STUDIES ON SOME PRELIMINARY TEST ESTIMATORS IN DOUBLE SAMPLING
(ABSTRACT)

by

Mr. Phrangstone Khongji

Department of Statistics

Submitted
In partial fulfillment of the requirement of the degree of
Doctor of Philosophy in Statistics
of North-Eastern Hill University
Shillong
(2009)

1.1 Introduction

It is a well known fact that for estimating the population mean μ_y of a random variable Y, precision of the estimator can be increased when information on an auxiliary variable X, highly correlated with Y is readily available on all the units of the population. When the relationship between Y and X is found to be approximately linear but does not pass through the origin, linear regression estimate may be used, which is given by (Cochran,1977)

$$t_1 = \bar{y} + b(\mu_x - \bar{x}) \qquad \ldots(1.1)$$

where b is an estimate of the change in y when x is increased by unity, $\bar{y}$ and $\bar{x}$ are the sample means of the variables Y and X respectively and μ_x is the population mean of X .

The regression estimate given in (1.1) require advance knowledge about μ_x, the population mean of the auxiliary variable X. When such information is lacking, it is sometimes relatively cheaper to take a large preliminary sample in which x_i alone is measured and used for estimating the population characteristic like mean, total or frequency distribution of x values. The purpose of this sample is to furnish a good estimate of μ_x or of the frequency distribution of x_i. Another independent or sub-sample to observe both (x_i, y_i) meant to estimate $(\bar{x}, \bar{y})$ for using it in the regression estimator. This technique is known as double sampling and was first formulated by Neyman (1938) in connection with the collection of information on strata sizes in a stratified sampling experiment.

1.2 Double sampling with partial information on auxiliary variables

To use the linear regression estimator t_1 it is usually assumed that population mean μ_x is known. However in certain practical situation, μ_x is not known a prior, in which case the technique of double sampling is applied. In the first preliminary sample of size n', we measure only x_i and use it for the estimation of μ_x ; in the second sample , a random sub sample of size n ($<n'$), from the preliminary sample we observed both x_i and y_i . Under double sampling the regression estimate (1.1) becomes

$$t_2 = \bar{y}_n + b(\bar{x}_{n'} - \bar{x}_n) \qquad \ldots\ldots\ldots\ldots\ldots \text{(1.2)}$$

where $\bar{x}_{n'}$ is the mean of x_i in the preliminary sample and $(\bar{y}_n, \bar{x}_n)$ are the means of x_i and y_i from the sub sample and b is the least square regression coefficient of Y on X which can be computed from the sub sample.

Han(1973) described that the precision of an estimator can be improved if auxiliary variables are used in a regression estimator based on double sampling with partial information on auxiliary variable. Sometimes there are situations where we have partial information about the mean μ_x of the auxiliary variable X. In order to utilize the partial information Han(1973) suggested the use of a preliminary test and constructed a preliminary test estimator using double sampling with partial information on the auxiliary variable as follows ;

$$t_3 = \begin{cases} (\bar{y}_n - \rho\bar{x}_n) & \text{if} \quad \left|\bar{x}_{n'}\right| \le Z_\alpha / \sqrt{n'} \\ (\bar{y}_n + \rho(\bar{x}_{n'} - \bar{x}_n)) & \text{if} \quad \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'} \end{cases} \qquad \ldots\ldots\ldots\ldots \text{(1.3)}$$

In estimating the population mean μ_y of the random variable Y, suppose that in addition to information on an auxiliary variable X, information on yet another auxiliary variable Z is available. When populations means μ_x and μ_z are not available, one can take a preliminary sample to estimate these by the use of double sampling. In such a situation an estimator using X and Z is being suggested by Mukerjee *et.al* (1987).

Das(1992), Das and Bez (1995) and Das(2003) suggested some preliminary test estimators for the population mean in double sampling with two auxiliary variables, alternative to the usual regression estimator, when the experimenter has partial information on one and /or both the auxiliary variables.

The present work is aimed to proceed in accordance to further enhance the work done by Han (1973) and Das(1995) and several other authors to find an appropriate estimator through the use of preliminary test estimation and double sampling procedures.

1.3 The combined regression preliminary test estimator(CRPTE) in double sampling.

It is known that stratified sampling consists of classifying the population units in a certain number of groups called strata and selecting samples independently from each group. The division of population into strata can be done in such a way that the values of the study variable are homogeneous within each stratum , in that the measurement vary little from one unit to another, a precise estimates of any stratum mean can be obtained from a small sample in that stratum. These estimates with best choices of sample sizes can

be combined into a precise estimate for the whole population. When appropriately used, the variance of the estimated mean of the study variable Y under stratification is usually less than that of the variance under simple random sampling (Cochran 1977).

Stratification can also be operationally convenient and economical if the sampling frame is available in the form of sub-frames. Stratification enables that the demarcation of the strata boundaries and the allocation of the total sample size to the strata may be done so as to make the estimator most efficient from the point of view of sampling variability and cost. Though the main advantage of using stratified sampling is the possible increase in efficiency per unit of cost in estimating the population characteristics, the method is also useful in situation when estimators are required with specific margins of errors not only for population as a whole but for certain groups of units.

Cochran(1977) suggested a regression estimate in stratified sampling which he called a combined regression estimator and is given by

$\bar{y}_{lrc} = \bar{y}_{st} + b(\mu_x - \bar{x}_{st})$, where

$\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

In this estimate the whole population is stratified into different classes and samples are selected from each stratum by simple random sampling and the stratum means are combined and used in a regression equation to obtain the desired mean. Here b is the estimate of combined regression coefficient and W_h is the stratum weight.

The combined linear regression estimator given by $\bar{y}_{lrc}$, can be utilized under three situations. Firstly when the population mean μ_x is known, as a consequence of which the study reduces to usual combined regression method of estimation. Secondly in certain practical situations μ_x is not known a prior, in which case the technique of double sampling can be applied wherein a preliminary sample is obtained to estimate μ_x and the estimator of μ_y is given by

$$t_4 = \bar{y}_{st} + b(\bar{x}_{n'} - \bar{x}_{st}) \text{ , where} \qquad \text{.....(1.4)}$$

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h \quad \text{and} \quad \bar{x}_{st} = \sum_h W_h \bar{x}_h$$

Here $\bar{x}_{n'}$ is the value of the mean of X obtained from the preliminary sample and is utilized to estimate μ_x. Thirdly when μ_x is partially known, then a preliminary test estimation using double sampling procedure can be used.

In the present study, the third case will be considered where partial information about the mean of the auxiliary variable will be used. The first sample is a stratified simple random sample of size n in which the pair (x_{hi}, y_{hi}) values are measured from n_h units drawn from each stratum and consequently estimating of the pair $(\bar{x}_{st}, \bar{y}_{st})$, with $n = \sum_h n_h$ .The second sample is a larger simple random sample of size $n'(=n+m)$ is obtained by supplementing m more independent observations on X where only x_i is measured and evaluates $\bar{x}_{n'}$ which is utilized to estimate μ_x.

In order to utilize the partial information, a preliminary test is done about the hypothesis

$$H_o : \mu_x = \mu_o \ , \quad \text{against} \qquad H_1 : \mu_x \neq \mu_0$$

where μ_0 is the value obtained from the partial information.

If H_0 is accepted then μ_0 is used to replace μ_x in the regression estimator $\bar{y}_{lrc}$ and if H_0 is rejected then the sample mean $\bar{x}_{n'}$ based on the preliminary sample is used.

We assume that the auxiliary variable X and the study variable Y and are jointly normally distributed with parameters given by $(\mu_x, \mu_y, \sigma_x^2, \sigma_y{}^2, \rho)$. The marginal distributions which is the distribution of the study variable Y and the auxiliary variable X will also follow normal distribution given as Y ~ $N(\mu_y, \sigma_y^2)$ and X ~ $N(\mu_x, \sigma_x^2)$. The strata population (X_h, Y_h) being carved out from the parent population are also jointly assumed to follow the bivariate normal distribution with parameters written as $\left(\mu_{x_h}, \mu_{y_h}, \sigma_{x_h}{}^2, \sigma_{y_h}{}^2, \rho\right)$. Also since the relationship between the pair (X, Y) is always maintained even within the stratum the strata correlations are assumed to be equal to the population correlation coefficient ρ . The regression estimator depends on weather the covariance matrix is known or not. If known, one may let $\sigma_x{}^2 = \sigma_y{}^2 = 1$ without loss of generality (WLOG).

Since the population is assumed to follow normal distribution, the preliminary sample utilize to collect information on the auxiliary variable for the

estimation of $\bar{x}_{n'}$, is also assume to follow normal distribution and therefore $\bar{x}_{n'} \sim N(\mu_x, \sigma_x^2/n')$ and under the assumption $\sigma_x{}^2 = \sigma_y{}^2 = 1$, $\bar{x}_{n'} \sim N(\mu_x, 1/n')$. Further marginal distributions of X_h and Y_h are also normal given as $X_h \sim N(\mu_{x_h}, \sigma_{x_h}^2)$ and $Y_n \sim N(\mu_{y_h}, \sigma_{y_h}^2)$.

For each stratum, the pair of variables (X_h, Y_h) for every h, follows a bivariate normal distribution with mean (μ_{x_h}, μ_{y_h}) and covariance matrix given by

$$\Sigma_h = \begin{pmatrix} \sigma_{x_h}{}^2 & \rho\sigma_{x_h}\sigma_{y_h} \\ \rho\sigma_{x_h}\sigma_{y_h} & \sigma_{y_h}{}^2 \end{pmatrix}$$

The regression estimator depends on whether Σ_h is known or not. If Σ_h is known, one may let $\sigma_{x_h}{}^2 = \sigma_{y_h}{}^2 = 1$, (WLOG).

Also, the stratum means are given by

$$\bar{x}_h = \sum_{i=1}^{n_h} x_{hi}/n_h \quad \text{and} \quad \bar{y}_h = \sum_{i=1}^{n_h} y_{hi}/n_h$$

are linear combinations of normally distributed random variables (X_h, Y_h)

Hence it can be easily observed that $\bar{x}_h$ and $\bar{y}_h$ also follow normal distribution with mean and variances given by

$$\bar{x}_h \sim N(\mu_{x_h}, \sigma_{x_h}^2/n_h) \quad \text{and} \quad \bar{y}_h \sim N(\mu_{y_h}, \sigma_{y_h}^2/n_h)$$

i.e. $\bar{x}_h \sim N(\mu_{x_h}, 1/n_h)$ and $\bar{y}_h \sim N(\mu_{y_h}, 1/n_h)$

under the assumption of $\sigma_{x_h}{}^2 = \sigma_{y_h}{}^2 = 1$

The joint distribution of $(\bar{x}_h, \bar{y}_h)$ is bivariate normal with mean as (μ_{x_h}, μ_{y_h}) and covariance matrix given by

$$\Sigma_c = \begin{pmatrix} \sigma_{x_h}^2/n_h & \rho\sigma_{x_h}\sigma_{y_h}/n_h \\ \rho\sigma_{x_h}\sigma_{y_h}/n_h & \sigma_{y_h}^2/n_h \end{pmatrix} = \begin{pmatrix} 1/n_h & \rho/n_h \\ \rho/n_h & 1/n_h \end{pmatrix}$$

In certain situations, the experimenter may have partial information about μ_x. In order to utilize the partial information, one can perform a preliminary test about the hypothesis

$$H_o : \mu_x = \mu_o \ , \quad \text{against} \quad H_1 : \mu_x \neq \mu_0$$

where μ_0 is the value obtained from the partial information and $\bar{x}_{n'}$ is the value of the mean of X obtained from the preliminary sample through the use of double sampling.

Now when μ_x is partially known, one can let $\mu_o = 0$ (WLOG), so that the hypothesis can be accepted, when

$$\left|(\bar{x}_{n'} - \mu_o)/SE(\bar{x}_{n'})\right| \leq Z_\alpha \qquad \Rightarrow \left|\bar{x}_{n'}\right| \leq Z_\alpha/\sqrt{n'}$$

where Z_α is the 100(1-α/2)% point of N (0,1) and α is the level of significance of the preliminary test.

