Jour. Ind. Soc. Ag. Statistics
43 (1), 1991 : 49 – 58

ESTIMATION OF POPULATION VARIANCE IN THE PRESENCE OF LARGE TRUE OBSERVATIONS

T.P. Tripathi
Indian Statistical Institute, Calcutta
and
N.P. Katyar and H.P. Singh
J.N. Agricultural University, Jabalpur
(Received : August, 1988)

Summary

The problem of estimating population variance is considered in the presence of large true observations. A class of estimators, which are linear function of s_t^2 and $\bar{y}_t^2$ is presented. General properties of the class are obtained. The results are in particular applied to estimation of variance of one-parameter family of exponential populations. Conditions are obtained to generate estimators, from the proposed class, which are better than those considered by Ojha and Srivastava [3], Ojha [4] and Singh [9].

Key words : Class of estimators, optimum weights, exponential population, outliers, truncated distribution.

Introduction

It is well known that based on a random sample of $\{y_1, y_2, ..., y_n\}$ of size n, the sample mean $\bar{y}$ and mean square $s^2 = (n-1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$ are unbiased for population mean μ and variance σ^2 respectively and are commonly used in practice.

It may be noted that in case of the populations satisfying

$$\sigma^2 = C^2 \mu^2 \tag{1.1}$$

where C is a known constant ; or, in other words, the population coefficient of variation σ/μ is known to be C, another unbiased estimator of σ^2 is given by $[nC^2/(n+C^2)]\,\bar{y}^2$. For example in case of one-parameter family of exponential distributions which satisfy the above condition with $C^2 = 1$, the estimator $n\bar{y}^2/(n+1)$ discussed by Ojha [4] is unbiased for the variance. We also refer to Lee [2] and Singh [8] in this context. One may thus consider a class of estimators

$$d^* = w_1^* y^2 + w_2^* s^2 \tag{1.2}$$

for estimating the variance, especially in case of the populations satisfying (1.1), where w_1^* and w_2^* are suitably chosen weights.

In case of the distributions skewed at right, if the outliers ($y_j > t$) are present or if the distribution F(y) is truncated on the right at t, the customary estimators $\bar{y}$ and s^2 may not be suitable for estimating μ and σ^2 respectively. In such situations where some 'extremely large' values $y_j > t$ are present in the sample, an estimator for μ defined by

$$\bar{y}_t = n^{-1} \left[\sum_{j=1}^{r} y_j + (n-r)t \right], \quad (r=0,1,2,...,n) \; ; \; (y_j \le t) \tag{1.3}$$

is given by Searls [7]. He showed that there exists a wide range of the values of t in which mean square error (MSE) of $\bar{y}_t$ is less than the variance of $\bar{y}$. In the similar circumstances Ojha and Srivastava [3] proposed an estimator

$$s_t^2 = (n-1)^{-1} \left[\sum_{j=1}^{r} y_j^2 + (n-r)t^2 - n\bar{y}_t^2 \right], (r=0,1, \ldots, n); \; (y_j \le t) \tag{1.4}$$

for the variance σ^2 and found that s_t^2 is better than usual unbiased estimator s^2 for a wide range of cut off point t. Following the approach adopted by Searls [6] and Hirano [1], recently Singh [9] modified the above estimator to

$$T_1 = w\, s_t^2 \tag{1.5}$$

where w is a suitably chosen weight. In case the coefficient of kurtosis $\gamma_2^* = \beta_2^* - 3$, $\beta_2^* (= \mu_4^*/\sigma^{*4})$ of the right truncated distribution and variance ratio $\lambda \left(= \sigma^{*2}/\sigma^2 \right)$ are known exactly, the estimator

$$T_2 = \frac{n(n-1)\, s_t^2}{\lambda[(n-1)\beta_2^* + n^2 - 2n + 3]} \tag{1.6}$$

with

$$MSE(T_2) = \sigma^4 \left[1 - \frac{n(n-1)}{(n-1)\beta_2^* + n^2 - 2n + 3} \right] \tag{1.7}$$

is found to be the best (in the sense of having minimum MSE) in the class of estimators T_1.

