AN ESTIMATOR OF THE POPULATION VARIANCE IN THE PRESENCE OF LARGE TRUE OBSERVATIONS

BY

V.P. OJHA[1] AND S.R. SRIVASTAVA[2]

(Received : May, 1977)

INTRODUCTION

Searls [5] proposed an estimator $\bar{y}_t$ for the population mean μ to reduce the effect of a few large observations, particularly in the samples of small sizes, which are not the outliers but are rather true observations. The effect of these offending observation is that the usual estimator $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} \bar{y}_i$ of the mean μ, though unbiased, gives an estimate which is considerably higher than the true value of the parameter μ. *Tukey* and McLaughlin [6], Grow [1], Dixon [2], Searls [3] etc. have suggested various techniques for estimation of mean when sample contains extreme or large observations. Searls [3] proposed an estimator $\bar{y}_t$ for mean which is constructed by replacing all sample values larger than a predetermined cutoff point t by the value of t itself. The cutoff point need not be any observed value, say $y_{(r)}$ or, one depending on the parameter to be estimated. It is assumed that the sampler has only some rough indication of the possible range in which the data may occur. It helps him to choose a cutoff point t and to construct the estimators for the parameters of the population. The object of the present paper is to show that the proposed estimator S_t^2 for variance σ^2 is more efficient then S^2 for a wide range of values of t.

[1] University of Gorakhpur, Gorakhpur.

[2] Banaras Hindu University, Varanasi.

2. The Estimator S_t^2

Let $y_1, y_2, \ldots y_n$, be the sample observations from a distribution with *p.d.f.* $f(y)$ and *c.d.f.* $F(y)$. Let $\bar{y} = \frac{1}{n} \sum_{j=1}^{n} y_j$ be the sample mean and $(n-1) S^2 = \sum_{j=1}^{n} (y_j - \bar{y})^2$. Then, we define

$$(n-1) S_t^2 = \sum_{j=1}^{r} y_j^2 + (n-r) t^2 - n \bar{y}_t^2 , \tag{2.1}$$

where r is the number of observations with values less than or equal to t,

$$\bar{y}_t = \frac{1}{n} \left[\sum_{j=1}^{r} y_j + (n-r) t \right] \tag{2.2}$$

and t is the cut off point where the distribution $F(y)$ gets truncated on the right. Let μ_t, σ_t^2 $\alpha_{3,\,t}$ and $\alpha_{4,\,t}$ be the mean, variance, third and fourth moments about the origin of this truncated distribution and $P[y_j < t] = F(t) = p$ and $q = 1 - p$. Then

$$\begin{aligned} E[\bar{y}_t] &= E\left[\frac{r}{n} \right] E\,[y_j \mid y_j < t] + t\, E\left[\frac{n-r}{n} \right] \\ &= p\, \mu_t + q\, t \end{aligned} \tag{2.3}$$

$$E\left[(n-1) S_t^2 \right] = E\left[\sum_{j=1}^{r} y_j^2 \right] + t^2 E\,[n-r] - n\, E\left[\bar{y}_t^2 \right] \tag{2.4}$$

$$E \left[\sum_{1}^{r} y_j^2 \right] = n\, p\, [\sigma_t^2 + \mu_t^2] \tag{2.5}$$

$$E\,[n-r] = nq \tag{2.6}$$

and

$$\begin{aligned} E\left[\bar{y}_t^2 \right] = \frac{1}{n} \Big[\Big\{ p \left(\sigma_t^2 + \mu_t^2 \right) + q\, t^2 \Big\} - \left(p\, \mu_t + q_t \right)^2 \Big] \\ + (p\, \mu_t + q_t)^2 \end{aligned} \tag{2.7}$$

Hence, from (2.3) to (2.7) we get

$$E\left[(n-1) S_t^2 \right] = (n-1) \left[p \left(\sigma_t^2 + \mu_t^2 \right) + q\, t^2 - (p\, \mu_t + q\, t)^2 \right]$$

or

$$E\left[\,S_t^2\,\right]=p\left(\,\sigma_t^2\,+\mu_t^2\,\right)+q\,t^2-(p\,\mu_t+q_t)^2$$

$$=p\left[\,\sigma_t^2\,+q\,(t-\mu_t)^2\,\right]$$

$$=\sigma^{*2}\text{ (say)} \tag{2.8}$$

3. Bias and Mean Square Error of S_t^2

$$\text{Bias}\left(\,S_t^2\,\right)=E\left[\,S_t^2\,\right]-\sigma^2$$

$$=\sigma^{*2}-\sigma^2$$

$$=-q[\sigma_t'^2+p\,(\mu_t'-t)^2+2p\,(t-\mu_t)\,(\mu_t'-t)] \tag{3.1}$$

where μ_t' and $\sigma_t'^2$ are the mean and variance of the left truncated distribution.