Under the above assumption the CRPTE in double sampling having partial information on the auxiliary variable X can be written as

$$t_5 = \begin{cases} (\bar{y}_{st} - \rho\bar{x}_{st}) & if \quad \left|\bar{x}_{n'}\right| \leq Z_\alpha/\sqrt{n'} \\ (\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) & if \quad \left|\bar{x}_{n'}\right| > Z_\alpha/\sqrt{n'} \end{cases} \quad \text{.....(1.5)}$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

and b from $\bar{y}_{lrc}$ reduces to $b=\rho(\sigma_y/\sigma_x)=\rho$ under the above assumptions.

1.4 Bias of the CRPTE in double sampling.

To evaluate the bias of t_5, we considered that the joint distribution of $(\bar{x}_{n'}, \bar{x}_{st}, \bar{y}_{st})$ which is a multivariate normal with mean (μ_x, μ_x, μ_y) and covariance matrix given by

$$\sum = \begin{pmatrix} Var(\bar{x}_{n'}) & Cov(\bar{x}_{n'}, \bar{x}_{st}) & Cov(\bar{x}_{n'}, \bar{y}_{st}) \\ Cov(\bar{x}_{st}, \bar{x}_{n'}) & Var(\bar{x}_{st}) & Cov(\bar{x}_{st}, \bar{y}_{st}) \\ Cov(\bar{y}_{st}, \bar{x}_{n'}) & Cov(\bar{y}_{st}, \bar{x}_{st}) & Var(\bar{y}_{st}) \end{pmatrix} \quad \text{.....(1.6)}$$

The derivation of the bias of the estimator t_5 involves conditional expectations, the condition being the acceptance or rejection of the hypothesis considered in the preliminary test. Further the expectations can be obtained from the integrals involving probability density functions which are assumed to be normal.

The variance of $\bar{x}_{st}$ and $\bar{y}_{st}$ being given by

$$\bar{x}_h \sim N(\mu_{x_h}, \sigma^2_{x_h} / n_h) \quad \text{and} \quad \bar{y}_h \sim N(\mu_{y_h}, \sigma^2_{y_h} / n_h)$$

i.e. $$\bar{x}_h \sim N(\mu_{x_h}, 1/n_h) \quad \text{and} \quad \bar{y}_h \sim N(\mu_{y_h}, 1/n_h)$$

under the assumption $\sigma_{x_h}{}^2 = \sigma_{y_h}{}^2 = 1$.

Also when the samples are selected with proportional allocation then the stratum weight is given by $W_h = (N_h / N) = (n_h / n)$

Thus $\sum_h W_h^2 / n_h = \sum_h W_h^2 / nW_h = (1/n)\sum_h W_h = (1/n)$ (as $\sum_h W_h = 1$)

Therefore the above covariance matrix in (1.6) reduces to

$$\sum = \begin{pmatrix} 1/n' & 1/n' & \rho/n' \\ 1/n' & 1/n & \rho/n \\ \rho/n' & \rho/n & 1/n \end{pmatrix} \quad \text{.....(1.7)}$$

The Bias of an estimator is defined as

$$Bias\ (t_5) = E(t_5) - \mu_y$$

where E(.) is the expectation

$$Bias\ (t_5) = E\left\{(\bar{y}_{st} - \rho\bar{x}_{st}) \text{..... } if \quad \left|\bar{x}_{n'}\right| \le Z_\alpha / \sqrt{n'}\right\} + E\left\{(\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) \text{..... } \text{....} if \quad \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'}\right\} - \mu_y$$

After routine derivation we get

$$Bias\ (t_5) = -\rho\mu_x\{\Phi(A) - \Phi(B)\} + \rho(1/\sqrt{n'})\{\varphi(A) - \varphi(B)\} \qquad \text{.....(1.8)}$$

where $\Phi(.)$ is the cumulative distribution function of N(0,1) and $\varphi(.)$ is its density function and $A = Z_\alpha - \sqrt{n'}\mu_x$, $B = -Z_\alpha - \sqrt{n'}\mu_x$.

The values of $Bias(t_5)$ can be easily computed for different values of μ_x. We notice that the bias is symmetric about $\mu_x = 0$, hence we need to consider only the case when $\mu_x \ge 0$. In order to get an idea about the behavior of the bias function with respect to μ_x, $Bias(t_5)$ is computed for a set of values of n, n', α and ρ. As μ_x is increases, the $Bias(t_5)$ increases to a maximum and then decreases to zero. This establishes the utility of the present study that the utilization of partial information and preliminary test of the auxiliary variable reduces the bias of the proposed estimator.

The above analytical method used for computing of the bias of the proposed estimator involves the evaluation of mathematical expectation of the random variables and consequently results in the computation of integrals. However, sometimes computation of integrals analytically may become cumbersome. Therefore an alternative method for the evaluation of the bias is

also sought with the help of numerical techniques. In the present study, attempt is made to evaluate the bias of the constructed estimator t_5 using numerical integration with programmes written in Fortran 77. When the results obtained numerically and that by analytical methods are compared, it is found that the bias function shows a similar pattern and difference wherever it exists is very negligible.

1.5 Mean square error of the CRPTE in double sampling.

The precision or a measure of the closeness of the sample estimates to the census count taken under identical conditions is judged in sampling theory by the variances of the estimators concerned. Here reliance is placed on the fact that with a small variance the probability of large deviations from the census count will be small. The general principle is to use estimators which will give the highest concentration of the sample estimates (in the sense of probability) around the valued aimed for. With unbiased estimators the method used for judging the degree of concentration is the variance of the estimators.

It may happen sometime that the degree of concentration of the sample around the valued aimed at is higher for the distribution of a biased estimator than for an unbiased one. In such a situation the biased estimator is preferable to the unbiased one. However in order to compare a biased estimator with an unbiased estimator, or two estimators with different amounts of bias, variance is not a satisfactory criterion, since it measures deviation from the expected value of the estimator, which is not the same as the population value. A useful

criterion is the mean square error (MSE) of the estimate, measured from the population value that is being estimated.

Formally, MSE(t) = Variance(t) + {Bias(t)}^{2}

To obtained MSE of t_5, we notice that

$$MSE\ (t_5) = \text{var}(t_5) + \{Bias\ (t_5)\}^2 \ = E(t_5^2) - \{E(t_5)\}^2 + \{Bias(t_5)\}^2$$

After routine derivations

$$MSE(t_5) = \{(1-\rho^2)/n + \rho^2/n'\} + (\rho^2/n')\{A\varphi(A) - B\varphi(B)\} - \rho^2(1/n' - \mu_x{}^2)\{\Phi(A) - \Phi(B)\} \quad \ldots.(1.9)$$

It is seen that the analytical method of determining the MSE involves evaluating the mathematical expectations of the random variables like $E(\bar{x}_n)$, $E(\bar{x}_{n'}{}^2)$, $E(\bar{x}_{n'}\bar{x}_{st})$ and $E(\bar{x}_{n'}\bar{y}_{st})$. The derivation of these expectation is done using moment generating function and also involves the application of single and double integration technique. In the process of evaluation which involves bivariate frequency distributions, a tedious substitution of variables is necessary to simplify the integrals. The above expectation is finally obtained by differentiating under the integral sign. As a consequence of the complexity involved in analytical deductions and the availability of numerical techniques, the MSE is evaluated using the numerical methods. In the numerical methods, the use of moment generating function and the substitutions involved can be avoided. The results of MSE obtained numerically show the similar pattern with that of the one derived by analytical methods for increasing values of the mean μ_x of the auxiliary variable. The differences in the values of MSE between analytical and numerical methods of computations are minimal.

1.6 Relative efficiency

The study of the mean square error of an estimator will not be completed unless it is compared with other estimators. Without the use of real life data, relative efficiency of an estimator can be obtained analytically as the ratio of the variance or mean square error of one estimator to that of the mean square error of the proposed estimator. If the relative efficiency is greater than 1, it can be concluded that the proposed estimator is more efficient in comparison to the other estimator.

In the present study, the mean square error of the proposed estimator is compared with other estimator t_4 and conclusion is drawn through the relative efficiency. Under similar assumptions a routine analysis gives

$$MSE\ (t_4) = (1/n)(1-\rho^2) + (1/n')\rho^2$$

Therefore the relative efficiency of t_5 to t_4 is given by

$$e_1 = [MSE(t_4)\ ] /\ [MSE(t_5)\ ]$$

In order to get an idea about the behavior of the relative efficiency function with respect to μ_x, e_1 is compute for a set of values of n, n', α and ρ. It is found in general that e_1 has a maximum at $\mu_x = 0$. This establishes the utility of the present study that the utilization of partial information and preliminary test increases the efficiency of the estimator. Graphically as μ_x is increases e_1 decreases to a minimum and then increases to unity. It is found that e_1 is very close to 1 at $\mu_x = 1$.

1.7 Optimum allocation.

In planning of a sample survey, a stage is always reached at which an important decision must be made about the size of the sample. Too large a sample implies a waste of resources, and too small a sample diminishes the precision of the estimators. Thus an optimum size of the sample is required so as to balance precision and cost involved in the survey. The optimum allocation of sample sizes are attained either by minimizing precision against a given cost or minimizing cost against given precision.

In obtaining optimum allocation of sample sizes for the proposed estimator, we consider a simple linear cost function C given by

$$C = c'n' + cn$$

where c is the cost per unit of observing the variable y and c' is the cost per unit of observing the variable x, assuming that the cost per unit is the same for all strata.

In general the values of μ_x are unknown, the experimenter has partial information about it. When $\mu_x = 0$, the mean square error of t_5 is least and the relative efficiency is largest. Thus it would be reasonable to let $\mu_x = 0$ in the $MSE\ (t_5)$ and obtain the values of n and n' under the optimum situation of minimizing precision against a given cost.

For a specific cost C* , routine mathematical derivation gives

$$M_{opt}(t_5) = \left[\sqrt{Kc} + \sqrt{c'K'}\right]^2 / C^*$$

In a similar way the optimum allocation for the estimator t_4 is given by

$$M_{opt}(t_4) = (\sqrt{Kc} + \sqrt{K''c'})/C^*$$

where $K = 1 - \rho^2$ $K' = \rho^2\{\alpha + 2Z_\alpha \varphi(Z_\alpha)\}$ and $K'' = \rho^2$

Analytically it can be seen that $(\alpha + 2Z_\alpha \varphi(Z_\alpha))$ is a decreasing function of Z_α with a maximum equal to unity at Z_α = 0.Therefore we can conclude that $M_{opt}(t_5) \leq MSE_{opt}(t_4)$ with equality holding for Z_α = 0 in which case the two estimators coincide.

Thus we have proved that under certain conditions, mean square error of a CRPTE in double sampling is smaller than the mean square error of combined regression estimator under double sampling.

1.8 Empirical Studies and Conclusion

Finally, in the last chapter empirical studies were made to show the applications of the proposed estimator as compared with other estimator. The empirical work is carried out with the help of both real life data as well as simulated data. Real life data were extracted from Rapid household survey – Reproductive child health (RHS-RCH project, phase 1, 1998).The data provides district wise demographic indicators for the Empowered Action Groups (EAG) States. From the data two distinct characteristics of the population were identified, namely complete child immunization and female literacy rate. Here the percent of complete child immunization is taken to be the dependent variable (Y) and female literacy rate is considered as an auxiliary variable (X). In the present analysis, the primary sample unit is set at district level and after

eliminating those districts in which data were not available partly or wholly the total size of the population reduces to N = 158. Since each state is governed by a distinct political and cultural system, which can have far reaching consequences on the economic, social status of the population within each state. This in turn can have an impact on the demographic characteristics of each state differently. So the data, on the selected variables can be homogeneous within each state and heterogeneous between states. Hence for the purpose the present study of the EAG states, each state is considered as a stratum. The data corresponding to different states and also the combined data are tested for normality and it was found that both X and Y follow normal distribution.

