A MODIFIED ESTIMATOR FOR POPULATION MEAN WHICH REDUCES THE EFFECT OF LARGE TRUE OBSERVATIONS

S. R. SRIVASTAVA, B. N. PANDEY
Banaras Hindu University
and
R. S. SRIVASTAVA
Gorakhpur University

(Received: March, 1982)

SUMMARY

A modified estimator of the population mean is suggested which reduces the effect of large true observations. The estimator makes use of a multiplier M which is so chosen that the mean squared error of the suggested estimator is least. The efficiency of the suggested estimator has also been studied with respect to Searls [2] and with simple mean estimator. The effect of the departure of the estimated M from the true M based on sample observations or on the guess value upon the efficiency of the estimator is also investigated.

Introduction

Let $y_1, y_2, \ldots, y_n$ be a random sample of size n from a population having mean μ and variance σ^2. If we are interested in the estimation of population mean μ, the sample mean $\bar{y}$ is the usual unbiased estimator. It may happen that out of these n observations, some observations may be very large. In this case the sample mean will always give an over estimate of the population mean. Searls [2] considered this problem and suggested an estimator

$$\bar{y}_t = \frac{\sum_{j=1}^{r} yj + (n-r)\,t}{n}, \quad r = 0, 1, 2, \ldots, n, \; yj \leqslant t. \tag{1}$$

Here y_j are independent random variables from the original distribution with p.d.f. $f(y)$ and cumulative distribution function $F(y)$ truncated on the right at t, t is the cut off point fixed by the experimenter according to his experience or by the behaviour of the system under considerations. He has shown that there exists a wide range of the values of t in which MSE $(\bar{y}_t)$ is less than $V(\bar{y})$. He has also obtained the optimum value of t for which MSE $(\bar{y}_t)$ is minimum. Some time it may happen that the cut off point t fixed by the experimenter may be beyond the range of the optimum value of t obtained in Searls [2]. In such situations the estimator proposed by Searls [2] may not give better result.

The proposed estimator is

$$\hat{y}_t = M\bar{y}_t \tag{2}$$

where M is so chosen that the mean squared error of the suggested estimator is least. The value of M will depend on the population parameters of the distribution namely μ, μ_t, μ_t', p and σ_t^2. If the distribution is specified the values of μ_t, μ_t', p and σ_t^2 can be obtained. We have considered exponential distribution and have obtained the value of M which minimizes MSE $(\hat{y}_t)$. We see that the value of M depends on the unknown parameter μ. If we replace the unknown parameters by their usual estimators, the estimated value of M can be obtained. The estimated value and the true value may be expressed as

$$\hat{M} = M\alpha \tag{3}$$

where M is the true value and α is any positive constant. We have also obtained the ranges of α for which the estimator

$$\hat{\hat{y}}_t = \hat{M}\,\bar{y}_t \tag{4}$$

will have smaller mean squared error than $\bar{y}_t$ and $\bar{y}$. The optimum value of t for which MSE $(\hat{y}_t)$ is least, is also obtained.

Estimator $\hat{y}_t$ and its properties

The proposed estimator is

$$\hat{y}_t = M\bar{y}_t. \text{ Now,}$$

$$E(\hat{y}_t) = M\,[p\mu_t + qt] \tag{5}$$

$$V(\hat{y}_t) = M^2\,\frac{p}{n}\,[\sigma_t^2 + q(t-\mu_t)^2] \tag{6}$$

where μ_t and σ_t^2 are the mean and variance of the truncated distribution on the right at t, $p = F(t)$ and $q = 1 - p$. From equations (5) and (6)

$$\text{Bias}\,(\hat{y}_t) = M[p\mu_t + qt] - \mu = -Mq(\mu_t' - t) + (M-1)\,\mu \tag{7}$$

$$\text{and MSE}(\hat{y}_t) = M^2\left[\frac{p}{n}\left\{\sigma_t^2 + q(t-\mu_t)^2\right\}\right] + q^2(\mu_t' - t)^2 - 2M(M-1)\,q\,\mu(\mu_t' - t) + (M-1)^2\,\mu^2. \tag{8}$$

where μ_t' is the mean of the left truncated distribution at the cut off point t. The value of M for which MSE$(\hat{y}_t)$ is least is given by

$$M\text{opt} = \frac{\mu^2 - \mu q(\mu_t' - t)}{\frac{p}{n}\left\{\sigma_t^2 + q(t-\mu_t)^2\right\} + q^2(\mu_t' - t)^2 + \mu^2 - 2\,\mu q(\mu_t' - t)} \tag{9}$$

