Ann. Inst. Statist. Math.
31 (1979), Part A, 225-233

DOUBLE STAGE ESTIMATION OF POPULATION VARIANCE

B. N. PANDEY

(Received July 20, 1972; revised Mar. 14, 1975)

Summary

Consider a normal population with mean μ and variance σ^2. We are interested in the estimation of population variance with the help of guess value σ_0^2 and a sample of observations. In this paper, a double stage shrinkage estimator $\hat{\sigma}_k^2$ based on the shrinkage estimator $ks_1^2+(1-k)\sigma_0^2$ if $s_1^2\in R$ and the usual estimator $s^2=\dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$ if $s_1^2\notin R$, where R is some specified region, have been proposed. The expressions for bias and mean squared error have been obtained. Comparison with the usual estimator s^2 have been made. It was found that though the largest gain is obtained for $k=0$, we can use $\hat{\sigma}_k^2$ with $0\leqq k\leqq 1/2$ even when σ^2 is very close to σ_0^2.

1. Introduction

Consider a normal population with mean μ and variance σ^2. Suppose that our a priori knowledge about the population variance σ^2 is in the form of an initial estimate σ_0^2. Let $x_1, x_2, \cdots, x_n$ be a random sample of size n from a normal population with mean μ and variance σ^2. The minimum variance unbiased estimator for σ^2 is s^2. We are interested in estimating σ^2 with the help of σ_0^2 and a random sample of observations. Thompson [5] considered a method of shrinking the minimum variance unbiased estimators towards the natural origin by multiplying it by a shrinkage factor c. He considered the estimation of population mean μ and propose the estimator

$$\hat{\mu}=\frac{(\bar{x}-\mu_0)^3}{(\bar{x}-\mu_0)^2+s^2/n}+\mu_0 .$$

The above estimator have higher efficiency when μ is very close to μ_0. Similarly the shrinkage estimator for population variance will be

$$\hat{\sigma}^2=\frac{(s^2-\sigma_0^2)^3}{(s^2-\sigma_0^2)^2+2s^4/n+1}+\sigma_0^2$$

which have higher efficiency when σ^2 is very close to σ_0^2. This suggest that we have to use $ks^2+(1-k)\sigma_0^2$ instead of s^2 when σ^2 is very close to σ_0^2. Therefore we can propose a preliminary test estimator

$$\hat{\sigma}_p^2=\begin{cases} ks^2+(1-k)\sigma_0^2 & \text{if } \sigma^2 \text{ is close to } \sigma_0^2 \\ s^2 & \text{if } \sigma^2 \text{ is not close to } \sigma_0^2. \end{cases}$$

The above estimator have been discussed in [4]. Here to test whether σ_0^2 is close to σ^2 or not, we apply the test statistic $(n-1)s^2/\sigma_0^2$, which follow the chi-square distribution with $(n-1)$ degrees of freedom. Let r_1, and r_2 be such that

$$\mathrm{P}\,[r_1\leqq(n-1)s^2/\sigma_0^2\leqq r_2]\geqq 1-\alpha\,.$$

Therefore we can say that σ_0^2 is close to σ^2 if $s^2\in R$. Hence the preliminary test estimator is

$$\hat{\sigma}_p^2=\begin{cases} ks^2+(1-k)\sigma_0^2 & \text{if } s^2\in R \\ s^2 & \text{if } s^2\notin R\,. \end{cases}$$

Katti [2] have considered a double stage scheme for estimating the mean μ when variance σ^2 is known and when an a priori estimate is given as μ_0. The estimator considered by Katti [2] is

$$\hat{\mu}=\begin{cases} \bar{x}_1 & \text{if } \bar{x}_1\in R_0 \\ \dfrac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2} & \text{if } \bar{x}_1\notin R_0 \end{cases}$$

where

$$R_0=\left[\mu_0-\frac{\sigma}{\sqrt{2n_1+n_2}},\ \mu_0+\frac{\sigma}{\sqrt{2n_1+n_2}}\right]=[c_1,\ c_2]$$

and is obtained by minimizing MSE $(\hat{\mu}/\mu_0)$. Similarly if we consider a double stage scheme for estimating the population variance σ^2, when an a priori estimate is given as σ_0^2, the estimator will be

$$\hat{\sigma}^2=\begin{cases} s_1^2 & \text{if } s_1^2\in R_0 \\ \dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} & \text{if } s_1^2\notin R_0 \end{cases}$$

where

$$R_0=\left[\sigma_0^2\left(1-\sqrt{\frac{2}{n_2+2n_1-3}}\right),\ \sigma_0^2\left(1+\sqrt{\frac{2}{n_2+2n_1-3}}\right)\right].$$