The partial information on X is obtained from census 1991 as female literacy rate is computed for the EAG states and given as $\mu_0 = 26.4$ %. After routine calculation, the estimates of the percentage of complete child immunization for the EAG states by the use of the combine regression preliminary test estimator in double sampling is 45.29%. As partial checks, the true population value of the mean of the auxiliary variable is 45%. Thus when reliable partial information on the mean of the auxiliary variable is not obtained as in present case, the hypothesis is rejected and as a result $\bar{x}_{n'}$ is utilized in the estimation of μ_x and consequently the proposed estimator reduces to the usual combined regression estimator under double sampling.

An attempt is also made to compute t_5 through the use of simulated data set. Bivariate normally distributed population data for different strata were

generated with the help of STATA 8.0 in which the input parameters are $(\mu_{x_h}, \mu_{y_h}, \sigma_{x_h}, \sigma_{y_h}, \rho)$. Four strata each of sizes N_1=35, N_2=40, N_3=50 and N_4=45 were considered and the stratification was done according to the mean of the value of the study variable Y. Further it is assume that there exist partial information about the mean of the auxiliary variable from certain sources and given by μ_0 = 75.0. Again routine calculation for the estimate μ_y the CRPTE in double sampling is 135.4. As partial checks, the true population value of the mean of the auxiliary variable is 73. Thus when a reliable partial information of the mean of the auxiliary variable is available, then the $MSE(t_5)$ is smaller than the $MSE(t_4)$ and consequently this increases the efficiency of the proposed estimator.

Thus we see that the empirical study also supports the analytical work of the present study that under the stated assumptions the CRPTE in double sampling is more efficient than the usual combined regression estimator, when a reliable information about the mean of the auxiliary variable is available.

References

Cochran, W.G. (1977). *Sampling Techniques,* Wiley Eastern Ltd. Third edition.

Das, G. (1992). Preliminary test estimators in double sampling with two auxiliary variables, Unpublished Ph.D. thesis, North Eastern Hill University, Shillong, India.

Das, G. (2003). A generalized study of preliminary test estimator in double sampling, *Assam Statistical review*, **17** (2), 127 - 138.

Das, G., and Bez, K. (1995). Preliminary test estimators In double sampling with two auxiliary variables, *Communications in statistics (Theory and Methods)*, **24** (5), 1211 - 1226.

Han, C.P. (1973). Double sampling with partial information on auxiliary variables. *Jour.Amer. Stat. Assoc.*, **68**, 914 - 918.

Mukerjee, R., Rao, T.J., and Vijayan,V.(1987). Regression type estimators using multiple auxiliary information. *Australian Journal of Statistics,* **29**, 244 - 254.

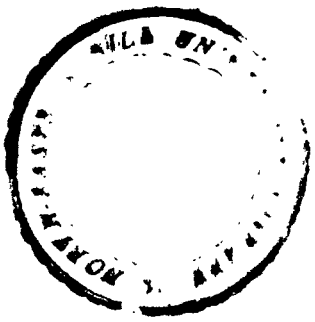
Neyman, J.(1938). Contribution to the theory of sampling human populations, *Jour.Amer. Stat. Assoc.*, **33**, 101 - 116.

STUDIES ON SOME PRELIMINARY TEST ESTIMATORS IN DOUBLE SAMPLING

by

Mr. Phrangstone Khongji

Department of Statistics



Submitted
In partial fulfillment of the requirement of the degree of
Doctor of Philosophy in Statistics
of North-Eastern Hill University
Shillong
(2009)

Declaration

North Eastern Hill University
December 2009

I, Mr. Pharangstone Khongji, hereby declare that the subject matter of this thesis is the record of work done by me, that the contents of this thesis did not form basis of the award of any previous degree to me or to the best of my knowledge to anybody else, and that the thesis has not been submitted by me for any research degree in any other University/Institute.

This is being submitted to North-Eastern Hill University for the degree of Doctor of Philosophy in Statistics.

18/12/2009

(Mr. Phrangstone Khongji)
Candidate

18/12/2009

(Prof. Gitasree Das)
Supervisor

Forwarded
Tkchakrabarty 18.12.2009
(T K. CHAKRABARTY)
Head
Department of Statistics

Acknowledgement

First of all, I would like to thank my supervisor Prof. G. Das for giving me the opportunity to pursue my Ph.D. work under her guidance in the Department of Statistics, NEHU. My serious interest in Sample Survey, which is my field of research, grew when I was a Post Graduate student at IIPS, Mumbai. The Institute is involved in large scale data collection on population studies and there we learnt about the different sampling techniques being used for these data collection.

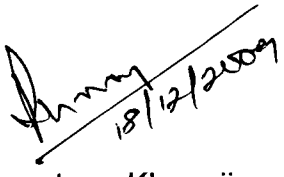
After my return from IIPS, I was eager to continue my studies to enhance my statistical skills in demographic studies. It was then that I came to the Center for Applied Statistics, NEHU (now a Department), where I met Prof. G.Das. We discussed at length about our common areas of interest and finally she agreed to let me work for a Ph.D. program at the University under her supervision and I am grateful to her for giving me this opportunity. She had been encouraging and helpful in every way in sharing ideas, giving valuable suggestions and comments throughout my research work.

To my family, I would say that it would be a dream comes true for my mother Smt. S.Khongji and father Mr. T.Diengdoh to see me completing my Ph.D. My mother has inspired, encouraged and supported me a lot for the same. I am also grateful to my wife and friend Mrs. S. Kharbhih for her co-operation and continuous moral support that has enabled me to complete this work.

My thanks also goes to Dr. M.K. Das, Dr. T.K. Chakrabarty and Dr. S.K. Jha from the Department of the Statistics, NEHU, who have encouraged and given valuable suggestions that enabled me to perfect my work. I express my gratefulness to all my colleagues and especially to Dr.A.K. Das, HoD in the Department of Basic Sciences and Social Sciences, NEHU, for their encouragement and support to be able to complete my work.

My sincere thanks goes to the library staff of the Delhi University who gave me the opportunity to access the library and from where I was able to obtain research articles related to my work. I am also grateful to Varghese, Deputy Librarian of the Indian Statistical Institute, Kolkata, for providing me with important research papers in connection with my work. I express my gratitude to the library staff of the Indian Agricultural Statistics Research Institute, New Delhi, who helped me to access the library and also reproduced the articles required by me for my research work. I am extremely grateful to the Library staff of my own University who had helped in every possible way, helped me to locate the journals, books, data and other important documents.

Lastly, my acknowledgement goes to the staff of the Department of Statistics, NEHU, Ms. Mariamma, Mr. Francis and Ms. Pershalyne and also to the staff of the Department of Basic Sciences and Social Sciences, NEHU, Ms. Donna, Mr. Steve and Mr. Maxwell for their kind co-operation.



December, 2009 Phrangstone Khongji

Contents

Chapter 3

Mean square error function of the CRPTE in double sampling with partial information on the auxiliary variable

Chapter 4

Relative efficiency and optimum allocation of the CRPTE in double sampling with partial information on the auxiliary variable

Chapter 5

Empirical studies based on the CRPTE in double sampling with partial information on the auxiliary variable

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

The purpose of sampling theory is to make sampling more efficient. It attempts to develop methods of sample selection and of estimation that provide, at the lowest possible cost, estimates that are precise enough for our purpose. So far as precision is concerned, one cannot foretell exactly how large an error will be present in an estimate in any specific situation, for this would require knowledge of the true value for the population. Instead, the precision of a sampling procedure is judged by examining the frequency distribution generated for the estimate if the procedure is applied again and again to the same population which is the standard technique by which precision is judged in statistical theory.

The precision or a measure of the closeness of the sample estimate to the census count taken under identical conditions is judged in sampling theory by the variance of the estimator concerned. Here reliance is placed on the fact that with a small variance the probability of large deviations from the census count will be small. The general principle is to use estimators which will give the highest concentration of the sample estimates (in the sense of probability) around the value that is aimed. With unbiased estimators the method used for judging the degree of concentration is the variance of the estimators.

It may happen sometimes that the degree of concentration of the sample around the value aimed at is higher for the distribution of a biased estimator than for an unbiased one. In such a situation the biased estimator is preferable to the unbiased one. However in order to compare a biased estimator with an

unbiased estimator, or two estimators with different amounts of bias, variance is not a satisfactory criterion, since it measures deviation from the expected value of the estimator, which is not the same as the population value. A useful criterion is the Mean Square Error (MSE) of the estimate, measured from the population value that is being estimated.

1.2 Auxiliary information

It is a well known fact that for estimating the population mean μ_y of a random variable Y, precision of the estimator can be increased when information on an auxiliary variable(s) is available. Murthy (1967), Cochran(1977) and Sukhatme (1984) have suggested number of estimation procedures using auxiliary information. Considerable amount of work has been done in the direction of providing improved estimators in survey sampling using auxiliary information.

For estimating the mean μ_y of a finite population utilizing the known population mean μ_x of an auxiliary variable X, Srivastava (1971) considered a class of estimators defined by

$$\bar{y}_h = \bar{y}\, h(\bar{x}/\bar{X}) \qquad \text{.....(1.1)}$$

where h(.) is a parametric function such that h(1) = 1, and $\bar{y}$ and $\bar{x}$ denote simple random sample means of the variables Y and X respectively. The lower bound for the asymptotic variance of estimators of the class (1.1) was shown to be equal to the asymptotic variance of the linear regression estimator.

Srivastava (1980) extended the class (1.1) of estimators to a much wider class and proved a similar result. A large class of estimators is considered for the mean of a finite population using information on an auxiliary variable. He has shown that members of this class of estimators are asymptotically no more efficient than the linear regression estimator.

Khare and Srivastava (1980) proposed an estimator using two auxiliary variables and also computed the expressions for the bias and variance. A comparison is made with the ratio, regression and product estimators. With an illustrated example, they have shown that the proposed estimator is more efficient than other estimators using two auxiliary variables.

Srivastava and Jhajj (1981) showed that for estimating the mean of a finite population using information on auxiliary variable, a class of estimators is defined as a function of the ratio of sample mean to population mean and the ratio of sample variance to population variance of the auxiliary variable. Asymptotic expression for the bias and mean square error are obtained. They also gave a condition for the minimum mean square error of estimators of this class to be smaller than those which used only the ratio of the sample mean to population mean of the auxiliary variable.

Isaki (1983) proposed and compared variance estimators under several sample design when auxiliary information is available. Improvement of the bias and mean square error over some estimators commonly used in practice is illustrated. He further presented a Monte Carlo comparison of the estimators.

Hidiroglou and Sarndal (1998) studied on two phase sampling designs that offer a variety of possibilities for use of auxiliary information when transformed into calibrated weights meant to construct efficient estimators of the population total. The calibration is done in two steps : (i) at the population level and (ii) at the level, of the first phase sample. The authors go on to show that the resulting calibrating estimators are also derivable via regression fitting in two steps. The estimators are examined for a special case of interest, namely, when auxiliary information is available for population sub groups called calibration groups. Estimation for domains of interest and variance estimation are also discussed. They illustrated the results by applying them to two important two phase designs at Statistics Canada. The general theory for using auxiliary information in two phase sampling is being incorporated into Statistics Canada's generalized estimation system.

Upadhayaya and Singh (1999) proposed an estimator using a transformed auxiliary variable. The bias and MSE of the proposed estimator have been obtained. The region of preference has been obtained under which it is better than the usual unbiased estimator $\bar{y}$, the ratio estimator $\bar{y}_R = \bar{y}(\bar{X}/\bar{x})$, Sisodia and Dwivedi (1981) estimator and Singh and Kakran (1993) estimator. An empirical study carried out by them also demonstrated the superiority of the suggested estimator over the others.