As $t \to \infty$ then $q \to 0$, $p \to 1$ and $M\text{opt} = n/n + v^2$. Thus if coefficient of variation of the distribution is known as an a *priori*, the estimator

$$T_1 = \frac{n}{n+v^2}\,\bar{y} \tag{10}$$

proposed in Searls [1] has uniformally smaller mean squared error than the usual estimator $\bar{y}$. In equation (9) the value of Mopt depends on the unknown parameters namely μ, μ_t, μ^t, p and σ_t^2 respectively. If we consider an exponential distribution having mean μ and variance μ^2, then Mopt obtained in equation (9) reduces to

$$M'_{\text{opt}} = \frac{np}{p(2-p) + np^2 - 2q\,\frac{t}{\mu}}. \tag{11}$$

The above value of M depends on the unknown parameters μ, cut off point t and the sample size n. In Table 1, we have calculated the value of M for different values of t/μ and n. From this table it is evident that for fixed values of t/μ, the value of M increases as we increase the sample size. Again as t/μ increases the value of M decreases. This justify the proposal of an estimator $M\bar{y}_t$ instead of $\bar{y}_t$. The reason is that at t/μ is small, the estimator $\bar{y}_t$ will underestimate, but due to constant M, which is greater than one in this case, the estimate is corrected upto certain extent. Similarly if t/μ is large, the estimate $\bar{y}_t$ will overestimate, but due to constant M, which is lesser than one in this case, the estimate is again corrected upto certain extent.

TABLE 1—VALUES OF M, FOR DIFFERENT SAMPLE SIZES AND INTEGRAL VALUES OF t/μ

Values of t/μ	*Sample sizes n*				
	5	*10*	*50*	*100*	*500*
1	1.486117	1.532560	1.571860	1.576913	1.577011
2	1.034642	1.092191	1.143053	1.149747	1.155157
3	0.911332	0.976798	1.036357	1.044316	1.050772
4	0.892171	0.951226	1.004418	1.011487	1.017216
5	0.846384	0.919643	0.988058	0.997333	1.004878
6	0.838888	0.913419	0.983309	0.992804	1.000533
7	0.836465	0.911322	0.981614	0.991169	0.998949
8	0.834266	0.909784	0.980811	0.990477	0.998348
9	0.833701	0.909263	0.980540	0.990240	0.998130
10	0.833485	0.909200	0.980453	0.990152	0.998057
∞	0.833333	0.909090	0.980396	0.990099	0.998004

Putting the value of M in equation (7) from equation (9) we get

$$\mathrm{MSE}(\hat{y}_t) = \mu^2 - \frac{\{\mu^2 - \mu q(\mu_t' - t)\}^2}{p/n\{\sigma_t^2 + q(t-\mu_t)^2\} + q^2(\mu_t' - t)^2 + \mu^2 - 2\,\mu q(\mu_t' - t)} \tag{12}$$

In particular if we take the exponential distribution having mean μ and variance μ^2, then $\mathrm{MSE}(\hat{y}_t)$ reduces to

$$\mathrm{MSE}(\hat{y}_t) = \mu^2\left[\frac{p(2-p) - 2q\dfrac{t}{\mu}}{p(2-p) - 2q\dfrac{t}{\mu} + np^2}\right]. \tag{13}$$

As $t \to \infty$, then $p \to 1$ and $q \to 0$ and

$$\mathrm{MSE}(\hat{y}_t) \to \frac{\mu^2}{n+1}. \tag{14}$$

The relative efficiency of the estimator $\hat{y}_t$ with respect to $\bar{y}_t$ is defined as

$$\mathrm{REF}(\hat{y}_t; \bar{y}_t) = \frac{\mathrm{MSE}(\bar{y}_t)}{\mathrm{MSE}(\hat{y}_t)}, \text{ where} \tag{14}$$

$$\mathrm{MSE}(\bar{y}_t) = \frac{p}{n}\left[\sigma_t^2 + q(t-\mu_t)^2\right] + q^2(\mu_t' - t)^2. \tag{15}$$

In exponential distribution

$$\mathrm{MSE}(\bar{y}_t) = \frac{\mu^2}{n}\left\{p(2-p) - 2q\frac{t}{\mu} + nq^2\right\}. \tag{16}$$