Arnold, J. C. and Bayyatti, Al. H. A. [1] have considered the estimation of population mean μ on double stage. They attempt to weight μ_0 and $\bar{x}_1$ by a constant k, $0 \leqq k \leqq 1$, such that the estimate is $k\bar{x}_1+(1-k)\mu_0$ if $\bar{x}_1 \in R$. k is a constant specified by the experimenter according to his beliefe in μ_0. A value of k close to zero implies a strong belief that μ_0 is near the true mean μ and a value near one causes the double stage estimator to be based essentially on the sample alone. The double stage estimator proposed by them is

$$\hat{\mu}_k=\begin{cases} k\bar{x}_1+(1-k)\mu_0 & \text{if } \bar{x}_1 \in R \\ \dfrac{n_1\bar{x}_1+n_2\bar{x}_2}{n_1+n_2} & \text{if } \bar{x}_1 \notin R\,. \end{cases}$$

Similarly the double stage shrinkage estimator for population variance will be

$$\hat{\sigma}_k^2=\begin{cases} ks_1^2+(1-k)\sigma_0^2 & \text{if } s_1^2 \in R \\ \dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} & \text{if } s_1^2 \notin R \end{cases}$$

where R can be obtained by minimizing the MSE $(\hat{\sigma}_k^2/\sigma_0^2)$.

In this paper we have also proposed a double stage shrinkage estimator $\hat{\sigma}_k^2$. We take a sample of size n_1 and compute

$$s_1^2=\frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x}_1)^2\,, \qquad \bar{x}_1=\frac{1}{n_1}\sum_{i=1}^{n_1}x_i\,.$$

If s_1^2 implies that our a priori estimate was reasonable, we stop sampling and shrink s_1^2 towards σ_0^2. If not so, we take additional sample of size n_2 and compute the pooled sample variance

$$s^2=\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$$

where

$$s_2^2=\frac{1}{n_2-1}\sum_{i=1}^{n_2}(x_i-\bar{x}_2)^2\,, \qquad \bar{x}_2=\frac{1}{n_2}\sum_{i=1}^{n_2}x_i\,.$$

Therefore the proposed double stage shrunken estimator is

$$\hat{\sigma}_k^2=\begin{cases} ks_1^2+(1-k)\sigma_0^2 & \text{if } \dfrac{r_1\sigma_0^2}{n_1-1} \leqq s_1^2 \leqq \dfrac{r_2\sigma_0^2}{n_1-1} \\ \dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} & \text{if } s_1^2<\dfrac{r_1\sigma_0^2}{n_1-1} \text{ and } s_1^2>\dfrac{r_2\sigma_0^2}{n_1-1}\,. \end{cases}$$

2. Bias and mean squared error

The proposed estimator is

$$\hat{\sigma}_k^2=\begin{cases} ks_1^2+(1-k)\sigma_0^2 & \text{if } s_1^2\in R \\ \dfrac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2} & \text{if } s_1^2\notin R \end{cases} \tag{2.1}$$

where

$$R=\left[\frac{r_1\sigma_0^2}{n_1-1},\ \frac{r_2\sigma_0^2}{n_1-1}\right].$$

The expected value of $\hat{\sigma}_k^2$ can be written as

$$\begin{aligned} \mathrm{E}\,(\hat{\sigma}_k^2)&=\iint \hat{\sigma}_k^2(s_1^2,\ s_2^2)p(s_1^2)p(s_2^2)ds_1^2ds_2^2 \\ &=\int_R \{ks_1^2+(1-k)\sigma_0^2\}\,p(s_1^2)ds_1^2 \\ &\quad+\int_0^\infty\int_{R^c}\frac{m_1s_1^2+m_2s_2^2}{m_1+m_2}\,p(s_1^2)p(s_2^2)ds_1^2ds_2^2 \end{aligned} \tag{2.2}$$

where $m_i=n_i-1,\ i=1,2$ and

$$p(s_i^2)=\frac{1}{\left(\dfrac{2\sigma^2}{m_i}\right)^{m_i/2}\Gamma\left(\dfrac{m_i}{2}\right)}\exp\left(-\frac{m_is_i^2}{2\sigma^2}\right)(s_i^2)^{m_i/2-1}.$$