Zhang (1999) in his article discussed that in the nonparametric setting, the standard bootstrap method is based on the empirical distribution function of a random sample. The author proposes, by means of the empirical likelihood

technique, an alternative bootstrap procedure under a nonparametric model in which one has some auxiliary information about the population distribution. By proving the almost sure weak convergence of the modified bootstrapped empirical process, the validity of the proposed bootstrap procedure is established. This new result is used to obtain bootstrap confidence bands for the population distribution function and to perform the bootstrap Kolmogorov test in the presence of auxiliary information. Other applications include bootstrapping means and variances with auxiliary information. He also presented three simulation studies to demonstrate the performance of the proposed bootstrap procedure for small samples.

Jhajj, Sharma and Grover (2006) showed that in practice, the information regarding the population proportion possessing certain attribute is easily available. So, for estimating the population mean μ_y of study variable Y, a family of estimators of μ_y has been proposed by using the known information of population proportion possessing an attribute (highly correlated with Y). The expressions for the mean square error of the estimators of the proposed family and its minimum value have been obtained. It has been shown that the optimum estimator of the proposed family of estimators μ_y is always more efficient than the mean per unit estimator. They have also extended the result for the case of the double sampling design. The results obtained have been illustrated numerically by taking some empirical populations considered in the literature.

Shabbir and Gupta (2007) showed that a family of estimators introduced by Jhajj *et al* (2006) introduced a family of estimators by using the known population proportion of elements possessing an attribute in simple random sampling and two phase sampling. The efficiency of this estimator is more than mean per unit estimator and is equal to the efficiency of the linear regression estimator. In this article, they proposed a ratio-type estimator by using the technique of Roy (2003). The proposed estimator has an improvement over mean per unit estimator as well as Jhajj *et al* (2006) estimator in simple random sampling and two phase sampling. The authors use three data sets to examine the improvement in efficiency.

1.3 Regression estimators

It is a well known fact that for estimating the population mean μ_y of a random variable Y, precision of the estimator can be increased when information on an auxiliary variable X, highly correlated with Y is readily available on all the units of the population. When the relationship between Y and X is found to be approximately linear but does not pass through the origin, linear regression estimate may be used. The simplest form of the linear regression estimate of μ_y, the population mean of Y is given by

$$\bar{y}_{lr} = \bar{y} + b(\mu_x - \bar{x}) \qquad \text{.....(1.2)}$$

where μ_x is the population mean of X and b is an estimate of the regression coefficient of Y on X.

Williams (1963) formulated a conditionally unbiased estimator of μ_y for the cases of one auxiliary variable and that of multivariate auxiliary situations also. It is observed that the variance of the proposed estimator T_k depends upon the form selected for the coefficients. While it is true that analytical bias statements that can be made about the usual regression estimator $\bar{y}_{lr}$ are only moderately enlightening, for example that it is $o(n^{-1})$, calculations that have been carried out by the author seem to reveal it is remarkably well behaved even for badly nonlinear data. For moderate and large sample sizes the estimator $\bar{y}_{lr}$ seems to hold a strong position even on the basis of mean square error.

Gunst and Mason (1977) used mean square error criteria to compare five estimators of the coefficients in a linear regression model viz, least square, principle components, ridge regression, latent root and a shrunken estimator. Each of the biased estimators is shown to offer improvement in mean square error over least squares for a wide range of choices of the parameter of the model. The results of a simulation involving all five estimators indicate that the principle components and latent root estimators performs best overall, but the ridge regression estimator has the potential of a smaller mean square error than either of these.

Singh and Srivastava (1980) proposed a sampling scheme for which the usual regression estimator is unbiased. Another sampling scheme with an unbiased regression-type estimator is also considered. On comparing the efficiencies of these sampling strategies with some existing strategies, they

found that the performance of the first of the new scheme is found to be highly satisfactory.

Ahmed (1998) showed that Kiregyera (1984), Mukerjee, Rao and Vijayan (1987), Tripathi and Ahmed(1995) considered a number of regression-type estimators where information on two auxiliary variables related to study variable is available at different levels. Mukerjee *et al* (1987) suggested three estimators and computed their mean square errors, but the computations seem to be incorrect. Ahmed's (1998) note corrects them, and finds their estimators are no better than that of Kiregyera (1984). The estimator suggested by Tripathi and Ahmed (1995) is the best in the sense of having the smallest mean square error.

Singh (2006) investigated a general set-up for inference from survey data that covers the estimation of variance of estimators of totals and distribution functions, using known first and second order moments of auxiliary information at the estimation stage. The traditional linear regression estimator of population total owed to Hansen, Hurwitz and Madow (1953) is shown to be unique in its class of estimators, and celebrates Golden Jubilee Year-2003 for its outstanding performance. Singh(2006) further states that although there is no need of simulation study to demonstrate the superiority of the proposed technique, because the theoretical results are crystal clear, but a small scale level simulation study have been designed to show the performance of the proposed estimators over the existing estimators in the literature.

1.4 Double sampling

A number of sampling procedures require advance knowledge about population mean of an auxiliary variable. When such information is lacking, it is sometimes relatively cheaper to take a large preliminary sample in which the auxiliary variable alone is measured and which is used for estimating the population characteristic like mean, total or frequency distribution of X values. The main sample is often a sub-sample of the initial sample but may be selected independently as well. Both X and Y are measured on the units in the main sample and estimates of population total (mean) of Y is obtained. This technique is known as double sampling or two phase sampling and was first formulated by Neyman (1938). These techniques are being widely discussed by Murthy (1967), Cochran (1977) and Sukhatme (1984).

Srivastava (1981) suggested a large class of ratio and product-type estimators in two-phase sampling. He showed that the asymptotic minimum variance for any estimator of this class is equal to that of the conventional linear regression estimator.

Sahoo and Swain (1989) suggested a sampling scheme for double sampling under two cases; one when the second phase sample is a sub-sample of the first phase sample and another when the second phase sample is independent of the first phase sample. They then considered two estimators and compared their efficiencies for both cases.

Sarndal and Swensson (1987) have given a general framework of estimation in two phase sampling assuming arbitrary probability of sampling

designs in both the phases where the selections are done without replacement. Srivenkataramana and Tracy (1989) in their paper discussed a specific with replacement version of the scheme where a ratio of two variables is suggested as a measure of size of the units and a mean of a ratio type estimator. A comparison is made with Raj (1965) scheme employing the difference method of estimation. The paper in particular illustrates the idea of using the ratio x/z as a measure of size and then using the mean of the ratio type estimator.

Sahoo and Sahoo (1999) showed that, if the experimenter knows the population mean of an additional auxiliary variable Z, and the population mean of the main auxiliary variable X is unknown, and is estimated by the use of double sampling procedures, then it is possible to formulate a class of estimator for the finite population mean μ_y.

Diana and Tommasi (2004) discussed that double sampling scheme is used when cheap auxiliary variables may be measured to improve the estimation of a finite population parameter. Several estimators for population mean, ratio of means and variance are available, when two dependent samples are drawn. In this paper both cases of dependent and independent samples are dealt with. A general approach for estimating a finite population parameter is given, showing that all the proposed estimators are particular cases of the same general class. The minimum variance bound for any estimator in this class is provided (at the first order of approximation). Further, an optimal estimator which reaches this minimum is found.

Senapati and Sahoo (2006) used double sampling procedure, to present a general class of estimators for the finite population mean when the population mean of the main auxiliary variable X is unknown but that of an additional auxiliary variable Z is known. The proposed class of estimators is superior to some of the previously studied classes under minimum variance criterion.

1.5 Double sampling for regression estimator

Ratio and regression estimates require the knowledge of the population mean μ_x. When such information is lacking, it is sometimes relatively cheaper to take a large preliminary sample in which x_i alone is measured and is used for estimating the population mean of x values. The purpose of this sample is to furnish a good estimate of μ_x. Another independent or sub-sample is drawn for measuring (x_i, y_i) to estimate $(\bar{x}, \bar{y})$ for being used in the regression estimator. The estimate of μ_y is given as ;

$$\bar{y}_{lr} = \bar{y} + b(\bar{x}' - \bar{x}) \qquad \text{.....(1.3)}$$

where $\bar{x}'$ and $\bar{x}$ are the means of x_i in the first (or preliminary sample) and second (or independent / sub-sample) , $\bar{y}$ is the mean of y_i and b is the least squares estimate of the regression coefficient of y_i on x_i computed from the second sample. Assuming random sampling

$$Var\ (\bar{y}_{lr}) = \{S_y^2(1 - \rho^2) / n\} + \{\rho^2 S_y^2 / n'\} - S_y^2 / N \qquad \text{(Cochran, 1977)}$$

Tamhane (1978) considered the problem of hypothesis testing using the regression estimator in double sampling. Test procedures are provided when the covariance matrix between the primary and the auxiliary variables is either partially known or completely unknown. For the latter case a new 'studentized' version of the regression estimator is proposed as a test statistic. The exact null distribution of this test statistic is derived in a special case. An approximation to the null distribution is derived in the general case and studied by means of Monte Carlo method. The problem of choosing between the double sampling regression estimator and the single sample mean estimator is also discussed.

Kiregyera (1984) used two auxiliary variables X and Z to construct two regression type estimators for the population mean of the study variable Y. The efficiency of the proposed estimator is investigated under a super-population model. A numerical study is done to demonstrate the practical use of different estimation formulae and empirically demonstrate the performance of the constructed estimators.

Mukerjee, Rao and Vijayan (1987) considered the problem of estimating the population mean of the study variable μ_y, when information on an auxiliary variable X highly correlated with Y is readily available on all units of the population. They mentioned that the ratio or regression type estimators could be used for increased efficiency by incorporating the knowledge of μ_x. The authors also state that, in certain practical situation μ_x is not known a priori in which case the technique of double sampling can be fruitfully applied. The values of X are assumed known over a large sample of size n'. They also

considered that in addition to the variable X, information on yet another auxiliary variable Z is available on all units of the population with mean μ_z. In this paper they observed that the expression for the MSE of the suggested estimator given by Kiregyera (1984) does not agree with theirs and then proceed to suggest a series of estimators which are quite normal to consider. Unlike Kiregyera (1984) made certain super population assumptions, the authors compared their estimators directly with respect to the MSE criterion and obtained the most suitable estimator in the present context. Next, they consider a more practical situation when Z is not known for the whole population but is known over a large sample of size n" and obtained the corresponding estimators and the comparison. They also gave a theoretical justification for the estimator chosen and extended the result to the case of multi auxiliary variables.

Mukherjee and Chaudhuri (1990) considered sampling a finite population in two phases with varying probabilities, choosing the first phase sample using known size measures and the second phase sub sample there from utilizing, in addition, ascertained auxiliary variate-values for the initial sample. Properties are investigated for proposed generalized regression estimators for the population mean of variable of interest with values observed only for the selected sub-sample. Referring to infinite sequences of finite populations with increasing sizes and postulating super population models, lower bounds for asymptotic model-expected design- 'mean square error' for the estimators are derived and optimal model based designs are characterized so as to attain these bounds.

Reich, Bonham and Kimberly (1993) performed a study conducted to determine whether a ratio or regression estimator should be used to estimate aboveground biomass of stands dominated by blue grama (*Bouteloua gracilis* (H.B.K.) Lag ex Steud.) in eastern Colorado. One hundred 0.25-m^{-1} circular plots were systematically located in a homogeneous stand of blue grama, and on each plot biomass was estimated visually and then clipped. Three methods (classical, jackknife, and bootstrap) of estimating the variance for double sampling with regression and ratio estimator were compared in a simulation study using sample sizes 10, 20, 30, 40, and 50 clipped plots. The ratio estimator consistently had smaller bias and is suggested to be used for estimating average clipped weight of blue grama. For n=10 clipped plots, the jackknife variance estimator is recommended for constructing confidence intervals. For n > 20 clipped plots, the classical variance estimate is suggested to be used to obtain reliable estimates of the population variance and in estimating confidence intervals.

