(2)})$$

Proof: This follows immediately from lemma 3.2 with $g = f_{t,W}^{(2)}$.

Remark: Since there always exists τ which minimizes the mean square error of $\bar{X}_t^{(1)}$ (Searls 1966), theorem 3.1 can be restated as follows: There exists τ such that

$$MSE(X_\tau^{(1)}) \leq MSE(\bar{X}_{t,W}^{(2)})$$

for all t,W. All the other theorems in this paper can be similarly restated.

Theorem 3.2: For any t there exists τ for which

$$MSE(\bar{X}_\tau^{(1)}) \leq MSE(\bar{X}_t^{(3)})$$

and

$$MSE(\bar{X}_\tau^{(1)}) \leq MSE(\bar{X}_t^{(4)}) .$$

Proof: Let

$$c_t(X) = \begin{cases} x & \text{if } 0 \leq x < t , \\ \mu_t & \text{if } x \geq t . \end{cases}$$

Then, by lemma 3.2 there exists τ satisfying $MSE(\bar{X}_\tau^{(1)}) \leq MSE[\overline{c_t(X)}]$. Furthermore, clearly $E[\overline{c_t(X)}|m_t] = E(\bar{X}_t^{(3)}|m_t) = E(\bar{X}_t^{(4)}|m_t) = \mu_t$. Consequently, to complete the proof we need only to show that

$$Var[\overline{c_t(X)}] \leq Var(\bar{X}_t^{(3)}) \tag{3.6}$$

and

$$Var[\overline{c_t(X)}] \leq Var(\bar{X}_t^{(4)}) . \tag{3.7}$$

To prove (3.6) we observe that

$$Var[\overline{c_t(X)}] = E(Var[\overline{c_t(X)}|m_t]) ,$$

$$Var(\bar{X}_t^{(3)}) = E[Var(\bar{X}_t^{(3)}|m_t)] ,$$

$$Var[\overline{c_t(X)}|m_t] = 0 \leq Var[\bar{X}_t^{(3)}|m_t] \quad \text{if } m_t = n ,$$

and

$$Var[\overline{c_t(X)}|m_t] = \frac{(n-m_t)\sigma_\tau^2}{n^2} \leq \frac{\sigma_t^2}{n-m_t}$$

$$= Var[\bar{X}_t^{(3)}|m_t] \quad \text{if } m_t < n .$$

To obtain (3.7) we simply note that

$$Var[\overline{c_t(X)}] = E(Var[\overline{c_t(X)}|m_t]) .$$

$$= \frac{E(n-m_t)\sigma_t^2}{n^2} \leq \frac{\sigma_t^2}{n} = Var(\bar{X}_t^{(4)}) .$$

4. COMPARISON OF $\bar{X}_t^{(1)}$, $\bar{X}_r^{(5)}$, $\bar{X}_r^{(6)}$, AND $\bar{X}_{r,W}^{(7)}$

We proceed to establish that $\bar{X}_\tau^{(1)}$ is for the optimal τ at least as efficient as $\bar{X}_r^{(5)}$, $\bar{X}_r^{(6)}$, and $\bar{X}_{r,W}^{(7)}$ for any r and W.

Lemma 4.1: If Y, Z are functions of X_1, ..., X_n with finite first and second moments, $E(Y) = E(Z)$ and

$$k_t = \min\{\ell : E(Z|m_t = \ell) \geq E(Y|m_t = \ell)\} ,$$

then $MSE(Y) \leq MSE(Z)$ provided the following hold for each $t \geq 0$:

(a) $E(Y|m_t = \ell)$ and $E(Z|m_t = \ell)$ are nondecreasing functions of ℓ;
(b) $E(Z|m_t = \ell) \geq E(Y|m_t = \ell)$ if $\ell \geq k_t$;
(c) $Cov(Z + Y, Z - Y|m_t = \ell) \geq 0$ if $\ell \geq k_t$.

and

(d) $E[Y^2 - Z^2)|m_t = \ell] < E[(Y + Z)|m_t = k_t]E[(Y - Z)|m_t = \ell]$ if $\ell < k_t$.