Let $\lambda=\dfrac{\sigma^{*2}}{\sigma^2}$. Then Bias

$\left(\,S_t^2\,\right)=(\lambda-1)\,\sigma^2<0$ [by virtue of (3.1)], so that $0<\lambda\leqslant1$,

The mean square error of S_t^2 can be written as

$$MSE\left(\,S_t^2\,\right)=E\left[\,S_t^4\,\right]-2\,\sigma^2\,E\left[\,S_t^2\,\right]+\sigma^4 \tag{3.2}$$

Now

$$E\left[\,S_t^4\,\right]=\frac{1}{(n-1)^2}\,E\left[\sum_{j=1}^{r}\,y_j^2\,+(n-r)\,t^2-n\,\bar{y}_t^2\,\right]^2$$

$$=\frac{1}{(n-1)^2}\left[\,E\left\{\left(\sum_{j=1}^{r}\,y_j^2\,\right)^2\right\}+t^4\,E\,(n-r)^2\right.$$

$$+n^2\,E\left(\,\bar{y}_t^2\,\right)^2+2\,t^2\,E\left\{(n-r)\sum_{j=1}^{r}\,y_j^2\,\right\}$$

$$-2\,nt^2\,E\left\{\,(n-r)\,\bar{y}_t^2\,\right\}-2n\,E\left\{\left(\sum_{j=1}^{r}\,y_j^2\,\right)\right.$$

$$\left.\left.\left(\,\bar{y}_t^2\,\right)\right\}\right] \tag{3.3}$$

$$E\left[\left(\sum_{j=1}^{r} y_j^2\right)\left(\bar{y}_t^2\right)\right]=\frac{1}{n}\Big[p\,\alpha_{4,t}+2(n-1)\,p\,\alpha_{3,t}\,(p\,\mu_t+q_t)$$
$$+(n-1)\,p^2\,(\alpha_t^2+\mu_t^2)\,(p\mu_t+qt)^2$$
$$+(n-1)\,p\,q_t^2\,(\sigma_t^2+\mu_t^2)\Big] \quad (3.4)$$

$$E\left[(n-r)\,\bar{y}_t^2\right]=\frac{1}{n}[(n-1)\,q\,\{p\,(\sigma_t^2+\mu_t^2)+q\,t^2\}$$
$$+2\,(n-1)\,q\,t\,(p\,\mu_t+q\,t)+(n-1)\,(n-2)\,q$$
$$(p\,\mu_t+qt)^2+q\,t^2] \quad (3.5)$$

$$E\left[(n-r)\sum_{j=1}^{r} y_j^2\right]=n\,(n-1)\,p\,q\,(\sigma_t^2+\mu_t^2) \quad (3.6)$$

$$E\,[(n-r)^2]=n\,(n-1)\,q^2+n\,q \quad (3.7)$$

$$E\left[\left(\sum_{j=1}^{r} y_j^2\right)^2\right]=n\,p\,\alpha_{4,t}+n\,(n-1)\,p^2\,(\sigma_t^2+\mu_t^2)^2 \quad (3.8)$$

and

$$E\left[\left(\bar{y}_t^2\right)^2\right.=\frac{1}{n^3}\,[p\,\alpha_{4,t}+4\,(n-1)\,p\,\alpha_{3,t}\,(p\mu_t+q\,t)$$
$$+6\,(n-1)\,(n-2)\,\{p\,(\sigma_t^2+\mu_t^2)+qt^2\}$$
$$(p\,\mu_t+qt)^2+n\,(n-1)\,(n-2)\,(n-3)$$
$$(p\,\mu_t+q\,t)^4+3\,(n-1)\,\{p\,(\sigma_t^2+\mu_t^2)$$
$$+q\,t^2\}^2+4\,(n-1)\,q\,t^3\,(p\,\mu_t+q_t)+q_t^4] \quad (3.9)$$