Dorfman (1994) mentioned that double sampling of a finite population occurs when a sample from the population is itself sampled, with the intend of measuring variates in the sub sample not already available in the sample. An important example is the regression estimator of means or totals, which uses values of an auxiliary variable from the full sample to estimate the mean of a variable of interest that is available only on the sub sample. This article concerns estimation of the variance of the regression estimator. The estimator of variance recommended by Cochran(1977) rely solely on the sub sample

data, at least to first order. This article note proposes variance estimators that make better use of the entire sample.

Singh, Katyar and Gangwar (1996) proposed a class of almost unbiased regression-type estimators with the help of Jack Knife technique developed by Quenouille (1956) for simple random sampling in two phases. The mean square error / variance expression of the resulting estimator is derived. Optimum estimator in the proposed class of estimators is also investigated and its mean square error / variance is compared with the usual biased linear regression estimator and it is found that they are approximately the same.

Sitter (1997) discussed that survey sampling depends on the possession of information on an auxiliary variable X or a vector of the auxiliary variables available for the entire population. This article focused on the variance estimators in the regression estimator in the aforementioned context and their use in constructing confidence interval. A design based linearization variance estimator that make more complete use of the sample data than the standard one is considered for two phase sampling. A jackknife variance estimator and its linearized version are obtained and showed to be design consistent. A bootstrap variance estimator is also shown to be design consistent. Unconditional and conditional repeated sampling properties of these variance estimators are studied through simulation. It is also shown that the linearization variance estimator displays superior unconditional properties, but the jackknife and its linearized version performs better conditionally.

Okafor and Lee (2000) in their article showed that Cochran (1977) proposed some ratio and regression estimators of the population mean using the Hansen and Hurwitz (1946) procedure of sub sampling the non-respondents assuming that the population mean of the auxiliary character is known. For the case where the population mean of the auxiliary character is not known in advance, some double sampling ratio and regression estimators are presented in this article. The potentially serious non response bias is eliminated by sub sampling the non respondents. The relative performances of the proposed estimators are compared with the estimator proposed by Hansen and Hurwitz (1946). The authors also derived optimum sample sizes for a given set of unit costs and compared theoretically and empirically the performance of the proposed estimators with that of the Hansen and Hurwitz(1946) estimator.

Roy (2003) considered the problems in the event of two phase sampling. He constructed a regression type estimator of the population mean of the study variable Y in the presence of available knowledge on the second auxiliary variable Z, when the population mean of the first auxiliary variable X is not known. The proposed estimator is found to be more efficient than the usual two phase ratio and regression estimators and also two phase ratio type and regression type estimators using two auxiliary variable suggested by Mohanty (1967), Chand (1975), Kiregyara (1980,1984) and Sahoo *et.al* (1993).

Pradhan (2005) considered the situation when the population mean of the study variable Y is estimated in a two phase sampling setup using three auxiliary variables with chain regression concept when the population mean of

one of the auxiliary variables is unknown and other auxiliary population means are known.

Samiuddin and Hanif (2007) considered estimation of the population mean of the variable of main interest in single and two phase sampling with or without additional information available in auxiliary variables. The class of ratio (including chain ratio) and regression estimators are further explored and expanded. Comparisons with other estimators suggested in the literature are also made.

1.6 Regression estimates in stratified sampling

It is known that in many of the large scale surveys, it is inevitable to adopt stratification for the purpose of preparing a frame from which the sample can be extracted. Stratification also produces gain in precision in the estimate of the characteristics of the whole population. If each stratum is homogeneous, in the sense that the measurement vary little from one another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum. These estimates can then be combined into a precise estimate for the whole population. Thus if intelligently used, stratification nearly always results in a smaller variance for the estimated mean or total that is given by a comparable simple random sample (Cochran, 1977).

If the resources are available for total sample of n units, they are distributed among the strata. One obvious choice is to divide the sample size proportionate to the stratum sizes. This type of stratification is described as

stratification with proportional allocation. Under certain conditions, proportional allocation can lead to gain in precision as compared to simple random sample.

Cochran (1977) has mentioned that as with ratio estimates, there are two types of regression estimates that can be constructed in stratified random sampling. In the first estimate $\bar{y}_{lrs}$, a separate regression estimate is computed for each stratum mean, that is,

$$\bar{y}_{lrh} = \bar{y}_h + b_h(\mu_x - \bar{x}_h), \qquad \text{.....(1.4)}$$

and with $W_h = (N_h / N)$

$$\bar{y}_{lrs} = \sum_h W_h \bar{y}_{lrh}$$

Where b_h is the within-stratum least square estimate of B_h .

It is stated that this estimate is appropriate when it is thought that the true regression coefficients B_h vary from stratum to stratum. In the second regression estimate $\bar{y}_{lrc}$, the combined regression estimate is computed as

$$\bar{y}_{lrc} = \bar{y}_{st} + b(\mu_x - \bar{x}_{st}) \quad \text{, where} \quad \text{.....(1.5)}$$

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h \quad \text{and} \quad \bar{x}_{st} = \sum_h W_h \bar{x}_h$$

b is an estimate of the regression coefficient of Y on X. Cochran states that $\bar{y}_{lrc}$ is appropriate when B_h are presumed to be the same in all strata.

Han (1990) showed that in stratified random sampling, one may construct either a separate regression estimator or a combined regression estimator for the population mean. The separate regression estimator is appropriate when the population regression coefficients are different from stratum to stratum; while the combined regression estimator is appropriate

when the strata regression coefficients are equal or not. In practice it may be uncertain whether the regression coefficients are equal. In such a case, one can use a preliminary test to test the equality of the population stratum regression coefficients. Then a preliminary test regression estimator can be constructed. Also one can use a weighted regression estimator which is a weighted average of the separate regression estimator and the combined regression estimator with the weights depending on the test statistic. A comparison of the various estimators is made by a Monte Carlo study.

Davidov and Chang (2002) provide a method for estimating the sample mean of a continuous outcome in a stratified population using a double sampling scheme. The stratified sample mean is a weighted average of stratum specific means. It is assumed that the fallible and true outcome data are related by a simple linear regression model in each stratum. The optimal stratified double sampling plan, i.e., the double sampling plan that minimizes the cost of sampling for fixed variances, or alternatively, minimizes the variance for fixed costs, is found and compared to a standard sampling plan. The design parameters are the total sample size and the number of doubly sampled units in each stratum. They further showed that the optimal double sampling plan is a function of the between-strata and within-strata cost and variance ratios. The efficiency gains, relative to standard sampling plans, under broad set of conditions, are considerable.

1.7 Allocation in stratified sampling

In planning of a sample survey, a stage is always reached at which a decision must be made about the size of the sample. This decision is important. Too large a sample implies a waste of resources, and too small a sample diminishes the precision of the estimators. Thus an optimum size of the sample is required, so as to balance precision and cost involved in the survey. The optimum allocation of sample sizes are attained either by minimizing precision against a given cost or minimizing cost against given precision.

Sukhatme and Tang (1975) proposed an allocation in stratified sampling subsequent to preliminary test of significance. They followed a procedure for allocation of sample sizes to different strata consisting of drawing a preliminary sample of fixed size from each stratum to estimate the strata variances and test their homogeneity. If the strata variances are found homogeneous, the sample sizes to be drawn from different strata are allocated according to proportional allocation; otherwise, they are allocated according to Neyman allocation using estimated variances. The efficiency of the proposed allocation based on preliminary test of significance with respect to proportional allocation and modified Neyman allocation is investigated.

Chernyak (2001) considers a stratified random sampling with the cost function given by $C = \sum_{k=1}^{L} c_k f(n_k)$ where, c_k is the cost per unit in the k^{th} stratum, and $f(x)$ be some function. The values of the sample sizes n_k in the respective strata are chosen by the sampler. They may be selected to minimize $Var(\bar{y}_{st})$ for a specific cost or to minimize the cost C for a specific value of $Var(\bar{y}_{st})$.

Under optimum condition, the size n of the sample is computed. Further the author considers the method of double sampling or two phase sampling where the first sample estimate the strata weights and the second sample estimates, the strata means $\bar{Y}_k$.In this case also, the sample size and the strata weights is computed by minimizing $Var(\bar{y}_{st})$ for a specific cost C or by minimizing the cost for a specific value V of $Var(\bar{y}_{st})$.The study were also supported by numerical computations.

1.8 Preliminary test estimators

One feature of theoretical statistics is the creation of a large body of theory that discusses how to make good estimates from data. In the development of theory specifically for sample surveys, relatively little use has been made of this knowledge. Cochran (1977) mentioned two principal reasons, the first of which he states that, in surveys that contain a large number of items, there is a great advantage, even with computers, in estimation procedures that require little more than simple addition. Whereas the superior methods of estimation in statistical theory, such as maximum likelihood, may necessitate a series of successive approximations before the estimate can be found. Secondly, he mentions that, there has been a difference in attitude in the two lines of research. Most of the estimation methods in theoretical statistics assumed that we know the functional form of the frequency distribution followed by the data in the sample, and the method of estimation is carefully geared to this type of distribution. The preference in sample survey theory has been to

make, at most, limited assumptions about this frequency distribution (that it is very skew or rather symmetrical) and to leave its specific functional form out of the discussion. This preference leads to the use of simple methods of estimation that work well under a range of types of frequency distributions. This attitude is a reasonable one for handling surveys in which the type of distribution may change from one item to another and when we do not wish to stop and examine all of them before deciding how to make each estimates.

In most of the work done, the estimator was constructed without the use of distribution theory. It is only afterwards that some researchers began to realize the importance of including the theory of distribution in the construction of estimator. Some of the authors who worked in this field are Han (1973) who published an important paper. In this paper he constructed a regression estimator to estimate the mean of the study variable μ_y, making use of the theory of distribution and by utilizing partial information on another auxiliary variable X. After Han (1973) some important works in this field are done by Sisodia (1981), Sisodia and Srivastava (1982), Das(1992), Das and Bez (1995), Das (2003) to name a few.

Introducing distribution theory complicates the mathematical derivation meant for in evaluating the bias and mean square error,etc., of the estimator as this involves complex integration, transformation of variables and summation of series. However an assumption about the distribution is worth considering, as it is conformable with realistic situation in which variables do follow some distribution or the other. Thus it goes without saying that without the use of

distribution theory, the process of construction of estimator becomes too simplistic.

Han (1973) in his paper described that the precision of an estimator can be improved in a regression model when a preliminary test is performed on mean of the auxiliary variable obtained from the preliminary sample in double sampling. He also considered that the study variable and the auxiliary variable jointly follow bivariate normal distribution. In a regression model the population mean μ_x of the auxiliary variable X is usually unknown. Han(1973) in his paper mentioned that there are situations in which partial information μ_0 about the mean of the auxiliary variable is available. In order to utilized the partial information about the mean of X, one can perform a preliminary test $H_0 : \mu_x = \mu_0$ against $H_1 : \mu_x \neq \mu_0$. If H_0 is accepted, μ_0 obtained from partial information will be used in the regression estimator and if H_0 is rejected, the sample mean $\bar{x}_{n'}$ based on the preliminary sample through double sampling is used. Based on the above assumptions he constructed a preliminary test estimator $\bar{y}_{lr}$. The bias, Mean square error and relative efficiency are obtained for the estimator.

Bock, Yancey and Judge (1973) were concerned with deriving the properties of the preliminary test estimator for general linear normal regression model, ascertaining the characteristics of the risk function over the parameter space, and determining the conditions necessary for the risk of this estimator to exceed or be less than the conditional one under squared error loss. A test procedure and the problem of choosing an optimal level of significance for the test are discussed. Some theorems and lemmas using in the evaluation the risk

of some properties of functions of the non central F distributions are developed in the appendices.

