PART III

CHAPTER 10

USE OF A PRIORI KNOWLEDGE IN THE ESTIMATION OF A PARAMETER FROM DOUBLE SAMPLES

10.1 Introduction

When there is no a priori information about the value of the population parameter, then various methods such as maximum likelihood, minimum variance etc., may be used to estimate the parameter. We suppose here that there exists an uniformly minimum variance unbiased estimator of the population parameter. In practical problems the experimenter possesses some guessed estimate for the value of the population parameter. Using this a priori information, Katti [22] obtained a better estimator of the population mean from double samples than the estimator obtained from the pooled sample when some guessed estimate for the value of the population mean was available. Here we generalise this method for obtaining a better estimator of the parameter from double samples when some guessed estimate of the value of the population parameter is available. The method developed here is applied to obtain a better estimator for the variance of a normal population from double samples than the estimator obtained from the pooled sample when some guessed estimate of the value of the variance is available. The method consists in constructing a region in the space of variance using the guessed estimate, using the estimater based on the single sample if the estimator belongs to the region and drawing a second sample and using the estimator based on both samples if the former estimator does not belong to the region. This region has been termed a preliminary test region (P.T. region) and the resultant estimator a preliminary test estimator (P.T. estimator) by Katti [22]. As pointed out by Katti [22], it is to be noted that this approach differs from the Bayesian approach in the sense that the guessed value is used only in constructing the region. The rest of the estimation procedure is free from the subjective judgement behind the guess. It is to be further observed that in such applications as agriculture, wherein samples can be drawn in succession, an acceptance of the estimate based on the first sample saves the second sample.

The procedure suggested here has some points of similarity with the two-sample procedure due to Stein

[52], in connection with determining the confidence intervals of preassigned length and confidence coefficients for the mean of a normal distribution with the unknown variance.

10.2 Preliminary test region and estimator

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the uniformly minimum variance unbiased estimators (UMVUE) of the population parameter θ based respectively on the first sample of size n_1 and the second sample of size n_2. Let $\hat{\theta} = a\hat{\theta}_1 + b\hat{\theta}_2$, where $a + b = 1$, be the best estimator of θ based on both samples, by best estimator it being understood that it is uniformly minimum variance unbiased estimator in the class of all unbiased estimators of the type $l\hat{\theta}_1 + m\hat{\theta}_2$, $l + m = 1$, $l \geq 0$, $m \geq 0$. Let Ω be the parameter space of θ and R' be the region of $\hat{\theta}_1$. R' will also be a region of $\hat{\theta}_2$. Ω and R' are intervals of the real line. Because $\hat{\theta} = a\hat{\theta}_1 + b\hat{\theta}_2$ is convex in $\hat{\theta}_1$ and $\hat{\theta}_2$ and only convex sets on the real line are intervals, it follows that $\hat{\theta}$ has the same range R'. The P.T. estimator consists of choosing a region R in the $\hat{\theta}_1$ space and accepting $\hat{\theta}_1$ as an estimate of θ if $\hat{\theta}_1 \in R$, and accepting the pooled estimate $\hat{\theta}$ if $\hat{\theta}_1 \in \bar{R}$, where $\bar{R}$ denotes the complement of R. The resultant P.T. estimator will be denoted by θ^P. The goodness of the estimator will be measured by its expected mean square (E.M.S.) i.e. $E(\theta^P - \theta)^2$.

Let $\hat{\theta}_1$ and $\hat{\theta}_2$ have continuous probability density functions $p_1(\hat{\theta}_1|\theta)$ and $p_2(\hat{\theta}_2|\theta)$. Let θ_0 be the guessed value of θ. Then if θ_0 is the true value of θ, the E.M.S. of θ^P is given by

$$\text{E.M.S.}(\theta^P|\theta_0)$$

$$= \int_{\hat{\theta}_1} \int_{\hat{\theta}_2} (\theta^P - \theta_0)^2 p_1(\hat{\theta}_1|\theta_0) p_2(\hat{\theta}_2|\theta_0) d\hat{\theta}_1 d\hat{\theta}_2$$

$$= \int_{\hat{\theta}_1 \in R} (\hat{\theta}_1 - \theta_0)^2 p_1(\hat{\theta}_1|\theta_0) d\hat{\theta}_1$$

$$(10.2.1) \qquad + \int_{\hat{\theta}_1 \in \bar{R}} \int_{\hat{\theta}_2} (\hat{\theta} - \theta_0)^2 p_1(\hat{\theta}_1|\theta_0) p_2(\hat{\theta}_2|\theta_0) d\hat{\theta}_1 d\hat{\theta}_2$$