Similarly the relative efficiency of $\hat{y}_t$ with respect to $\bar{y}$ is defined as

$$\mathrm{REF}\,(\hat{y}_t, \bar{y}) = \frac{\mathrm{MSE}(\bar{y})}{\mathrm{MSE}(\hat{y}_t)} \tag{17}$$

In Table 2 and Table 3 we have calculated the relative efficiencies of $\hat{y}_t$ relative to $\bar{y}_t$ and $\hat{y}_t$ relative to $\bar{y}$ for different values of t/μ and n in the

TABLE 2—RELATIVE EFFICIENCIES OF $\hat{y}_t$ RELATIVE TO $\bar{y}_t$

(in percentages)

Values of t/μ	*Sample sizes n*				
	5	*10*	*50*	*100*	*500*
1	265.87	474.45	2154.50	4264.00	21487.70
2	100.95	112.09	232.96	388.14	1630.92
3	106.12	100.72	107.95	123.25	250.78
4	110.30	103.71	100.13	101.82	120.21
5	117.38	108.06	100.77	100.08	101.21
6	118.91	109.22	101.43	100.52	100.01
7	119.44	109.63	101.78	100.81	100.05
8	119.83	109.87	101.92	100.93	100.14
9	119.93	109.95	101.96	100.98	100.19
10	100.97	109.98	101.99	100.99	100.19

case when the parent population is assumed to be exponential. From Table 2, we see that $\hat{y}_t$ has uniformally smaller mean squared error than $\bar{y}_t$. From Table 3, we see that the estimator $\hat{y}_t$ is also better than the simple mean estimator $\bar{y}$ for those values of t/μ where $\bar{y}_t$ is not better than $\bar{y}$. Thus the proposed estimator is preferable if M is known as *a priori.* But since M depends on the unknown parameters, therefore in practical problems generally M will be unknown. If we replace the unknown

TABLE 3—RELATIVE EFFICIENCIES OF $\hat{y}_t$ RELATIVE TO $\bar{y}$

(in percentage)

Values of t/μ	Sample sizes n				
	5	10	50	100	500
1	221.73	209.23	198.21	196.14	193.39
2	184.84	171.83	155.38	151.76	148.42
3	149.10	138.99	130.19	128.43	125.14
4	129.05	121.03	115.05	114.75	117.44
5	125.83	115.84	107.72	106.78	105,97
6	122.60	112.60	104.47	103.53	102.71
7	120.99	111.05	103.10	102.12	101.35
8	120.42	110.42	102.43	101.43	100.64
9	120.17	110.17	102.16	101.18	100.39
10	120.09	110.09	102.09	101.09	100.29

parameters by their usual estimators then M can be estimated and the proposed estimator will be

$$\hat{\hat{y}}_t = \hat{M}\bar{y}_t. \text{ Suppose} \tag{18}$$

$$\hat{M} = M\alpha \tag{19}$$

where M is the true value and α is a positive constant. In order to have

$$\text{MSE}(\hat{\hat{y}}_t) \leqslant \text{MSE}(\bar{y}_t) \text{ and}$$

$$\text{MSE}(\hat{\hat{y}}_t) \leqslant \text{MSE}(\bar{y}_t)$$

we should have the following in equality,

$$(1-\alpha^2M^2)\,\text{MSE}(\bar{y}_t) + 2\alpha M(\alpha M-1)\,\mu q(\mu_t'-t) - (\alpha M-1)^2\mu_t^2 \geqslant 0 \tag{20}$$

and

$$\frac{\mu^2}{n} - \alpha^2M^2\,\text{MSE}(\bar{y}_t) + 2\alpha M(\alpha M-1)\,\mu q(\mu_t'-t) - (\alpha M-1)^2\mu^2 \geqslant 0 \tag{21}$$

respectively.

In Table 4, for the exponential distribution, we have calculated the

TABLE 4—THE RANGES OF α (IN PERCENTAGES) FOR DIFFERENT VALUES OF t/μ, n AND THE VALUES OF M GIVEN IN TABLE 1 FOR WHICH EQUATIONS (20) AND (21) HOLDS