Johnson, Bancroft and Han (1977) showed that when two or more regression equations are the same, it is advantageous to pool the data for making inferences about the population regression model of interest. This paper derives the biases and mean square errors of estimation procedures subsequent to preliminary test for the cases of pooling two or more linear regression lines, and pooling two multiple regressions. Relative efficiency of the sometimes-pool predictor to the never-pool predictor are obtained, and recommendations of the levels of preliminary tests are made so that the efficiency of predictions with the final fitted regression will be at a prescribed level.

Pandey and Singh (1977) suggested some estimators of the variance of a normal population using prior information. These estimators are based on a guessed value of the population variance. The expression for the bias and mean square error has been obtained. These estimators have been compared with the usual unbiased estimators and found that the modified version of the preliminary test shrunken estimator has higher efficiency when sample size n is small.

Han (1978) in his paper studied nonnegative estimators of variance components involving preliminary tests. The restricted maximum likelihood estimator and other truncated estimators are being viewed as preliminary test

estimators. A new preliminary test estimator is proposed and studied and also recommendations of the levels of preliminary test are made.

Grimes and Sukhatme (1980) considered that if the data on the auxiliary variable X correlated with the variable Y under study are available, regression-type estimators are often used to estimate the population mean μ_y. An estimator based on preliminary test of significance that chooses between the difference estimator and the regression estimator has been proposed. This article investigates the efficiency of the proposed regression type estimator with respect to other regression type estimators.

Sisodia (1981) suggested a preliminary test estimator for the population mean on the current occasion in case of sampling over two occasions is built up which depends on the outcome of the preliminary test. Both the cases are considered when variance-covariance of the variables on both the occasions are known and unknown. In both the cases the preliminary test estimators are found to be better than usual estimators for large value of correlation coefficient ρ depending upon the proper choice of level of significance α and the probability q.

Sisodia and Srivastava (1982) proposed a modified regression estimator with a preliminary test in double sampling, alternative to the usual regression estimator for a population mean in double sampling. On the basis of a preliminary test of a simple hypothesis about the auxiliary variate-mean. Two phase sampling is assumed from a bivariate normal population. Gain in efficiency is investigated theoretically and empirically.

Esimai and Han (1982) suggested a regression estimator in double sampling and partial information on the multivariate auxiliary variable. They evaluated the bias and for comparison with other estimator, the MSE was computed and hence the relative efficiency evaluated. The results show that the suggested estimator is more efficient than the estimator when partial information is not used.

Han and Bancroft (1983) studied the estimation of variance in the normal distribution under statistical inference based on conditional specification. They mentioned that, when there are two samples available for estimating the variance and it is not certain whether the two samples are from the same population, the experimenter usually uses the test to resolve this uncertainty. When the tests are not significant, the samples are pooled to obtain the pooled estimator otherwise the individual sample variance is used. The bias and mean square of such a preliminary test estimator are studied. It is shown that the preliminary test estimator has a smaller mean square error than the usual unbiased estimator when the level of significance for the preliminary test is appropriately chosen.

Saleh and Kibria (1993) mentioned that the problem of estimation of the regression coefficients in a multiple regression model is considered under multi collinearity situation when it is suspected that the regression coefficients may be restricted to a subspace. They presented the estimators of the regression coefficients combining the idea of preliminary test and ridge regression methodology. Accordingly, it is considered three estimators, namely, the

unrestricted ridge regression estimator (URRE), the restricted ridge regression estimator (RRRE), and finally, the preliminary test ridge regression estimator (PTRRE). The biases, variance covariance matrices and mean square error of the estimators are derived and compared with the usual estimators. Regions of optimality of the estimators are determined by studying the MSE criterion. The conditions of superiority of the estimators over the traditional estimators as in Saleh and Han (1990) and Ali and Saleh (1991) have also been discussed.

Das and Bez (1995) in their article, suggested some estimators for the population mean μ_y in double sampling with two auxiliary variables X and Z, alternative to the usual regression estimator. They considered that (X,Y,Z) follow joint trivariate normal distribution with mean (μ_x, μ_y, μ_z) and covariance matrix Σ in which the variances are denoted by σ_x^2, σ_y^2 and σ_z^2. They considered that partial information about one of the auxiliary variable, say μ_z is available. In order to utilize the partial information, they performed a preliminary test about the hypothesis $H_0 : \mu_z = \mu_0$ where μ_0 is the value obtained from the partial information. After the preliminary sample of size n' is obtained, $H_0 : \mu_z = \mu_0$ can be tested against $H_0 : \mu_z \neq \mu_0$. If H_0 is accepted, μ_0 will be used in the regression estimator; if H_0 is rejected, the sample mean based on the preliminary sample is used. Since μ_x is totally unknown, it is estimated from another preliminary sample of size n'. It is also assumed that Σ is known. The assumptions are incorporated into the regression model to obtain the suggested estimator. The bias, mean square error, relative efficiency and optimum allocation of sample sizes are obtained for the suggested estimator. It is shown

that under certain conditions preliminary test estimator in double sampling with two auxiliary variable having partial information on only one auxiliary variable is more efficient than regression estimators in double sampling with two auxiliary variables.

Next they proceeded to construct another preliminary test estimator in double sampling with two auxiliary variables having partial information on both the auxiliary variables X and Z. In this case also the bias, mean square error, relative efficiency and optimum allocation of sample sizes are obtained for the suggested estimator. It is shown that under certain conditions preliminary test estimator in double sampling with two auxiliary variable having partial information on both auxiliary variables is more efficient than preliminary test estimator in double sampling with two auxiliary variable having partial information on only one auxiliary variable.

Kibria (1996) in his article studied the properties of the preliminary test, restricted and unrestricted ridge regression estimators of the linear regression model with non-normal disturbances. He presented the estimators of the regression coefficients combining the idea of preliminary test and ridge regression methodology, when it is suspected that the regression coefficients may be restricted to a subspace and the regression error is distributed as multivariate t. Accordingly they considered three estimators, namely the unrestricted ridge regression estimator, the restricted ridge regression estimator and finally the preliminary test ridge regression estimator. The biases and the mean square error of the estimators are derived under the null and alternative

hypotheses and compared with the usual estimators. By studying the MSE criterion, the regions of optimality of the estimators are determined.

Das (2003) proposed a generalized study of preliminary test estimators in double sampling, in the same lines with that given in Das and Bez (1995).The only difference in this case is that the size of the preliminary sample that arise in the estimation of the means μ_x and μ_z are considered to be different.

In another work, Kibria and Saleh (2004) studied the preliminary test ridge regression estimators (PTRRE) based on the Wald (W), Likelihood Ratio (LR) and Lagrangian Multiplier (LM) tests for estimating the regression parameters. Here the authors considered the multiple regression model with student t error distribution. The bias and the mean square errors (MSE) of the proposed estimators are derived under both null and alternative hypothesis. By studying the MSE criterion, the regions of optimality of the estimators are determined. Under the null hypothesis, the PTRRE based on LM test has the smallest risk followed by the estimators based on LR and W tests. However, the PTRRE based on W test performs the best followed by the LR and LM based estimators when the parameter moves away from the subspace of the restrictions. The conditions of superiority of the proposed estimators for shrinkage parameter, k and the departure parameter, D are provided. Some tables for the maximum and minimum guaranteed efficiency of the proposed estimators have been given, which allows the authors to determine the optimum level of significance corresponding to the optimum estimator. Finally, they

conclude that the estimator based on Wald test dominates the other two estimators in the sense of having highest minimum guaranteed efficiency.

1.9 Objectives of the study

The main objective of the thesis is to suggest alternative preliminary test estimators of the population mean in double sampling. The performance of the bias function of the suggested estimator will be considered. In order to compare the efficiency of the suggested estimator(s) to that of the existing estimator(s), the Mean Square Error (MSE) will be used as a useful criteria. The determination of bias and MSE for the suggested estimator involves conditional expectations. These will be derived under certain assumptions about the joint distribution of the parent populations of the variables from where the samples are drawn.

It is well known that stratified sampling consists of classifying the population units in a certain number of groups called strata and selecting samples independently from each group. The division of population into strata can be done in such a way that the variability of the study variable is homogeneous within each stratum , in that the measurement vary little from one unit to another, a precise estimates of any stratum mean can be obtained from a small sample in that stratum. These estimates with best choices of sample sizes can be combined into a precise estimate for the whole population. When appropriately used, the variance of the estimated mean of the study variable Y

under stratification is usually less than the variance under simple random sampling.

It is also well known fact that for estimating the population mean μ_y of a random variable Y, precision of the estimator can be increased when information on an auxiliary variable X, highly correlated with Y is readily available on all the units of the population. In the present investigation an attempt will be made to construct preliminary test estimators in double sampling under stratified sampling scheme. Attempt will be made to determine whether or not the suggested estimator leads to higher efficiency when compared with other known estimator under the similar assumptions.

Cochran (1977) suggested a regression estimate in stratified sampling which he called a combined regression estimate and is given by

$$\bar{y}_{lrc} = \bar{y}_{st} + b(\mu_x - \bar{x}_{st}), \text{ where}$$

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h \quad \text{and} \quad \bar{x}_{st} = \sum_h W_h \bar{x}_h$$

In this estimate the whole population is stratified into different classes and samples are selected from each stratum by simple random sampling and the stratum means are combined and used in a regression equation to obtain the desired mean. Here b is the estimate of combined regression coefficient and W_h is the stratum weight.

The combined linear regression estimator given by $\bar{y}_{lrc}$, can be utilized under three situations. Firstly when the population mean μ_x is known, as a consequence of which the study reduces to usual combined regression method

of estimation. Secondly in certain practical situations μ_x is not known a prior, in which case the technique of double sampling can be applied wherein a preliminary sample is obtained to estimate μ_x and the estimator of μ_y is given by

$$t_4 = \bar{y}_{st} + b(\bar{x}_{n'} - \bar{x}_{st}) \text{ , where}$$

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h \quad \text{and} \quad \bar{x}_{st} = \sum_h W_h \bar{x}_h$$

Here $\bar{x}_{n'}$ is the value of the mean of X obtained from the preliminary sample and is utilized to estimate μ_x. Thirdly when μ_x is partially known, then a preliminary test estimation using double sampling procedure can be used.

In the present study, the third case will be considered where partial information about the mean of the auxiliary variable will be used. The first sample is a stratified simple random sample of size n in which the pair (x_{hi}, y_{hi}) values are measured from n_h units drawn from each stratum and consequently estimating the pair $(\bar{x}_{st}, \bar{y}_{st})$, with $n = \sum_h n_h$ .The second sample is a larger simple random sample of size $n'(=n+m)$ is obtained by supplementing m more independent observations on X where only x_i is measured and evaluates $\bar{x}_{n'}$ which is utilized to estimate μ_x.

In order to utilize the partial information, a preliminary test is done about the hypothesis

$$H_o : \mu_x = \mu_o \text{ , against } \quad H_1 : \mu_x \neq \mu_0$$

where μ_0 is the value obtained from the partial information.

If H_0 is accepted then μ_0 is used to replace μ_x in the regression estimator $\bar{y}_{lrc}$ and if H_0 is rejected then the sample mean $\bar{x}_{n'}$ based on the preliminary sample is used.

The bias function of the suggested estimator will then be derived. For comparision of a biased estimator with an unbiased estimator, or two estimators with different amounts of bias, mean square error of the estimator shall be evaluated. Further, the behavior of the efficiency of the suggested estimator can be evaluated through the relative efficiency. In order to determine the optimum MSE of the suggested estimator, cost function approach will be followed. This can be done by suggesting a suitable linear cost function C and then conditions will be determined to minimize the MSE for specified C.

Many authors have constructed preliminary test estimators in double sampling and derivation of the bias, MSE of these estimators involves complex mathematical techniques like conditional expectations, integrations and complicated substitutions of functions. Significant amount of energy were to be expended on the technique to evaluate the bias, mean square error etc. An alternative method for the evaluation of the bias, MSE is also sought with the help of numerical techniques. Here the complicated integrations that arise can be solved by numerical integration that can be obtained by using quadrature formulae and the most widely used amongst these is the Simpson's 1/3 rule.