$$= \int_{\hat{\theta}_1 \in R} (\hat{\theta}_1 - \theta_0)^2 p_1(\hat{\theta}_1|\theta_0) d\hat{\theta}_1$$

$$+ \int_{\hat{\theta}_1 \in \bar{R}} \left[a^2(\hat{\theta}_1 - \theta_0)^2 + b^2 V(\hat{\theta}_2|\theta_0)\right] p_1(\hat{\theta}_1|\theta_0) d\hat{\theta}_1$$

$$= a^2 V(\hat{\theta}_1|\theta_0) + b^2 V(\hat{\theta}_2|\theta_0) +$$

$$+ \int_{\hat{\theta}_1 \in R} \left[(1-a^2)(\hat{\theta}_1-\theta_0)^2-b^2V(\hat{\theta}_2|\theta_0)\right] p_1(\hat{\theta}_1|\theta_0)d\hat{\theta}_1,$$

where $V(\hat{\theta}_i|\theta_0)$ denotes the variance of $\hat{\theta}_i$, $i = 1,2$, if θ_0 is the true value of θ.

We choose R such that E.M.S.$(\theta^P|\theta_0)$ is minimum. Minimising E.M.S.$(\theta^P|\theta_0)$, we get

$$(10.2.2) \qquad (1 - a^2)(\hat{\theta}_1 - \theta_0)^2 - b^2V(\hat{\theta}_2|\theta_0) = 0.$$

Hence R is given by

$$(10.2.3)\ R=\left[\theta_0-b\sqrt{V(\hat{\theta}_2|\theta_0)/(1-a^2)},\ \theta_0+b\sqrt{V(\hat{\theta}_2|\theta_0)/(1-a^2)}\right]$$

and the P.T. estimator is given by

$$(10.2.4) \qquad \begin{aligned} \theta^P &= \hat{\theta}_1, \quad \text{if} \quad \theta_1 \in R, \\ &= \hat{\theta} = a\hat{\theta}_1 + b\hat{\theta}_2, \quad \text{if} \quad \theta_1 \in R. \end{aligned}$$

10.3 Properties of P.T. estimator

(a) Efficiency of P.T. estimator. When θ_0 is the true value of θ, then it follows from (10.2.1) that

E.M.S. of θ^P for the P.T. region is given by

$$\text{E.M.S.}(\theta^P|\theta_0)$$

$$(10.3.1) = V(\hat{\theta}|\theta_0)+(1-a^2)\int_{\hat{\theta}_1 \epsilon R}(\hat{\theta}_1-\theta_0)^2 p_1(\hat{\theta}_1|\theta_0)d\hat{\theta}_1$$

$$- b^2 V(\hat{\theta}_2|\theta_0)\int_{\hat{\theta}_1 \epsilon R} p_1(\hat{\theta}_1|\theta_0)d\hat{\theta}_1.$$

If however, θ_0 is not equal to the true value of θ, the E.M.S. of θ^P is given by

$$\text{E.M.S.}(\theta^P|\theta)$$

$$(10.3.2) = V(\hat{\theta}|\theta)+(1-a^2)\int_{\hat{\theta}_1 \epsilon R}(\hat{\theta}_1-\theta)^2 p_1(\hat{\theta}_1|\theta)d\hat{\theta}_1$$

$$- b^2 V(\hat{\theta}_2|\theta)\int_{\hat{\theta}_1 \epsilon R} p_1(\hat{\theta}_1|\theta)d\hat{\theta}_1.$$

The efficiency of θ^P relative to the best unbiased estimator based on both samples will be measured by

$$(10.3.3) \qquad E = [\text{E.M.S.}(\hat{\theta}|\theta)]/[\text{E.M.S.}(\theta^P|\theta)].$$

We give bounds for the efficiency E of the P.T. estimator in the following lemma.

Lemma 10.3.1. The efficiency E of the P.T. estimator θ^P relative to the best unbiased estimator θ based on both samples satisfies the inequality

$$[1 + (1-a^2)mC - b^2BC]^{-1} \geq E \geq [1 + (1-a^2)MC - b^2BC]^{-1} ,$$

where $A = b\sqrt{V(\hat{\theta}_2|\theta_0)/(1-a^2)}$, $B = V(\hat{\theta}_2|\theta)$,

$C = \int_{\hat{\theta}_1 \in R} p_1(\hat{\theta}_1|\theta)d\hat{\theta}_1/V(\hat{\theta}|\theta)$ and m and M are respectively lower and upper bounds of the function $f(\hat{\theta}_1|\theta) = (\hat{\theta}_1 - \theta)^2$ in the interval $(\theta_0 - A, \theta_0 + A)$, θ_0 being the guessed value of θ.