In case of the populations satisfying (1.1), one may consider estimator, for σ^2 defined by

$$T_3 = w^* \bar{y}_t^2 \tag{1.8}$$

where w* is a suitably chosen weight. In case of one-parameter family of exponential populations truncated at t an estimator for variance due to Ojha [4] is given by

$$T_4 = \frac{n}{n+1} \bar{y}_t^2 \tag{1.9}$$

It is natural to consider a combined estimator

$$d = w_1 \bar{y}_t^2 + w_2 s_t^2 \tag{1.10}$$

for estimating σ^2 in the presence of 'extremely large' observations ($y_j > t$) where w_1 and w_2 are suitably chosen non-random weights whose sum need not be unity.

In the present paper, first we study the properties of the above estimator in general case and then in the particular case of one- parameter exponential family of distributions.

2. *Properties of the Proposed Class of Estimators*

The estimators in the proposed class d defined by (1.10) are, in general, biased and their biases and MSE are, respectively, given by

$$B(d) = w' \underline{Q} - \sigma^2 \tag{2.1}$$

and

$$M(d) = w' G w - 2\sigma^2 w' Q + \sigma^4 \tag{2.2}$$

where $w' = (w_1, w_2)$; $\underline{Q} = (E\bar{y}_t^2, Es_t^2)$ and $G = (g_{ik})$ is a positive definite matrix with

$$g_{11} = E(\bar{y}_t^4),\ g_{12} = g_{21} = E(\bar{y}_t^2 s_t^2),\ g_{22} = E(s_t^4)$$

After algebraic simplification, the optimum values $w_0 = (w_{01}, w_{02})'$ and minimum MSE are given by

$$w_{01} = \sigma^2 [E(s_t^4) E(\bar{y}_t^2) - E(s_t^2) E(\bar{y}_t^2 s_t^2)]/D(\bar{y}_t^2, s_t^2), \tag{2.3}$$

$$w_{02} = \sigma^2[\, E(s_t^2)\, E(\bar{y}_t^4) - E(\bar{y}_t^2) E(\bar{y}_t^2 s_t^2)]/D(\bar{y}_t^2, s_t^2),$$

and
$$M_0(d) = \sigma^4 \left[1 - \frac{N(\bar{y}_t^2, s_t^2)}{D(\bar{y}_t^2, s_t^2)}\right] \tag{2.4}$$

respectively and the resulting bias is given by

$$B_0(d) = -M_0^{(d)}/\sigma^2 \tag{2.5}$$

where
$$D(\bar{y}_t^2, s_t^2) = E(\bar{y}_t^4)\, E(s_t^4) - \left\{ E(\bar{y}_t^2, s_t^2) \right\}^2 \tag{2.6}$$

$$N(\bar{y}_t^2, s_t^2) = \left\{E(\bar{y}_t^2)\right\}^2 E(s_t^4) + \left\{E(s_t^2)\right\}^2 E(\bar{y}_t^4) - 2E(\bar{y}_t^2)\, E(s_t^2)\, E(\bar{y}_t^2, s_t^2)$$

Let μ_t and σ_t^2 denote the mean and variance and $\alpha_{3,t}$ and $\alpha_{4,t}$ denote third and fourth order moments about origin of the distribution F(y) truncated on the right at the point t. It is noted that the random variable r, the number of y_j's in the random sample $\{y_1,...,y_n\}$ satisfying $y_j \le t$ $(j = 1, \ldots, n)$, will have the binomial distribution with parameters n and $p = p[\, y_j \le t\,] = F(t)$. Let $q=1-p$. It is found that

$$E(\bar{y}_t^2) = \frac{\mu^{2*}}{n}(n + C^{*2}) \; ; \; E(s_t^2) = \sigma^{*2}$$

$$E(\bar{y}_t^2, s_t^2) = \frac{\mu^{*2}\sigma^{*2}}{n^2}[(\beta_2^* + n - 3)C^{*2} + 2n\sqrt{\beta_1^*}\, C^* + n^2] \tag{2.7}$$

$$E(s_t^4) = \frac{\sigma^{*4}}{n(n-1)}[\, \beta_2^*(n-1) + n^2 - 2n + 3\,]$$

$$E(\bar{y}_t^4) = \frac{\mu^{*4}}{n^3}[\{\beta_2^* + 3(n-1)\}C^{*4} + 4n\sqrt{\beta_1^*}\, C^{*3} + 6n^2C^{*2} + n^3]$$

where
$$\mu^* = p\mu_t + qt \; ; \; \sigma^{*2} = p(\sigma_t^2 + \mu_t^2) + qt^2 - \mu^{*2}$$

$$\mu_3^* = p\,\alpha_{3,t} + q\,t^3 - 3(\sigma^{*2} + \mu^{*2})\mu^* + 2\mu^{*3}$$

$$\mu_4^* = p\alpha_{4,t} + qt^4 - 4(p\alpha_{3,t} + qt^3)\mu^* + 6(\sigma^{*2} + \mu^{*2})\mu^{*2} - 3\mu^{*4}$$

$$C^* = \frac{\sigma^*}{\mu^*};\ \beta_1^* = \left(\frac{\mu_3^*}{\sigma^{*3}}\right)^2;\ \beta_2^* = \frac{\mu_4^*}{\sigma^{*4}}$$