If we use the results (3.3) through (3.9) and simplify, we get

$$E\left[\left(S_t^2\right)^2\right]=\frac{1}{n^2\,(n-1)^2}\,[n\,(n-1)^2\,\{p\,\alpha_{4,t}+q\,t^4)$$
$$-4\,(p\,\alpha_{3,t}+q\,t^3)+6\,(p\,(\sigma_t^2+\mu_t^2)+q\,t^2)$$
$$(p\,\mu_t+q\,t)^2-3\,(p\,\mu_t+q\,t)^4+n\,(n-1)$$
$$(n^2-2n+3)\,\{p\,(\sigma_t^2+\mu_t^2)+q\,t^2$$
$$-(p\,\mu_t+q\,t)^2\}^2] \quad (3.10)$$

If μ_4^* and σ^{*2} are the fourth central moment and variance respectively of the population truncated on the right at the point t, then

$$\mu_4^*=(p\alpha_{4,t}+q_t^4)-4\,(p\,\alpha_{3,t}+q_t^3)\,(p\,\mu_t+q_t)$$
$$+6\,\{p(\sigma_t^2+\mu_t^2)+q_t^2\}\,(p\,\mu_t+q_t)^2-3\,(p\,\mu_t+q_t)^4$$
$$(3.11)$$

$$\sigma^{*2}=p\,(\sigma_t^2-\mu_t^2)+q\,t^2-(p\mu_t+q\,t)^2$$

$$=p\,\{\sigma_t^2+q\,(t-\mu_t)^2\} \qquad (3.12)$$

$$\therefore\quad E\left[\;S_t^4\;\right]\;=\;\frac{\mu_4^*}{n}+\frac{n^2-2n+3}{n\,(n-1)}\,(\sigma^{*2})^2 \qquad (3.13)$$

Hence

$$MSE\,(S_t^2)=\left[\frac{\mu_4^*}{n}+\frac{n^2-2n+3}{n\,(n-1)}\;\sigma^{*4}\right]-2\sigma^{*2}\,\sigma^2+\sigma^4$$

$$=\left[\frac{\mu_4^*}{n}+\frac{3-n}{n\,(n-1)}\;\alpha^{*4}\right]+(\sigma^{*2}-\sigma^2)^2 \qquad (3.14)$$

It may be noted that when t approaches the upper limit of the distribution,

$$p\longrightarrow 1,\qquad q\longrightarrow 0,\qquad \mu_4^*\longrightarrow\mu_4,\qquad \sigma^{*2}\longrightarrow\sigma^2$$

and $\qquad MSE\,(S_t^2)\longrightarrow \mathrm{Var}\,(S^2)$

4. The Relative Efficiency of S_t^2

The relative efficiency of S_t^2 with respect to S^2 may be defined by

$$REF\,(S_t^2)=\frac{\mathrm{Var}\,(S^2)}{MSE\,(S_t^2)}\times 100$$

$$=\frac{\dfrac{\beta_2}{n}+\dfrac{3-n}{n\,(n-1)}}{\left[\dfrac{\beta_2^*}{n}+\dfrac{3-n}{n\,(n-1)}\right]\lambda^2+(\lambda-1)^2}\times 100 \qquad (4.1)$$

where

$$\beta_2=\frac{\mu_4}{\sigma^4},\qquad \beta_2{}^*=\frac{\mu_4^*}{\sigma^{*4}}$$

The relative efficiency of S_t^2 is greater than unity if and only if the *r. h. s.* of (4.1) is greater than unity, that is,

$$\varphi(\lambda)=\left[\frac{\beta_2^*}{n}+\frac{3-n}{n\,(n-1)}+1\right]\lambda^2-2\lambda$$

$$-\left[\frac{\beta_2}{n}+\frac{3-n}{n\,(n-1)}-1\right]\leqslant 0 \tag{4.2}$$

which may be rearranged as

$$\lambda^2\,\beta_2^* \leq \beta_2-n\,(1-\lambda)^2-\frac{n-3}{n-1}\,(1-\lambda^2) \tag{4.3}$$

Since t and σ^{*2} are connected by equation (3.12); therefore, a region of t for which $MSE\left(S_t^2\right) <$ Var (S^2) may be found out by solving (4.2) as a quadratic equation (φ $(\lambda)=0$, in λ and consider only those values of λ which are in the interval (0, 1).