Proof. The result of Lemma 10.3.1 immediately follows from the following mean value theorem:

"If ϕ (x) be positive in the interval $(a < x < b)$ and $f(x)$ is integrable, then

$$m\int_a^b \phi(x)dx \leq \int_a^b \phi(x)f(x)dx \leq M \int_a^b \phi(x)dx,$$

where m and M are respectively lower and upper bounds of $f(x)$ in the interval (a,b),"

and letting $\phi(\hat{\theta}_1) = p_1(\hat{\theta}_1|\theta)$, $f(\hat{\theta}_1|\theta) = (\hat{\theta}_1 - \theta)^2$, $a = \theta_0 - A$ and $b = \theta_0 + A$.

We discuss below the efficiency of θ^P relative to θ for four cases: (i) $\theta \leq \theta_0 - A$, (ii) $\theta_0 - A \leq \theta \leq \theta_0$, (iii) $\theta_0 \leq \theta \leq \theta_0 + A$, and (iv) $\theta_0 + A \leq \theta$, because for each case m and M are easily found. In cases (i), (ii), (iii) and (iv), let

$$X = [1 + (1-a^2)(\theta_0-A-\theta)^2C - b^2BC]^{-1},$$

$$Y = [1 + (1-a^2)(\theta_0+A-\theta)^2C - b^2BC]^{-1},$$

$$Z = [1 - b^2BC]^{-1}.$$

Case (i). If $\theta \leq \theta_0 - A$, then E satisfies the inequality

(10.3.4) $$X \geq E \geq Y.$$

Case (ii). If $\theta_0 - A \leq \theta \leq \theta_0$, then E satisfies the inequality

(10.3.5) $$Z \geq E \geq Y.$$

Case (iii). If $\theta_0 \leq \theta \leq \theta_0 + A$, then E satisfies the inequality

(10.3.6) $$Z \geq E \geq X.$$

We can combine cases (ii) and (iii) to obtain wider bounds on E as

$$[1 - b^2BC]^{-1} \geq E \geq [1 + 4A^2(1 - a^2)C - b^2BC]^{-1}$$

whenever $|\theta - \theta_0| \leq A$.

Case (iv). If $\theta_0 + A \leq \theta$, then E satisfies the inequality

(10.3.7) $$Y \geq E \geq X.$$

Corollary 10.3.1. If $\theta = \theta_0$, i.e., when the guessed value θ_0 is equal to the true value θ, then E satisfies the inequality

$$[1 - b^2BC]^{-1} \geq E \geq 1.$$

This follows from (10.3.5) or (10.3.6) by putting $\theta = \theta_0$ and noting that $(1 - a^2)A^2 = b^2B$.

Corollary 10.3.2. The asymptotic efficiency E_∞ when $\theta \to \pm \infty$ is equal to unity.

This follows from (10.3.4) and (10.3.7) by noting that $BC \to 0$ and $C(\theta_0 \pm A - \theta)^2 \to 0$ as $\theta \to \pm \infty$.

(b) The probability of avoiding the second sample and the expected percentage of overall sample saved.

If we use P.T. estimator, then the probability of avoiding the second sample is given by

$$P = \Pr(\hat{\theta}_1 \in R) = \int_{\theta_0 - A}^{\theta_0 + A} p_1(\hat{\theta}_1 | \theta_0) d\hat{\theta}_1 \tag{10.3.8}$$

and the expected percentage of overall sample saved is

$$[n_2/(n_1 + n_2)] P \times 100, \tag{10.3.9}$$

when $\theta = \theta_0$ is true.

Katti's [22] results (3), (5) and (6) about the estimation of mean from double samples follow from the results (10.2.4), (10.3.1) and (10.3.3) of the present chapter. However, the results (5) and (6) of Katti [22] contain error or misprint; there should be $1/\sqrt{2+u}$ in place of $1/\sqrt{1+u}$.

10.4 Application to the estimation of the variance of a normal population

Let x_{1i}, $(i = 1, 2, \ldots, n_1)$ be the first sample and x_{2i}, $(i = 1, 2, \ldots, n_2)$ be the second sample. It is assumed here that x ' s are all independently normally distributed with variance σ^2. Let σ_0^2 be the guessed estimate of σ^2.

Let $s_1^2 = \sum_{i=1}^{n_1} (x_{1i} - \bar{x}_1)^2 / (n_1-1)$ and $s_2^2 = \sum_{i=1}^{n_2} (x_{2i} - \bar{x}_2)^2 / (n_2 - 1)$ be uniformly minimum variance unbiased estimators of σ^2 based respectively on the first and second samples.