Now,

$$\begin{aligned} \mathrm{Bias}\,(\hat{\sigma}_k^2)&=\mathrm{E}\,(\hat{\sigma}_k^2)-\sigma^2 \\ &=\left(k-\frac{m_1}{m_1+m_2}\right)\sigma^2Q'+(1-k)\sigma_0^2R'-\frac{m_2}{m_1+m_2}R' \end{aligned} \tag{2.3}$$

where

$$Q'=I\left(\frac{r_2\sigma_0^2}{2\sigma^2},\ \frac{m_1}{2}\right)-I\left(\frac{r_1\sigma_0^2}{2\sigma^2},\ \frac{m_1}{2}\right)$$

$$R'=I\left(\frac{r_2\sigma_0^2}{2\sigma^2},\ \frac{m_1}{2}-1\right)-I\left(\frac{r_1\sigma_0^2}{2\sigma^2},\ \frac{m_1}{2}-1\right)$$

and

$$I(x,\ p-1)=\frac{1}{\Gamma(p)}\int_0^x e^{-t}t^{p-1}dt \qquad \text{(Incomplete Gamma Integral).}$$

If $k=0$, the proposed estimator is

$$\hat{\sigma}_0^2=\begin{cases}\sigma_0^2 & \text{if } s_1^2\in R\\ \dfrac{m_1s_1^2+m_2s_2^2}{m_1+m_2} & \text{if } s_1^2\notin R\,.\end{cases} \tag{2.4}$$

Now,

$$\text{Bias}\,(\hat{\sigma}_0^2)=-\frac{m_1}{m_1+m_2}\sigma^2Q'+\sigma_0^2R'-\frac{m_2}{m_1+m_2}R'\,. \tag{2.5}$$

If $k=1$, the proposed estimator is

$$\hat{\sigma}_1^2=\begin{cases}s_1^2 & \text{if } s_1^2\in R\\ \dfrac{m_1s_1^2+m_2s_2^2}{m_1+m_2} & \text{if } s_1^2\notin R\,.\end{cases} \tag{2.6}$$

Now,

$$\text{Bias}\,(\hat{\sigma}_1^2)=\frac{m_2\sigma^2}{m_1+m_2}Q'-\frac{m_2\sigma^2}{m_1+m_2}R'\,. \tag{2.7}$$

Again we have,

$$\begin{aligned}\text{MSE}\,(\hat{\sigma}_k^2)&=\text{E}\,(\hat{\sigma}_k^2-\sigma^2)^2\\ &=\iint(\hat{\sigma}_k^2-\sigma^2)^2(s_1^2,\,s_2^2)p(s_1^2)p(s_2^2)ds_1^2ds_2^2\\ &=\int_R\{ks_1^2+(1-k)\sigma_0^2-\sigma^2\}^2p(s_1^2)ds_1^2\\ &\quad+\int_0^\infty\int_{R^c}\left\{\frac{m_1s_1^2+m_2s_2^2}{m_1+m_2}-\sigma^2\right\}^2p(s_1^2)p(s_2^2)ds_1^2ds_2^2\,.\end{aligned} \tag{2.8}$$

After simplification we get

$$\begin{aligned}\text{MSE}\,(\hat{\sigma}_k^2)=&\frac{2\sigma^4}{m_1+m_2}+\left\{k^2-\frac{m_1^2}{(m_1+m_2)^2}\right\}\frac{m_1+2}{m_1}\sigma^4P'\\ &+\left\{2k\sigma^2(\sigma_0^2-\sigma^2)-2k\sigma_0^2\sigma^2+\frac{2m_1^2\sigma^4}{(m_1+m_2)^2}\right\}Q'\\ &+\left\{(\sigma_0^2-\sigma^2)^2+k^2\sigma_0^4-2k\sigma_0^2(\sigma_0^2-\sigma^2)-\frac{(m_1^2+2m_2)\sigma^4}{(m_1+m_2)^2}\right\}R'\end{aligned} \tag{2.9}$$

which will be equal to

$$\begin{aligned}&\frac{2\sigma^4}{m_1+m_2}-\frac{(m_1+2)m_1}{(m_1+m_2)^2}\sigma^4P'+\frac{2m_1^2\sigma^4Q'}{(m_1+m_2)^2}\\ &\qquad+\left\{(\sigma_0^2-\sigma^2)^2-\frac{(m_1^2+2m_2)\sigma^4}{(m_1+m_2)^2}\right\}R'\qquad\text{for } k=0\end{aligned} \tag{2.10}$$

and

$$\text{(2.11)}\quad \frac{2\sigma^4}{m_1+m_2}+\left\{1-\frac{m_1^2}{(m_1+m_2)^2}\right\}\frac{m_1+2}{m_1}\sigma^4 P' + \left\{\frac{2m_1^2\sigma^4}{(m_1+m_2)^2}-2\sigma^4\right\}Q'+\left\{\sigma^4-\frac{m_1^2+2m_2}{(m_1+m_2)^2}\sigma^4\right\}R' \qquad \text{for } k=1$$