Proof: From the relation $E(Y) = E(Z)$ it follows that

$$\mathrm{MSE}(Y) \leq \mathrm{MSE}(Z) \leftrightarrow E(Y^2) \leq E(Z^2)$$

$$\leftrightarrow E[(Y^2 - Z^2)|m_t < k_t]\Pr(m_t < k_t) \leq E[(Z^2 - Y^2)|m_t \geq k_t]\Pr(m_t \geq k_t) .$$

To obtain this last inequality we first note that by (d),

$$E[(Y^2 - Z^2)|m_t < k_t] \leq E[(Y + Z)|m_t = k_t]E[(Y - Z)|m_t < k_t] . \tag{4.1}$$

Furthermore, if $\ell \geq k_t$ then by (c), (a), and (b)

$$E[(Z^2 - Y^2)|m_t = \ell] \geq E[(Z + Y)|m_t = k_t]E[(Z - Y)|m_t = \ell] ,$$

and hence

$$E[(Z^2 - Y^2)|m_t \geq k_t] \geq E[(Z + Y)|m_t = k_t]E[(Z - Y)|m_t \geq k_t] . \tag{4.2}$$

Finally, we combine (4.1) and (4.2) with the relation

$$E[(Y - Z)|m_t < k_t]\Pr(m_t < k_t) = E[(Z - Y)|m_t \geq k_t]\Pr(m_t \geq k_t) ,$$

which follows since $E(Y) = E(Z)$.

We next note the following relations for use in the proof of theorems 4.1 - 4.3:

$$\mathrm{Cov}\,(X_i, X_j|m_t = \ell) \geq 0 \quad \text{for } i, j=1,\dots, n. \tag{4.3}$$

$$(X_{(i)}|m_t = \ell) \text{ and } (X_{(j)}|m_t = \ell) \text{ are independent if } i \leq n - \ell < j. \tag{4.4}$$

$$E(\sum_{i=1}^{n-r} X_i|m_t = \ell) \leq (n-r)\mu_t \text{ if } r \geq \ell. \tag{4.5}$$

To establish (4.3) we observe that in case $i \leq n-\ell$ and $j \leq n-\ell$, then X_i and X_j are order statistics from the distribution of X truncated on the right at t, and consequently (4.3) follows from the fact that under very general conditions the covariance of two order statistics is nonnegative (David 1970, Ex. 3.1.11). Similarly (4.3) holds if $i > n-\ell$ and $j > n-\ell$. On the other hand, if exactly one of i,j exceeds $n-\ell$, then (4.3) and (4.4) both follow since X_i and X_j are then order statistics from independent distributions, namely X truncated on the right at t and X truncated on the left at t.

To obtain (4.5) we simply note that if $r \geq \ell$, then

$$\frac{E(\sum_{i=1}^{n-r} X_i|m_t = \ell)}{n-r} \leq \frac{E(\sum_{i=1}^{n-\ell} X_i|m_t = \ell)}{n-\ell} = \mu_t$$

Theorem 4.1: For any r there exists τ for which $\mathrm{MSE}(X_\tau^{(1)}) < \mathrm{MSE}(\bar{X}_r^{(5)})$.

Proof: By lemma 3.1 there exists τ satisfying $E(\bar{X}_\tau^{(1)}) = E(\bar{X}_r^{(5)})$. We will prove the theorem by showing that conditions (a) - (d) of lemma 4.1 hold with $Y = \bar{X}_\tau^{(1)}$ and $Z = X_r^{(5)}$.

Clearly (a) and (b) both hold and $k_\tau = r + 1$. To obtain (c) we apply (4.3) after first noting that if $\ell \geq k_\tau = r + 1$, then

$$[(\bar{X}_r^{(5)} + \bar{X}_\tau^{(1)})|m_\tau = \ell] = \frac{1}{n}(2\sum_{i=1}^{n-\ell} X_{(i)} + \sum_{i=n-\ell+1}^{n-r} X_{(i)} + rX_{(n-r)} + \ell\tau), \tag{4.6}$$

and

$$[(\bar{X}_r^{(5)} - \bar{X}_\tau^{(1)})|m_\tau = \ell]$$

$$= \frac{1}{n}\left(\sum_{i=n-\ell+1}^{n-r} X_{(i)} + rX_{(n-r)} - \ell\tau\right) .$$

To prove (d) we observe that if $\ell < k_\tau = r + 1$, then

$$[(\bar{X}_\tau^{(1)} + \bar{X}_r^{(5)})|m_\tau = \ell]$$

$$= \frac{1}{n}\left(2\sum_{i=1}^{n-r} X_{(i)} + \sum_{i=n-r+1}^{n-\ell} X_{(i)} + \ell\tau + rX_{(n-r)}\right)$$

$$\leq \frac{1}{n}\left(2\sum_{i=1}^{n-r} X_{(i)} + 2r\tau\right)$$

and

$$[(\bar{X}_\tau^{(1)} - \bar{X}_r^{(5)})|m_\tau = \ell]$$

$$= \frac{1}{n}\left(\sum_{i=n-r+1}^{n-\ell} X_{(i)} + \ell\tau - rX_{(n-r)}\right) ,$$

which together with (4.3) and (4.5) imply that

$$E([(\bar{X}_\tau^{(1)})^2 - (\bar{X}_r^{(5)})^2]|m_\tau = \ell)$$

$$\leq \frac{1}{n}[2(n-r)\mu_\tau + 2r\tau]E[(\bar{X}_\tau^{(1)} - \bar{X}_r^{(5)})|m_\tau = \ell] .$$

Furthermore, from (4.6) with $\ell = k_\tau = r+1$ we obtain

$$E[(\bar{X}_\tau^{(1)} + \bar{X}_r^{(5)})|m_\tau = k_\tau]$$

$$\geq \frac{1}{n}[2(n-r-1)\mu_\tau + 2(r+1)\tau] \geq \frac{1}{n}[2(n-r)\mu_\tau + 2r\tau] .$$

<u>Theorem 4.2:</u> For any r, there exists τ for which $MSE(\bar{X}_\tau^{(1)}) \leq MSE(\bar{X}_r^{(6)})$.