Let $s^2 = \left[(n_1-1)s_1^2 + (n_2-1)s_2^2\right] / (n_1+n_2-2)$ be the best estimator of σ^2 based on both the samples. Let the preliminary test (P.T.) estimator of σ^2 be denoted by σ_P^2. Noting that $a = (n_1-1)/(n_1+n_2-2)$, $b = (n_2-1)/(n_1+n_2-2)$ and $V(s_2^2 | \sigma_0^2) = 2\sigma_0^4 / (n_2-1)$, it follows from (10.2.3) and (10.2.4) that the P.T. region and P.T. estimator are given by

$$(10.4.1) \quad R = \left[\sigma_0^2\left(1-\sqrt{2/(2n_1+n_2-3)}\right), \sigma_0^2\left(1+\sqrt{2/(2n_1+n_2-3)}\right)\right],$$

$$(10.4.2) \quad \begin{aligned} \sigma_P^2 &= s_1^2, \text{ if } s_1^2 \in R, \\ &= s^2, \text{ if } s_1^2 \in \bar{R}. \end{aligned}$$

(a) Efficiency of σ_P^2 when σ_0^2 is the true value of ^{2}.

When σ_0^2 is the true value of the variance, then it follows from (10.3.1) that E.M.S. of σ_P^2 is given by

$$\text{(10.4.3)} \qquad \text{E.M.S.}(\sigma_P^2 | \sigma_0^2) = \left[2\sigma_0^4 / V_1(1+u)\right] \times T,$$

where

$$T = \left[1 + \frac{u(u+2)(V_1+2)}{2(u+1)} \left\{Q(t_1|V_1+4) - Q(t_2|V_1+4)\right\} - \frac{(u+2)V_2}{(u+1)} \left\{Q(t_1|V_1+2) - Q(t_2|V_1+2)\right\} + \frac{u(2V_1 + V_2 - 2)}{2(u+1)} \left\{Q(t_1|V_1) - Q(t_2|V_1)\right\}\right],$$

$V_1 = n_1-1$, $V_2 = n_2-1$, $u = V_2/V_1$, $t_1 = V_1\left[1-\sqrt{2/(2V_1+V_2)}\right]$, $t_2 = V_1\left[1+\sqrt{2/(2V_1+V_2)}\right]$ and

$$Q(t|V) = 2^{-V/2}(\Gamma V/2)^{-1} \int_t^\infty (\chi^2)^{\frac{V}{2}-1} e^{-\chi^2/2} d\chi^2,$$

which is tabulated in [32].

Noting that the variance of the best estimator s^2

based on both the samples is $2\,\sigma_0^4 / (n_1 + n_2 - 2)$, when σ_0^2 is the true value of σ^2, the efficiency $E(\sigma_P^2|\sigma_0^2)$ of σ_P^2 when σ_0^2 is the true value of σ^2 is from (10.3.3) given by

$$(10.4.4) \qquad E(\sigma_P^2|\sigma_0^2) = 1/T,$$

where T has the same meaning as in (10.4.3).

In Table 10.4.1, the efficiency $E(\sigma_P^2|\sigma_0^2)$ of σ_P^2 when σ_0^2 is the true value of the variance, is given for various values of u and V_1. That the efficiency is greater than 1 follows from the corollery 10.3.1. Table 10.4.1 confirms this proved result. This shows that when the guessed value σ_0^2 is the true value of σ^2, then the P.T. estimator σ_P^2 is better than the best estimator s^2 based on both the samples. The efficiency is maximum in the neighbourhood of $u = 2.5$ and the maximum average value of efficiency is near to 1.211 which implies that for maximum efficiency one should plan the sizes of the first and second samples in such a way that the degrees of freedom for the unbiased estimator of population variance in the second sample is nearly 2.5 times that in the first sample.

TABLE 10.4.1

EFFICIENCY OF σ_P^2 WHEN σ_0^2 IS THE TRUE VALUE OF THE VARIANCE

V_1	Values of u												
	0	.5	1	1.5	2	2.5	3	5	10	15	20	50	100
6	1	1.121	1.173	1.198	1.207	1.211	1.209	1.195	1.157	1.139	1.119	1.077	1.060
8	1	1.120	1.176	1.198	1.207	1.211	1.210	1.196	1.161	1.138	1.120	1.074	1.040
10	1	1.122	1.174	1.201	1.208	1.211	1.209	1.198	1.165	1.138	1.132	1.079	1.059
20	1	1.120	1.175	1.198	1.209	1.212	1.211	1.197	1.163	1.140	1.129	1.090	1.066
50	1	1.120	1.175	1.199	1.210	1.211	1.209	1.203	1.155	1.118	1.088	1.069	1.045
Average value	1	1.121	1.175	1.199	1.208	1.211	1.210	1.198	1.160	1.135	1.116	1.078	1.054

We also note that the efficiency remains fairly constant for fixed value of u. From (10.4.4), it follows that the efficiency tends to 1 as $u \to \infty$. The numerical results of Table 10.4.1 confirm this result.