Further, due to availability of high sped computation facilities, it is much faster and easier to use the numerical techniques. Using numerical techniques and computers to obtain the bias, MSE etc, one can approach these

calculations without recourse to simplifying assumptions or time-intensive techniques. Although analytical solutions are extremely valuable both for problem solving and for providing insight, numerical methods represent alternatives that greatly enlarge one's own capabilities to confront and solve problems. As a result more time is available for use of creative skills for other meaningful purposes. Thus, more emphasis can be placed on problem formulation and interpretation of the solution. In addition to analytical solution, the present study will also attempt to find the bias and MSE by the use of numerical techniques with programs written in Fortran 77.

Finally, the thesis also aims at investigating empirically all the results that are being derived by analytical methods. In the present study the empirical investigation will be done by using both real life data and also data generated through simulation. Thus an attempt shall be made to show the applications of the suggested estimators under practical situations and also to demonstrate the performance of these vis-à-vis other existing estimators.

1.10 Plan of chapters

The second chapter will attempt to construct the proposed estimator by the use of preliminary test in double sampling. As it is well known that an auxiliary variable increases the precision of estimating the mean of the study characteristic, a linear regression model can be considered for the present study. Again in many of the modern and complex surveys, the information on characteristics is conveniently extracted when the population units are classified

in a certain number of groups called strata and selecting samples independently from each group. The division of population into strata can be done in such a way that the variability of the study variable is homogeneous within each stratum. Cochran (1977) mentioned that an appropriate combined regression estimator $\bar{y}_{lrc}$ for the population as a whole can be obtained by suitably combining the stratum-wise estimators of the characteristics under consideration. In the estimator $\bar{y}_{lrc}$, it can be assumed that mean of the auxiliary variable is completely unknown for which case double sampling procedure can be used in the estimation of the mean of the auxiliary variable. However in certain practical situation, the experimenter can have partial information about the mean of the auxiliary variable. In order to utilize the partial information, a preliminary test can be performed. Then a combined regression estimator based on this preliminary test can be defined. Assumptions can also be made about the joint distribution of the parent population of the characteristics under study and from where the samples are drawn and this joint distribution can be considered to be a multivariate normal. The bias function of the proposed estimator will then be derived. The derivation can involve conditional expectations, the conditions being the acceptance or rejection of the hypothesis considered in the preliminary test. Behavior of the bias function shall be studied under different levels of the preliminary test. The evaluation of the bias can also be attempted by numerical methods and the results are then compared with that obtained by analytical method.

In the third chapter, most of the work shall be devoted on the derivation of an useful criterion viz. Mean Square Error. The mean square error of the combined regression preliminary test estimator will be deduced in order to compare its efficiency with other existing estimator(s) in double sampling through stratification. The derivation of the MSE may involve conditional expectations, the conditions being the acceptance or rejection of the hypothesis considered in the preliminary test. To determine this function, the joint distribution of the parent population of the characteristics under study will be considered to follow a multivariate normal distribution. The evaluation of the MSE will also be attempted by numerical methods and the results are then compared with that obtained by analytical method.

In the fourth chapter, relative efficiency of the proposed estimator in comparison with other estimators shall be considered. This chapter will also discuss about the problem of optimum allocation of the size of the sample in planning of a sample survey. The optimum allocation of sample sizes are attained either by minimizing precision against a given cost or minimizing cost against given precision and consequently the corresponding optimum mean square of the proposed estimator can be compared with other compatible estimators.

In the fifth chapter, an attempt will be made to investigate empirically all the theoretical results that are derived in the previous chapters. The empirical works shall be carried out with the help of both real life data as well as data simulated using some standard statistical software package viz. STATA 8.0.The

validity of the assumptions about the distributions are tested using SYSTAT 12.0.

Chapter 2

Combined regression preliminary test estimator in double sampling with partial information on the auxiliary variable.

2.1 Introduction

It is a well known fact that for estimating the population mean μ_y of a random variable Y, precision of the estimator can be increased when information on an auxiliary variable X, highly correlated with Y is readily available on all the units of the population. When the relationship between Y and X is found to be approximately linear but does not pass through the origin, linear regression estimate may be used, and the estimator is given by (Cochran, 1977)

$$t_1 = \bar{y} + b(\mu_x - \bar{x}) \qquad \text{.....(2.1)}$$

where b is an estimate of the change in y when x is increased by unity, $\bar{y}$ and $\bar{x}$ are the sample means of the variables Y and X respectively and μ_x is the population mean of X .

The regression estimator given in (2.1) requires advance knowledge about μ_x, the population mean of the auxiliary variable X. When such information is lacking, double sampling technique can be utilized, wherein it is sometimes considered relatively cheaper to take a large preliminary sample in which x_i alone is measured and is used for estimating the population characteristic like mean and total. The purpose of this sample is to furnish a good estimate of μ_x. Another independent or sub-sample observes both (x_i, y_i) meant to estimate $(\bar{x}, \bar{y})$ for using it in the regression estimator.

To use the linear regression estimator t_1 it is usually assumed that population mean μ_x is known. However in certain practical situation, μ_x is not

known a prior, in which case the technique of double sampling is applied. In the first preliminary sample of size n', we measure only x_i and use it for the estimation of μ_x ; in the second sample, a random sub sample of size n ($<n'$), from the preliminary sample, we observed both x_i and y_i. Under double sampling the regression estimate (2.1) becomes

$$t_2 = \bar{y}_n + b(\bar{x}_{n'} - \bar{x}_n) \qquad \ldots\ldots\ldots\ldots \quad (2.2)$$

where $\bar{x}_{n'}$ is the mean of x_i in the first sample and $(\bar{x}_n, \bar{y}_n)$ are the means of x_i and y_i in the second sample and b is the least square regression coefficient of Y on X which can be computed from the second sample.

Han (1973) described that the precision of an estimator can be improved if auxiliary variable is used in a regression estimator based on double sampling with partial information on auxiliary variable. Sometimes there are situations where we have partial information about the mean μ_x of the auxiliary variable X. In order to utilize the partial information, Han(1973) suggested the use of a preliminary test and constructed a preliminary test estimator using double sampling with partial information on the auxiliary variable as follows;

$$t_3 = \begin{cases} (\bar{y}_n - \rho\bar{x}_n) & if \quad \left|\bar{x}_{n'}\right| \le Z_\alpha / \sqrt{n'} \\ (\bar{y}_n + \rho(\bar{x}_{n'} - \bar{x}_n)) & if \quad \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'} \end{cases} \qquad \ldots\ldots\ldots\ldots (2.3)$$

where Z_α is the 100(1-α/2)% point of N (0,1) and α is the level of significance of the preliminary test .

In estimating the population mean μ_y of the random variable Y, suppose that in addition to information on an auxiliary variable X, information on yet another auxiliary variable Z is available. When μ_x and μ_z are not available, one can take a preliminary sample to estimate these by the use of double sampling. In such a situation an estimator using X and Z is being suggested by Mukerjee *et.al* (1987).

Das(1992), Das and Bez(1995) and Das(2003), suggested some preliminary test estimators for the population mean in double sampling with two auxiliary variables, alternative to the usual regression estimator, when the experimenter has partial information on one and /or both auxiliary variables.

The present work is aimed to proceed in accordance to further enhance the work done by Han (1973) and Das (1995) and several other authors to find an appropriate estimator through the use of preliminary test estimation and double sampling procedures.

2.2 The combined regression preliminary test estimator (CRPTE) in double sampling

It is known that in many of the large scale surveys, it is inevitable to adopt stratification for the purpose of preparing a frame from which the sample can be extracted. Stratification produces gain in precision in the estimate of the characteristics of the whole population. It consists of classifying the population units in a certain number of groups called strata and selecting samples independently from each group. The division of population into strata can be done in such a way that the values of the study variable are homogeneous

within each stratum. Stratification can also be operationally convenient and economical if the sampling frame is available in the form of sub-frames.

An appropriate estimator for the population as a whole can be obtained by suitably combining the stratum-wise estimators of the characteristics under consideration. Stratification enables that the demarcation of the strata boundaries and the allocation of the total sample size to the strata may be done so as to make the estimator most efficient from the point of view of sampling variability and cost. Though the main advantage of using stratified sampling is the possible increase in efficiency per unit of cost in estimating the population characteristics, the method is also useful in situation when estimators are required with specific margins of errors not only for population as a whole but for certain groups of units. When appropriately used, the variance of the estimated mean of the study variable Y under stratification is usually less than that of the variance under simple random sampling (Cochran,1977).

The present study attempts to proceed in accordance to further enhance the work done by Han (1973) and Das (1995), by utilizing double sampling and use preliminary test on the partial information on auxiliary variable in stratified sampling. Cochran(1977) mentioned that as with ratio estimates, there are two types of regression estimates that can be constructed in stratified random sampling. In the first estimate, $\bar{y}_{lrs}$, a separate regression estimate is computed for each stratum mean, that is,

$$\bar{y}_{lrh} = \bar{y}_h + b_h(\mu_x - \bar{x}_h), \quad \text{for every h}$$

and with $W_h = (N_h / N)$ $\quad \bar{y}_{lrs} = \sum_h W_h \bar{y}_{lrh}$

where b_h is the within-stratum least square estimate of B_h and W_h is the stratum weight, $(\bar{y}_h, \bar{x}_h)$ are the stratum means of Y and X respectively. Further, he discussed that this estimate is appropriate when it is thought that the true regression coefficients B_h vary from stratum to stratum.

In the second regression estimate, $\bar{y}_{lrc}$, the combined regression estimate is computed as

$$\bar{y}_{lrc} = \bar{y}_{st} + b(\mu_x - \bar{x}_{st}) \quad \text{, where}$$

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h \quad \text{and} \quad \bar{x}_{st} = \sum_h W_h \bar{x}_h$$

where b is the estimate of combined regression coefficient and W_h is the stratum weight. Cochran states that $\bar{y}_{lrc}$ is appropriate when b_h an estimate of B_h are presumed to be the same in all strata.

In many of the studies done in stratification, it is known that though the within stratum regression coefficients B_h might differ slightly from stratum to stratum, but as such the relationship between the pair (X,Y) is always maintained. Considering the combined regression estimate, the whole population is stratified into different classes and samples are selected from each stratum by simple random sampling. The mean from each stratum are calculated and utilizing the stratum weight which is estimated by proportional allocation, the strata means are combined to obtain the desired combined regression estimate.

The combined linear regression estimator $\bar{y}_{lrc}$ can be utilized under three situations. Firstly when the population mean μ_x is known as a consequence of which, the study reduces to usual combined regression method of estimation. Secondly in certain practical situations μ_x is not known a prior, in which case the technique of double sampling can be applied wherein a preliminary sample is obtained to estimate μ_x and the estimator of μ_y is given by

$$t_4 = \bar{y}_{st} + b(\bar{x}_{n'} - \bar{x}_{st}), \qquad \text{.....(2.4)}$$

where

$$\bar{y}_{st} = \sum_h W_h \bar{y}_h \quad \text{and} \quad \bar{x}_{st} = \sum_h W_h \bar{x}_h$$

Here $\bar{x}_{n'}$ is the value of the mean of X obtained from the preliminary sample and is utilized to estimate μ_x. Thirdly, in certain situations, the experimenter may have partial information about μ_x. Under such circumstances a preliminary test estimator using double sampling procedure can be used.

In the present study, the third case will be considered where partial information about the mean of the auxiliary variable will be used. The first sample is a stratified simple random sample of size n in which the pair (x_{hi}, y_{hi}) values are measured from n_h units drawn from each stratum and consequently estimating the pair $(\bar{x}_{st}, \bar{y}_{st})$, with $n = \sum_h n_h$ .The second sample is a larger simple random sample of size $n'(=n+m)$ which is obtained by supplementing m

more independent units on X where only x_i is measured and evaluates $\bar{x}_{n'}$ which is utilized to estimate μ_x.