It may be noted that the estimators T_1 to T_4 being particular members of the class d in (1.10), each of them is dominated by the optimum estimator

$$d_0 = w_{01}\,\bar{y}_t^2 + w_{02}\,s_t^2$$

whatever weights are used in T_1 and T_3. However the optimum estimators $T_{01} = w_0\,s_t^2$, $T_{03} = w_0^*\,\bar{y}_t^2$ and d_0 are not quite useful from application point of view as the optimum weights depend on the unknown parameters C^*, β_1^*, β_2^* and $\lambda = (\sigma^*/\sigma)^2$. In such a situation the technique suggested by Tripathi, Maiti and Sharma [10] may be quite useful for generating estimators from d better than s_t^2 and T_1, T_2, T_3 and T_4 without depending on the exact optimum weights.

Using (2.2) it may be shown that the estimator $d = T_1 + W_1\,\bar{y}_t^2$ would be better than T_1

iff w_1 lies between 0 and $2w_{01}^*$ (2.9)

where $w_{01}^* = [\sigma^2\,E(\bar{y}_t^2) - wE(\bar{y}_t^2, s_t^2)] / E(\bar{y}_t^4)$

is the optimum choice of w_1, for fixed w, in d. In particular we may set

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

where $0 < \lambda \le \lambda^{(1)}$, $\beta_2^{*(2)} \ge \beta_2^*$, so that $T_1 = w^{**}\,s_t^2$ is better than s_t^2.

Further using (2.2) it may be shown that $d = T_3 + w_2\,s_t^2$ would be better than T_3

iff w_2 lies between 0 and $2w_{02}^*$ (2.10)

where $w_{02}^* = [\sigma^2\,E(s_t^2) - w^*\,E(\bar{y}_t^2\,s_t^2)] / E(s_t^4)$

is the optimum choice of w_2, for fixed w^* in d.

In the following section we discuss the properties of the proposed class of estimators for one-parameter family of exponential distributions.

3. *Properties of the Proposed Class of Estimators for Exponential Population*

It is well known that in case of one-parameter family of exponential distributions with probability density function

$$f(y,\theta) = \begin{cases} \frac{1}{\theta} e^{-y/\theta} ; & y > 0 , \theta > 0 \\ 0 & \text{otherwise} \end{cases} \qquad (3.1)$$

the mean and variance are given by θ and θ^2 respectively satisfying the condition (1.1) with $C^2=1$.

For such populations we obtain

$$E(\bar{y}_t^2) = \frac{\theta^2}{n}(\lambda + np^2) ; E(s_t^2) = \theta^2 \lambda ; E(\bar{y}_t^4) = \frac{\theta^3}{n^3}(A+B)$$

$$E(\bar{y}_t^2 s_t^2) = \frac{\theta^4}{n^2}(E+F) ; E(s_t^4) = \frac{\theta^4}{n(n-1)}(C-D) \qquad (3.2)$$

where $p = F(t) = p(y_j \le t) = 1 - \exp(-t/\theta) ; q = 1-p; \lambda = 1-2q(1/\theta) - q^2$

$$A = 3\lambda^2 (n-2) + 6\lambda(n^2p^2 + 2npq + 2)$$

$$B = np^3(n^2p + 4q + 8) - 12npq\,(1/\theta)^2 - 4q\,(1/\theta)^3$$

$$C = \lambda^2 (n^2 - 5n + 6) + 12\lambda(n-1) ; \; D = 4(n-1)\, q(1/\theta)^3$$