<u>Theorem 4.3:</u> For any r, W there exists τ for which $MSE(\bar{X}_\tau^{(1)}) \leq MSE(\bar{X}_{r,W}^{(7)})$.

The proofs of theorems 4.2 and 4.3 have been omitted due to lack of space. These proofs are available from the author.

5. AN EXAMPLE

The following tables illustrate the results of the previous sections for the exponential distribution with sample sizes of 10, 100, and 1,000. The exponential distribution was chosen for reasons of computational simplicity and because it is a positively skewed distribution. As has been proven $\bar{X}_t^{(1)}$ attains the highest efficiency among the seven estimators. In this example $\bar{X}_t^{(1)}$, $\bar{X}_{t,W}^{(2)}$, $\bar{X}_t^{(3)}$, $\bar{X}_t^{(4)}$, and $\bar{X}_{r,W}^{(7)}$ are all, for the optimal choice of parameters, more efficient than the ordinary sample mean, $\bar{X}$. This is true for these five estimators for all continuous random variables X which take only nonnegative values as was proven by Searls (1963). This result does not always hold for $\bar{X}_r^{(5)}$ and $\bar{X}_r^{(6)}$. In the particular case of the exponential distribution $MSE(\bar{X}_r^{(5)})$ and $MSE(\bar{X}_r^{(6)})$ increase as r increases and

$$MSE(\bar{X}_1^{(6)}) > MSE(\bar{X}_1^{(5)}) = Var(\bar{X}) .$$

It is also interesting to note that for the exponential distribution $MSE(\bar{X}_{r,W}^{(7)})$ decreases as r increases, for optimal W, and hence is minimal if r = n-1. However, if the restriction that r < n is removed, then $\bar{X}_{r,W}^{(7)}$ will attain its maximal efficiency when r = n, W = n/(n+1).

1. Parameter Values Which Yield Maximal Relative Efficiency of Estimators With Respect to $\overline{X}$ for Samples From the Exponential Distribution With Mean μ

Estimator	Sample Size 10	100	1000
$\overline{X}_t^{(1)}$	2.10μ	3.53μ	5.32μ
$\overline{X}_{t,W}^{(2)}$	2.62μ, 0.517	4.20μ, 0.644	6.01μ, 0.729
$\overline{X}_t^{(3)}$	3.40μ	5.48μ	7.68μ
$\overline{X}_t^{(4)}$	3.32μ	5.44μ	7.66μ
$\overline{X}_r^{(5)}$	1	1	1
$\overline{X}_r^{(6)}$	1	1	1
$\overline{X}_{r,W}^{(7)}$	9, 0.908	99, 0.990	999, 0.999

2. Relative Efficiency of Estimators with Respect to $\overline{X}$ for Samples from the Exponential Distribution With Mean μ When Optimal Parameter Values are Used

Estimator	Sample Size 10	100	1000
$\overline{X}_t^{(1)}$	1.6112	1.1401	1.0292
$\overline{X}_{t,W}^{(2)}$	1.5466	1.1298	1.0275
$\overline{X}_t^{(3)}$	1.3395	1.0757	1.0145
$\overline{X}_t^{(4)}$	1.3796	1.0796	1.0150
$\overline{X}_r^{(5)}$	1.0000	1.0000	1.0000
$\overline{X}_r^{(6)}$	0.8605	0.9009	0.9701
$\overline{X}_{r,W}^{(7)}$	1.0999	1.0100	1.0010

REFERENCES

Bershad, Max A. (1960), "Some Observations on Outliers," unpublished memorandum, Statistical Research Division, U.S. Bureau of the Census.

David, H. A. (1970), _Order Statistics_, New York: John Wiley and Sons.

Fuller, Wayne A. (1970) "Simple Estimators for the Mean of Skewed Populations," unpublished manuscript prepared for U.S. Bureau of the Census.

Searls, Donald T. (1963), "On the 'Large' Observation Problem," unpublished Ph.D. thesis, North Carolina State University.

—— (1966), "An Estimator for a Population Mean Which Reduces the Effect of Large True Observations," _Journal of the American Statistical Association_, 61, 1200-1204.