The optimum value of t can be obtained by solving the equation

$$\frac{d}{dt}\,MSE\,(S_t^2)=0. \tag{4.4}$$

5. An Example

In this example, we have chosen the underlying distribution to be exponential

$$f(y)=\frac{1}{\theta}\;e^{-\frac{y}{\theta}}\;,\,0<y<\infty \tag{5.1}$$

for the following reasons :

1. It is positively skewed.

2. The computations involved are simpler than for other distributions.

Here,

$$REF\;(S_t^2)=\frac{(5n-3)\times 100}{(n^2-5n+6)\,\lambda^2-2(n-6)\,(n-1)\,\lambda+n\,(n-1)-3\,(n-1)\,p^2\,(1-q^2)} \tag{5.2}$$

The table below gives relative efficiency of S_t^2 for samples of sizes 5, 10, 20, 50, 100 from the exponential distributions. It can be seen from this table that the gains in relative efficiency are achieved for wide range of t. The gains become modest for large sample sizes and large values of t. This result is in conformity with that of Searls. The utility of estimator can be seen from the fact that in case of a sample of size 5, the expected number of observations greater than twice the mean is 0.677, ($=nq$ with $n=5$, $q=0.13524$) and the expected percentage of samples with one or more observations greater than twice the mean is 52.5% ($=100\ (1-p^n)$ with $n=5$, $q=.13524$) with relative efficiency 138.4%.

TABLE 1

Relative Efficiencies (%) of S_t^2 for samples from the Exponential distribution

Value of $\frac{t}{\theta}$	$\lambda=\frac{\sigma^{*2}}{\sigma^2}$	Sample Size 5	10	20	50	100
1	0.129	123.0	58.2	28.4	11.1	5.5
2	0.441	138.4	93.4	59.4	28.0	14.9
3	0.699	124.1	110.3	90.2	60.8	39.2
4	0.853	112.1	108.5	103.3	90.9	76.0
5	0.933	105.6	104.8	103.8	101.4	97.9
6	0.970	102.5	102.2	101.9	101.3	100.4
7	0.987	101.1	101.0	100.9	100.7	100.6
8	0.995	100.5	100.4	100.4	100.4	100.4
9	0.998	100.2	100.2	100.2	100.2	100.2
10	0.9999	100.0	100.0	100.0	100.0	100.0

SUMMARY

An estimator S_t^2 of the population variance σ^2 is presented which reduces the effect of large true observations. The proposed estimator is constructed by replacing all sample observations larger than a predetermined cutoff point t by the value of t in the usual unbiased estimator S^2 of σ^2. It has been demonstrated that a region of t exists where the mean square error of S_t^2 is smaller than the variance of S^2. An example has been given to show that there exists a wide range of cutoff point t for which S^2 is better than S^2.

6. ACKNOWLEDGEMENT

The authors are grateful to the referee for his suggestions which led to an improvement of the paper.

REFERENCES

[1] Crow, Edwin L., (1964), : "The Statistical construction of in single standard from several available standards". IEEE Transactions on Instrumentation and Measurements, 13 180-5.

[2] Dixon W.J., (1960), : "Simplified estimation from censored normal samples" Annals. of Mathematical Statistics, 31 385-91.

[3] Searls, Donald T., (1964), : "The utilisation of a Known Coefficient of variation in the estimation procedures" Journal Amer. Statist. Assocn, 59, 1225-6.

[4] ——— (1963), : "On the large observation problem" Ph.D. thesis, North Carolina State University.

[5] ——— (1966), : "An Estimator for a Population Mean which Reduces the Effect of Large True Observations". Journ. Amer. Statist. Assocn., 61, 1200-1204.

[6] Tukey J.W. and McLauglin, D.H.(1963), : "Less Vulnerable Confidence and Signficance procedures for location based on a single sample : trimming/Winsorization 1". Sankhya Series A, 25, 331-52.