respectively. Now differentiating (2.9) with respect to k and putting the derivative equal to zero, we get

$$\text{(2.12)}\qquad k\frac{m_1+2}{m_1}\sigma^4 P'-(\sigma_0^4-k\sigma_0^4+\sigma^2\sigma_0^2)R'+(\sigma^2\sigma_0^2-2k\sigma^2\sigma_0^2-\sigma^4)Q'=0$$

which gives

$$\text{(2.13)}\qquad k=\frac{(\sigma_0^4-\sigma^2\sigma_0^2)R'+(\sigma^4-\sigma^2\sigma_0^2)Q'}{\frac{m_1+2}{m_1}\sigma^4 P'+\sigma_0^4 R'-2\sigma^2\sigma_0^2 Q'}\,.$$

Again differentiating (2.12) with respect to k we get

$$\text{(2.14)}\qquad \frac{m_1+2}{m_1}\sigma^4 P'+\sigma_0^4 R'-2\sigma^2\sigma_0^2 Q'$$

which will be always positive. Hence the value of k obtained in (2.13) will give the minimum mean squared error. The value of k depends on the unknown values of σ^2, which can be estimated by s_1^2. Therefore the estimated value of k, on the basis of first sample will be

$$\text{(2.15)}\qquad k_0=\hat{k}=\frac{(m_1+2)\{\sigma_0^4-s^2\sigma_0^2\}R''+[m_1 s_1^4-s_1^2(m_1+2)\sigma_0^2]Q''}{(m_1+2)[s_1^4 P''+\sigma_0^4 R''-2s_1^2\sigma_0^2 Q'']}$$

where

$$P''=I\left(\frac{r_2\sigma_0^2}{2s_1^2},\ \frac{m_1}{2}+1\right)-I\left(\frac{r_1\sigma_0^2}{2s_1^2},\ \frac{m_1}{2}+1\right)$$

$$Q''=I\left(\frac{r_2\sigma_0^2}{2s_1^2},\ \frac{m_1}{2}\right)-I\left(\frac{r_1\sigma_0^2}{2s_1^2},\ \frac{m_1}{2}\right)$$

$$R''=I\left(\frac{r_2\sigma_0^2}{2s_1^2},\ \frac{m_1}{2}-1\right)-I\left(\frac{r_1\sigma_0^2}{2s_1^2},\ \frac{m_1}{2}-1\right).$$

Therefore the proposed estimator will be

$$\text{(2.16)}\qquad \hat{\sigma}_{k_0}^2=\begin{cases} k_0 s_1^2+(1-k_0)\sigma_0^2 & \text{if } s_1^2\in R \\ \dfrac{m_1 s_1^2+m_2 s_2^2}{m_1+m_2} & \text{if } s_1^2\notin R\,. \end{cases}$$

The expressions for bias and mean squared error can be derived. It will involve complicated algebra.

3. Comparison

The relative efficiency of $\hat{\sigma}_k^2$ with respect to s^2 is defined as

$$(3.1)\quad \mathrm{REF}\,(\hat{\sigma}_k^2, s^2)=\frac{\mathrm{MSE}\,(s^2)}{\mathrm{MSE}\,(\hat{\sigma}_k^2)}$$

$$=\Bigg[1+\left(k^2-\frac{m_1^2}{(m_1+m_2)^2}\right)\frac{(m_1+m_2)(m_1+2)}{2m_1}P'$$

$$+\left\{\frac{k(\sigma_0^2-\sigma^2)(m_1+m_2)}{\sigma^2}-\frac{k(m_1+m_2)\sigma_0^2}{\sigma^2}+\frac{m_1^2}{m_1+m_2}\right\}Q'$$

$$+\left\{\frac{(m_1+m_2)(\sigma_0^2-\sigma^2)^2}{2\sigma^4}+\frac{(m_1+m_2)\sigma_0^4}{2\sigma^4}\right.$$

$$\left.-\frac{k(m_1+m_2)(\sigma_0^2-\sigma^2)\sigma_0^2}{\sigma^4}-\frac{m_1^2+2m_2}{m_1+m_2}\right\}R'\Bigg]^{-1}.$$

The numerical calculations of the relative efficiency have been shown in the following tables.