t/μ \ n	5	10	50	100	500
1	67.3 ~ 139.9 (61.7 ~ 145.5)	65.2 ~ 142.3 (72.8 ~ 134.7)	63.6 ~ 144.2 (84.7 ~ 123.1)	63.5 ~ 144.5 (86.7 ~ 121.2)	63.4 ~ 145 0 (88.6 ~ 119.8)
2	96.6 ~ 103.4 (67.5 ~ 132.5)	91.6 ~ 108.4 (78.3 ~ 121.68)	87.5 ~ 112.5 (90.8 ~ 109.2)	87.0 ~ 113.0 (93.5 ~ 106.5)	85.7 ~ 115.0 (93.8 ~ 106.9)
3	88.9 ~ 109.7 (73.1 ~ 125.5)	96.9 ~ 102.2 (83.3 ~ 115.9)	96.5 ~ 103.3 (93.6 ~ 106.3)	95.8 ~ 104.2 (95.6 ~ 104.4)	95.2 ~ 104.8 (98.0 ~ 101.9)
4	81.3 ~ 112.1 (73.9 ~ 119.4)	91.3 ~ 105.1 (84.8 ~ 111.6)	98.4 ~ 100.9 (94.7 ~ 104.6)	99.0 ~ 100.6 (96.6 ~ 103.0)	98.3 ~ 101.7 (98.6 ~ 101.4)
5	81.3 ~ 118.5 (78.4 ~ 121.0)	90.9 ~ 108.7 (88.3 ~ 111.3)	98.7 ~ 101.2 (96.7 ~ 103.3)	99.7 ~ 100.3 (97.9 ~ 102.1)	99.5 ~ 100.5 (99.1 ~ 100.9)
6	80.4 ~ 119.2 (79.3 ~ 120.3)	90.3 ~ 109.5 (89.3 ~ 110.4)	98.2 ~ 101.7 (96.6 ~ 103.3)	99.2 ~ 100.7 (98.5 ~ 101.5)	99.9 ~ 100.1 (99.4 ~ 100.6)
7	80.2 ~ 119.6 (78.9 ~ 120.8)	90.1 ~ 109.7 (89.5 ~ 110.4)	98.1 ~ 101.9 (97.5 ~ 102.5)	99.1 ~ 100.9 (98.6 ~ 101.5)	99.9 ~ 100.1 (99.5 ~ 100.5)
8	80.1 ~ 119.9 (79.8 ~ 120.2)	90.1 ~ 109.9 (89.8 ~ 110.2)	98.0 ~ 102.0 (97.8 ~ 102.2)	99.0 ~ 101.0 (98.8 ~ 101.2)	99.8 ~ 100.2 (99.6 ~ 100.4)
9	80.1 ~ 119.9 (79.9 ~ 120.1)	90.0 ~ 110.0 (89.9 ~ 110.1)	98.0 ~ 102.0 (97.9 ~ 102.1)	97.1 ~ 101.0 (98.5 ~ 101.0)	99.8 ~ 100.2 (99.5 ~ 100.3)
10	80.0 ~ 120.0 (79.9 ~ 120.1)	90.0 ~ 110.0 (89.9 ~ 110.1)	98.0 ~ 102.0 (97.9 ~ 102.1)	99.0 ~ 101.0 (98.9 ~ 101.1)	99.8 ~ 100.2 (99.7 ~ 100.3)
	80.0 ~ 120.0 (79.9 ~ 120.1)	89.4 ~ 110.1 (89.8 ~ 110.1)	98.0 ~ 102.0 (97.9 ~ 102.1)	99.0 ~ 101.0 (98.9 ~ 101.1)	99.8 ~ 100.2 (99.7 ~ 100.3)

ranges of α (%) for different values of t/μ, n and the values of M given in Table 1. From this table we see that the proposed estimator has also smaller mean squared error than $\bar{y}_t$ and $\bar{y}$ for some estimated M also. So we can prefer this estimator in the situations when some large true observations are present.

The optimum choice of t for which MSE$(\hat{\bar{y}}_t)$ is minimum is given by

$$t = \frac{\frac{\mu}{M} - \mu + \frac{p}{n}\mu_t + q\mu_t'}{\frac{p}{n} + q}. \tag{22}$$

ACKNOWLEDGEMENT

The authors are thankful to the referees for their valuable suggestions. They are also thankful to Prof. B. G. Verma, Head of the Department of Mathematics & Statistics for encouragement.

REFERENCES

[1] Searls, D. T. (1964): The utilization of a known coefficient of variation in the estimation procedure, *Jour. Amer. Stat. Assoc.*, **59**; 1225-26.

[2] Searls, D. T. (1966): An estimator for a population mean which reduces the effect of large true observations, *Jour. Amer. Stat. Assoc.*, **61**; 1200-1204.