(b) The probability of avoiding the second sample and the expected percentage of overall sample saved.

When the guessed value σ_0^2 is equal to the true value σ^2, the probability of avoiding the second sample and accepting s_1^2 as the estimate of σ^2 is obtained from (10.3.8) and (10.4.1) as

$$\begin{aligned} P = \Pr(s_1^2 \in R) &= \int_{t_1}^{t_2} p_1(s_1^2 | \sigma_0^2) ds_1^2 \\ &= Q(t_1 | V_1) - Q(t_2 | V_1), \end{aligned} \qquad (10.4.5)$$

where t_1, t_2, Q and V_1 have the same meanings as in (10.4.3). The expected percentage of overall sample in saved in this case is from (10.3.7) given by

$$[(uV_1 + 1)P \times 100] \; / \; [V_1(u + 1) + 2], \qquad (10.4.6)$$

where V_1 and u have the same meanings as in (10.4.3) in Table 10.4.2, we give the expected percentage of

TABLE 10.4.2

EXPECTED PERCENTAGE OF OVERALL SAMPLE SAVED IF P.T. ESTIMATOR IS USED

V_1	Values of u								
	1	2	3	5	10	15	20	50	100
6	21.93	24.82	25.09	23.75	20.00	17.59	15.64	10.49	7.54
8	21.94	25.00	25.28	24.06	20.09	17.58	15.84	10.50	7.72
10	21.86	25.14	25.37	24.04	20.29	17.56	15.83	10.61	7.70
20	21.96	25.29	25.69	24.38	20.52	17.92	15.86	10.88	7.72
50	21.75	25.41	25.66	24.46	20.67	17.82	16.12	10.89	7.86
Average value	21.89	25.13	25.42	24.14	20.31	17.69	15.86	10.67	7.71

overall sample saved for various values of u and V_1, if P.T. estimator is used.

From Table 10.4.2, we observe that if P.T. estimator is used, then the expected percentage of overall sample saved (i) remains fairly constant for fixed value of u and (ii) is largest when u is in the neighbourhood of 3. This implies that if it is possible to take a sample of size 102, then for best results, one should take a sample of size 26 to start with, followed by another of size 76 only if the preliminary test rejects the guessed value.

(c) Efficiency of σ_P^2 when σ_0^2 is not the true value of σ^2.

When σ_0^2 is not the true value of the variance, then from (10.3.2), we obtain E.M.S. of σ_P^2 as

$$\text{(10.4.7)} \qquad \text{E.M.S.}(\sigma_P^2 \mid \sigma^2) = [2\,\sigma^4 / V_1(1+u)]T',$$

where

$$T' = \Big[1 + \frac{u(u+2)(V_1+2)}{2(u+1)} \Big\{Q(t_1k \mid V_1+4) - Q(t_2k \mid V_1+4)\Big\}$$

$$- \frac{(u+2)V_2}{(u+1)} \Big\{ Q(t_1k \mid V_1+2) - Q(t_2k \mid V_1+2) \Big\} +$$

$$+ \frac{u(2V_1+V_2-2)}{2(u+1)} \left\{ Q(t_1k|V_1)-Q(t_2k|V_1) \right\} \Bigg] ,$$

where V_1, V_2, t_1, t_2, u and Q have the same meanings as in (10.4.3) and $k= \sigma_0^2 / \sigma^2$. Noting that the variance of the best estimator s^2 of σ^2 based on both the samples is $2\sigma^4 / (n_1 + n_2 - 2)$, the efficiency $E(\sigma_P^2 | \sigma^2)$ of σ_P^2 when σ_0^2 is not the true value of variance is from (10.3.3) given by

$$(10.4.8) \qquad E(\sigma_P^2 \mid \sigma^2) = 1/T',$$

where T' has the same meaning as in (10.4.7).

The behaviour of the function $E(\sigma_P^2 \mid \sigma^2)$ for various values of $k = \sigma_0^2 / \sigma^2$ and $V_1 = 6$, 10 and 20 was studied through graphs and it was found that the efficiency is greater than unity as $k \to \infty$. In all the three cases, there is an interval within which if k lies the bad guess will result in a heavy loss. It is believed that these observations will be true in the general case as well.