Table 3.1 $\alpha=.20$, $\sigma^2=\sigma_0^2=4$

n_1	n_2	k					
		.1	.2	.3	.5	.8	1.0
5	10	275.09	254.57	226.45	167.28	129.85	75.19
8	12	221.86	215.51	205.66	184.80	136.91	112.33
15	10	230.20	219.62	203.97	170.52	109.97	84.00
20	15	196.09	190.73	182.51	160.31	123.61	102.05
31	40	188.38	185.97	182.10	170.82	148.41	132.36

The above table shows that the relative efficiency is a decreasing function of k. The proposed estimator is better than the usual estimator s^2 if $0\leqq k\leqq .8$.

Table 3.2 $\alpha=.05$, $\sigma^2=\sigma_0^2=4$

n_1	n_2	k					
		.1	.2	.3	.5	.8	1.0
5	10	687.37	472.00	310.17	147.86	64.91	36.91
8	12	585.59	542.18	357.56	171.09	85.64	49.67
15	10	509.45	432.08	344.97	209.60	107.15	73.85
20	15	924.40	404.21	325.29	200.19	103.30	71.41
31	40	952.36	614.30	412.63	170.07	72.84	47.68

The proposed estimator is better than the usual estimator s^2 if $0\leqq k\leqq .6$.

Table 3.3 $\alpha=.01$, $\sigma^2=\sigma_0^2=4$

n_1	n_2	k					
		.1	.2	.3	.5	.8	1.0
5	10	1495.44	1341.35	319.38	123.65	49.99	31.83
8	12	1907.51	1109.40	453.92	152.13	58.03	36.96
15	10	1424.05	869.56	527.61	233.52	98.99	64.64
20	15	1605.46	903.73	523.03	222.76	92.74	60.34
31	40	356.49	291.15	222.98	127.49	62.37	42.39

The proposed estimator is better than the usual estimator s^2 if $0\leqq k\leqq .5$.

Hence from the above tables we can say that the relative efficiency is a decreasing function of α and k and the largest gain is obtained for $k=0$.

Table 3.4 $\alpha=.05$, $\sigma_0^2=4$, $\sigma^2=5$

n_1	n_2	k					
		.1	.2	.3	.5	.8	1.0
5	10	231.55	188.80	167.28	133.14	87.68	80.38
8	12	159.38	155.12	146.96	123.92	88.19	69.29
15	10	237.11	169.03	154.33	147.16	82.38	63.56
20	15	104.90	105.44	106.08	103.97	95.96	87.39
31	40	106.78	106.56	104.45	95.99	77.98	65.82

Table 3.5 $\alpha=.05$, $\sigma_0^2=4$, $\sigma^2=5$

n_1	n_2	k					
		.1	.2	.3	.5	.8	1.0
5	10	146.72	153.35	143.99	104.84	54.37	36.22
8	12	110.38	121.48	125.57	109.29	66.56	45.93
15	10	81.01	90.31	98.60	107.59	94.65	76.28
20	15	61.69	70.49	79.40	93.49	90.31	74.25
31	40	34.64	39.97	45.52	67.37	53.33	43.38

The above tables show that we can use the proposed estimator with $0\leqq k\leqq .5$ even when σ_0^2 is close to σ^2.

Table 3.6 $\alpha=.01$, $\sigma_0^2=4$, $\sigma^2=16$

n_1	n_2	k					
		.1	.2	.3	.5	.8	1.0
5	10	41.80	43.78	45.88	50.07	50.19	57.79
8	12	43.45	45.63	47.94	56.52	61.34	67.47
15	10	69.07	70.68	72.30	75.55	80.37	83.47
20	15	79.20	80.49	81.51	83.90	87.65	90.27
31	40	93.67	95.34	97.24	98.63	101.25	106.34

The above table shows that the proposed estimator $\hat{\sigma}_k^2$ is worse than s^2 if there is vast difference in σ_0^2 and σ^2.

Acknowledgement

The author is grateful to Dr. J. Singh for his advices and to the referee for his useful comments.

BANARASU HINDU UNIVERSITY

REFERENCES

[1] Arnold, J. C. and Bayyatti, Al. H. A. (1970). On double stage estimation of the mean using prior knowledge, *Biometrics*, **26**, 787-800.

[2] Katti, S. K. (1962). Use of some a priori knowledge in the estimation of means from double samples, *Biometrics*, **18**, 139-147.

[3] Pearson, E. S. and Hartley, H. A. (1939). *The Biomerika Tables*, Cambridge University Press, London.

[4] Pandey, B. N. and Singh, J. (1977). Estimation of variance of normal population using prior information, *J. Indian Statist. Ass.*, **15**, 141-150.

[5] Thompson, J. R. (1968). Some shrinkage techniques for estimating the mean, *J. Amer. Statist. Ass.*, **63**, 113-123.