$$E = \lambda^2 (n-6) + \lambda(n^2p^2 + 6npq + 12)$$

$$F = 2np \left\{ p^2 (q+2) - 3q(1/\theta)^2 \right\} - 4q(1/\theta)^3$$

From (2.1) to (2.6) and (3.2) we obtain

$$B(d) = \frac{\theta^2}{n}[w_1(\lambda + np^2) + n\lambda w_2 - n],$$

$$M(d) = \frac{\theta^4}{n^3(n-1)}\left[w_1^2 (n-1)(A+B) + w_2^2 n^2 (C-D) + 2w_1 w_2 n(n-1)(E+F)\right.$$

$$\left. -2w_1 n^2 (n-1)(\lambda + np^2) - 2n^3 (n-1)\lambda w_2 + n^3 (n-1)\right] \qquad (3.3$$

$$w_{01} = \frac{n^2[(\lambda + np^2)(C-D) - (n-1)\lambda(E+F)]}{(A+B)(C-D) - (n-1)(E+F)^2} \tag{3.4}$$

$$w_{02} = \frac{n(n-1)[(A+B)\lambda - (\lambda + np^2)(E+F)]}{(A+B)(C-D) - (n-1)(E+F)^2}$$

$$M_0(d) = \theta^4\left[1 - \frac{1}{n}\left\{(\lambda + np^2)w_{01} + n\lambda w_{02}\right\}\right] \tag{3.5}$$

and $\quad B_0(d) = -M_0^{(d)}/\theta^2$

As t approaches the upper limit of the distribution i.e. as $t \to \theta$ we have $p \to 1, q \to 0$ giving

$$B(d) = \frac{\theta^2}{n}[(n+1)w_1 + nw_2 - n]$$

and $$M(d) = \theta^4\left[\frac{w_1^2(n^3 + 6n^2 + 11n+6)}{n^3} + \frac{w_2^2(n^2 + 7n - 6)}{n(n-1)} + 1\right.$$

$$\left. + 2w_1w_2\frac{(n^2 + 5n + 6)}{n^2} - \frac{2w_1(n+1)}{n} - 2w_2\right] \tag{3.6}$$

which is same as obtained by Pandey and Singh [5] in case of the original distribution $f(y, \theta)$ in (3.1).

It is interesting to note that M(d) in (3.6) is minimized for

$$w_1 = \frac{n^2}{n^2 + 5n + 6}, \; w_2 = 0 \tag{3.7}$$

which indicates that one need not include the component s^2 in the estimators (1.2) for estimating the variance θ^2. However the same does not hold true in general for the estimators (1.10) suitable for estimating θ^2 in case of presence of outliers $y_j > t$. The optimum estimator

$$d_0 = w_{01}\bar{y}_t^2 + w_{02}s_t^2$$

is always better than the optimum estimators

$$T_{01} = w_0 s_t^2, \; w_0 = n(n-1)\lambda/(C-D)$$

$$T_{03} = w_0^* \bar{y}_t^2, \quad w_0^* = n^2(\lambda + np^2) / (A+B) \tag{3.8}$$

It may be noted that as $t/\theta \to \infty$

$$w_{01} \to \frac{n^2}{n^2 + 5n + 6}, \quad w_{02} \to 0 \tag{3.9}$$

and $$w_0 \to \frac{n(n-1)}{n^2 + 7n - 6}, \quad w_0^* \to \frac{n^2(n+1)}{n^3 + 6n^2 + 11n + 6}$$

It is noted that the optimum estimators d_0, T_{01} and T_{03} are not easy to apply in practice as the optimum weights involve the quantities p and t/q which may not be known exactly. However one can generate estimators from d better than T_{01} and T_{03} with the help of (2.9), (2.10) and (3.2) without knowing the exact optimum weights.

In Table-1, we present the optimum values of the weights w_1 and w_2 in the proposed estimator d for different cut-off points t in the range $1 \le t/\theta \le 10$ and for samples of sizes n=5, 10, 20 and 50. The table reveals that for a given n the values of w_{02} decrease monotonically as t/θ increases and for a fixed t/θ, they increase monotonically with increasing n. On the other hand it is found that in the range $1 \le t/\theta \le 2$, w_{01} is a monotonically decreasing function while in the range $3 \le t/\theta \le 10$ it is a monotonically increasing function of t/θ and n for given n and t/θ respectively.

Further it is observed that in the range $4 \le t/\theta \le 10$ and $5 \le n \le 50$

$0.33 < w_{01} < 0.91$; $0.005 < w_{02} < 0.52$ and

$0.44 < w_{01}+w_{02} < 1.03$.

Moreover in the range $4 \le t/\theta \le 10$,

for $n = 5$: $0.33 < w_{01} < 0.45$; $0.45 < w_{01}+w_{02} < 0.62$

$n = 10$: $0.41 < w_{01} < 0.64$; $0.64 < w_{01}+w_{02} < 0.81$

$n = 20$: $0.47 \le w_{01} < 0.79$; $0.79 < w_{01}+w_{02} < 0.94$

$n = 50$: $0.51 < w_{01} < 0.91$; $0.91 < w_{01}+w_{02} < 1.03$

It is noted that in the range $1 \le t/\theta \le 3$, $w_{02} > w_{01} \ge 0.138$, while in the range $4 \le t/\theta \le 10$, $0.005 \le w_{02} < w_{01}$. It indicates that role of the component

$w_2 s_t^2$ in d is quite important for $1 \leq t/\theta \leq 3$ but diminishes as t/θ increases and one may in practice use only the component $w_1 \bar{y}_t^2$ of d for $t/\theta > 10$.

The above observations may be quite helpful in making suitable choices of the values of w_1 and w_2 which may not be optimum but near optimum at least.

Table 1 : Optimum Values of Weights w_1 and w_2 in $d = w_1 \bar{y}_t^2 + w_2 s_t^2$

t/θ	n	5	10	20	50
1	w_{01}	1.071	0.990	0.940	0.907
	w_{02}	3.757	4.388	4.705	4.894
2	w_{01}	0.279	0.257	0.190	0.138
	w_{02}	1.208	1.596	1.834	1.993
3	w_{01}	0.295	0.286	0.266	0.245
	w_{02}	0.554	0.773	0.936	1.051
4	w_{01}	0.336	0.415	0.470	0.511
	w_{02}	0.275	0.383	0.460	0.516
5	w_{01}	0.378	0.515	0.621	0.704
	w_{02}	0.140	0.187	0.219	0.240
6	w_{01}	0.407	0.574	0.705	0.808
	w_{02}	0.072	0.092	0.104	0.112
7	w_{01}	0.425	0.607	0.748	0.859
	w_{02}	0.037	0.045	0.050	0.052
8	w_{01}	0.435	0.624	0.770	0.884
	w_{02}	0.019	0.022	0.024	0.025
9	w_{01}	0.441	0.632	0.781	0.896
	w_{02}	0.009	0.011	0.011	0.011
10	w_{01}	0.444	0.637	0.786	0.902
	w_{02}	0.005	0.005	0.005	0.005

ACKNOWLEDGEMENT

The authors are thankful to the referee for his valuable suggestions which have led to the considerable improvement in the presentation of the paper.

REFERENCES

[1] Hirano, K., 1973. Biased efficient estimators utilizing apriori information, *J. Japan Statist. Soc.* **4**, 1, 11-13.

[2] Lee, K.H. 1981. Estimation of variance of mean using known coefficient of variation. *Comm. Statist. Theor. Meth.* **A(10)** (5), 503-514.

[3] Ojha, V.P. and Srivastava, S.R., 1979. An estimator of the population variance in the presence of large true observations. *Jour. Ind. Soc. Agri. Statist.* **31**, (1), 77-84.

[4] Ojha, V.P., 1982. A note on estimation of variance in exponential density. *Jour. Ind. Soc. Agri. Statist.* **34**, (3), 82- 88.

[5] Pandey, B.N. and Singh, J., 1977. A note on estimation of variance in exponential density. *Sankhya, Sr. B*, **39**, (3), 294- 298.

[6] Searls, D.T., 1964. The utilization of a known coefficient of variation in the estimation procedure. *Jour. Amer. Statist. Assoc.* **59**, 1225-1226.

[7] Searls, D.T., 1966. An estimator for a population mean which reduces the effect of large true observations. *J. Amer. Statist. Assoc.* **61**, 1200-1204.

[8] Singh, H.P., 1986. A note on the estimation of variance of sample mean using knowledge of coefficient of variation in normal population. *Comm. Statist. Theory Meth*, **15**(12), 3737-3746.

[9] Singh, H.P. 1987. A modified estimator for population variance in the presence of large true observations. *Guj. Statist, Rev.* **15**,2, 15-30.

[10] Tripathi, T.P., Maiti, P. and Sharma, S.D., 1983. Use of prior information on some parameters in estimating population mean. *Sankhya, Sr. A*, **45**,3, 372-376.