In order to utilize the partial information a preliminary test is done about the hypothesis

$$H_o : \mu_x = \mu_o \quad \text{against} \quad H_1 : \mu_x \neq \mu_0$$

where μ_0 is the value obtained from the partial information. If H_0 is accepted then μ_0 is used to replace μ_x in the regression estimator $\bar{y}_{lrc}$ and if H_0 is rejected then the sample mean $\bar{x}_{n'}$ based on the preliminary sample is used in $\bar{y}_{lrc}$.

We assume that the auxiliary variable X and the study variable Y are jointly normally distributed with parameters given by $(\mu_x, \mu_y, \sigma_x^2, \sigma_y{}^2, \rho)$. The marginal distributions which is the distribution of the study variable Y and the auxiliary variable X follow normal distribution given as Y ~ $N(\mu_y, \sigma_y^2)$ and X ~ $N(\mu_x, \sigma_x^2)$. The regression estimator depends on weather the covariance matrix is known or not. If known, one may let $\sigma_x{}^2 = \sigma_y{}^2 = 1$ without loss of generality (WLOG). The strata population (X_h, Y_h) being carved out from the parent population, are also jointly assumed to follow bivariate normal distribution with parameters written as $\left(\mu_{x_h}, \mu_{y_h}, \sigma_{x_h}{}^2, \sigma_{y_h}{}^2, \rho\right)$. The correlation coefficient within each stratum between the pair of variables (X_h, Y_h) might differ slightly from strata to strata, but such a relationship of the pair (X, Y) is maintained in (X_h, Y_h) for every stratum h. Hence the strata

correlations are assumed to be equal to the population correlation coefficient ρ .

Since the population is assumed to follow normal distribution, the preliminary sample utilized to collect information on the auxiliary variable for the estimation of $\bar{x}_{n'}$ is also assumed to follow normal distribution and therefore $\bar{x}_{n'} \sim N(\mu_x, \sigma_x^2 / n')$ and under the assumption $\sigma_x^2 = \sigma_y^2 = 1$, $\bar{x}_{n'} \sim N(\mu_x, 1/n')$. Further marginal distributions of X_h and Y_h are also normal given as $X_h \sim N(\mu_{x_h}, \sigma_{x_h}^2)$ and $Y_n \sim N(\mu_{y_h}, \sigma_{y_h}^2)$.

For each stratum, the pair of variables (X_h, Y_h) for every h, follows a bivariate normal distribution with mean (μ_{x_h}, μ_{y_h}) and covariance matrix given by

$$\sum_h = \begin{pmatrix} \sigma_{x_h}^2 & \rho\sigma_{x_h}\sigma_{y_h} \\ \rho\sigma_{x_h}\sigma_{y_h} & \sigma_{y_h}^2 \end{pmatrix}$$

The regression estimator depends on whether $\sum_h$ is known or not. If $\sum_h$ is known, one may let $\sigma_{x_h}^2 = \sigma_{y_h}^2 = 1$, (WLOG).

The stratum means are given by

$$\bar{x}_h = \sum_{i=1}^{n_h} x_{hi} / n_h \quad \text{and} \quad \bar{y}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$$

are linear combination of normally distributed random variables (X_h, Y_h). Hence it can be easily observed that $\bar{x}_h$ and $\bar{y}_h$ also follow normal distribution given by

$$\bar{x}_h \sim N(\mu_{x_h}, \sigma_{x_h}^2 / n_h) \qquad \text{and} \qquad \bar{y}_h \sim N(\mu_{y_h}, \sigma_{y_h}^2 / n_h)$$

i.e. $\bar{x}_h \sim N(\mu_{x_h}, 1/n_h)$ and $\bar{y}_h \sim N(\mu_{y_h}, 1/n_h)$

under the assumption of $\sigma_{x_h}^2 = \sigma_{y_h}^2 = 1$

Since it is assumed that the joint distribution of the pair (X,Y) is normal, so it follows that the joint distribution of $(\bar{x}_h, \bar{y}_h)$ is bivariate normal with mean as (μ_{x_h}, μ_{y_h}) and covariance matrix given by

$$\sum_c = \begin{pmatrix} \sigma_{x_h}^2/n_h & \rho\sigma_{x_h}\sigma_{y_h}/n_h \\ \rho\sigma_{x_h}\sigma_{y_h}/n_h & \sigma_{y_h}^2/n_h \end{pmatrix} = \begin{pmatrix} 1/n_h & \rho/n_h \\ \rho/n_h & 1/n_h \end{pmatrix}$$

In certain situation, the experimenter may have partial information about μ_x. In order to utilize the partial information, one can perform a preliminary test about the hypothesis

$H_o : \mu_x = \mu_o$, against $H_1 : \mu_x \neq \mu_0$

where μ_0 is the value obtained from the partial information and $\bar{x}_{n'}$ which is the value of the mean of X obtained from the preliminary sample through the use of double sampling is being used to test the hypothesis.

Now, when μ_x is partially known, one can let $\mu_o = 0$ (WLOG), so that the hypothesis can be accepted when,

$$\left|(\bar{x}_{n'} - \mu_o)/SE(\bar{x}_{n'})\right| \leq Z_\alpha$$

$$\Rightarrow \left|\bar{x}_{n'}/(1/\sqrt{n'})\right| \leq Z_\alpha$$

$$\Rightarrow \left|\bar{x}_{n'}\right| \leq Z_\alpha/\sqrt{n'}$$

where Z_α is the 100(1-α/2)% point of N (0,1) and α is the level of significance of the preliminary test.

Under the above assumption the Combined regression preliminary test estimator (CRPTE) in double sampling having partial information on the auxiliary variable X can be written as

$$t_5 = \begin{cases} (\bar{y}_{st} - \rho \bar{x}_{st}) & if \quad |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'} \\ (\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) & if \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'} \end{cases} \quad \text{.....(2.5)}$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

and the regression coefficient b from $\bar{y}_{lrc}$ reduces to $b=\rho(\sigma_y/\sigma_x)=\rho$ under the above assumptions.

2.3 Bias of the CRPTE

To evaluate the bias of t_5, we consider that the joint distribution of $(\bar{x}_{n'}, \bar{x}_{st}, \bar{y}_{st})$ is a multivariate normal with mean (μ_x, μ_x, μ_y) and covariance matrix given by

$$\sum = \begin{pmatrix} Var(\bar{x}_{n'}) & Cov(\bar{x}_{n'}, \bar{x}_{st}) & Cov(\bar{x}_{n'}, \bar{y}_{st}) \\ Cov(\bar{x}_{st}, \bar{x}_{n'}) & Var(\bar{x}_{st}) & Cov(\bar{x}_{st}, \bar{y}_{st}) \\ Cov(\bar{y}_{st}, \bar{x}_{n'}) & Cov(\bar{y}_{st}, \bar{x}_{st}) & Var(\bar{y}_{st}) \end{pmatrix} \quad \text{.....(2.6)}$$

Now

The variance of $\bar{x}_{st}$ and $\bar{y}_{st}$ being given by

$$Var(\bar{x}_{st}) = \sum W_h^2 \sigma_{x_h}^2 / n_h \qquad Var(\bar{y}_{st}) = \sum W_h^2 \sigma_{y_h}^2 / n_h \quad \text{(Cochran 1977)}$$

which under the assumption $\sigma_{x_h}^2 = \sigma_{y_h}^2 = 1$, becomes

$Var(\bar{x}_{st}) = \sum W_h^2 / n_h$ $\qquad$ $Var(\bar{y}_{st}) = \sum W_h^2 / n_h$

When the samples are selected with proportional allocation then the stratum weight is given by $W_h = (N_h / N) = (n_h / n)$

Thus $\qquad \sum_h W_h^2 / n_h = \sum_h W_h^2 / n W_h = (1/n) \sum_h W_h = (1/n) \qquad (\text{ as } \sum_h W_h = 1)$

Hence the covariance matrix (2.6) reduces to

$$\sum = \begin{pmatrix} 1/n' & 1/n' & \rho/n' \\ 1/n' & 1/n & \rho/n \\ \rho/n' & \rho/n & 1/n \end{pmatrix} \qquad \text{.....(2.7)}$$

The derivation of bias of t_5 involves conditional expectations, the condition being the acceptance or rejection of the hypothesis considered in the preliminary test. Further the expectations can be obtained from the integrals involving probability density functions which are assumed to be normal. The bias of the estimator t_5 is derived as follows ;

The Bias of an estimator is defined as

$$Bias\ (t_5) = E(t_5) - \mu_y \qquad \text{.....(2.8)}$$

where E(.) is the mathematical expectation

$$Bias\ (t_5) = E\left\{(\bar{y}_{st} - \rho \bar{x}_{st}) \quad \text{..... } if \quad |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'}\right\}$$
$$+ E\left\{(\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) \quad \text{..... } if \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right\} - \mu_y$$

$$= E(\bar{y}_{st} - \rho \bar{x}_{st}) + \left\{E(\rho \bar{x}_{n'}) \quad \text{..... } if \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right\} - \mu_y$$

$$\Rightarrow Bias\ (t_5) = \sum W_h \mu_{y_h} - \rho \sum W_h \mu_{x_h} + E\{\rho \bar{x}_{n'} / |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\} - \mu_y$$

It is given that

$\sum W_h \mu_{x_h} = \mu_x$ and $\sum W_h \mu_{y_h} = \mu_y$ (Cochran, 1977)

Thus,

$$Bias\ (t_5) = -\rho\mu_x + E\{\rho \bar{x}_{n'} \quad \ldots\ldots\ldots\ if \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$$

$$= -\rho\mu_x + E\{\rho \bar{x}_{n'} \ .. \ if \ \bar{x}_{n'} > Z_\alpha / \sqrt{n'}\} + E\{\rho \bar{x}_{n'} \quad if \ \bar{x}_{n'} < -Z_\alpha / \sqrt{n'}\}$$

$$\Rightarrow Bias\ (t_5) = -\rho\mu_x + \rho\{\int_{Z_\alpha/\sqrt{n'}}^{\infty} \bar{x}_{n'} f(\bar{x}_{n'}) d\bar{x}_{n'} + \int_{-\infty}^{-Z_\alpha/\sqrt{n'}} \bar{x}_{n'} f(\bar{x}_{n'}) d\bar{x}_{n'}\}$$

.....(2.9)

where $f(\bar{x}_{n'})$ is the probability density function of $\bar{x}_{n'}$ which follows $N(\mu_x, 1/n')$ under the assumption $\sigma_x^2 = 1$.

Let

$$I = \int_{Z_\alpha/\sqrt{n'}}^{\infty} \bar{x}_{n'} f(\bar{x}_{n'}) d\bar{x}_{n'} + \int_{-\infty}^{-Z_\alpha/\sqrt{n'}} \bar{x}_{n'} f(\bar{x}_{n'}) d\bar{x}_{n'}$$

$$= (\sqrt{n'}/\sqrt{2\pi}) \left[\int_{z_\alpha/\sqrt{n'}}^{\infty} \bar{x}_{n'} Exp\{(-1/2)((\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}))^2\} d\bar{x}_{n'} + \int_{-\infty}^{-Z\alpha/\sqrt{n'}} \bar{x}_{n'} Exp\{(-1/2)((\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}))^2\} d\bar{x}_{n'} \right.$$

Putting $w = (\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}) \Rightarrow dw = \sqrt{n'}d\bar{x}_{n'}$, we have

When $\bar{x}_{n'} = Z_\alpha/\sqrt{n'}$ then $w = \sqrt{n'}(Z_\alpha/\sqrt{n'} - \mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$ and

When $\bar{x}_{n'} = -Z_\alpha/\sqrt{n'}$ then $w = \sqrt{n'}(-Z_\alpha/\sqrt{n'} - \mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

Therefore,

$$I = (1/\sqrt{2\pi})\{\int_A^\infty (\mu_x + (w/\sqrt{n'}))Exp((-1/2)w^2)dw + \int_{-\infty}^B (\mu_x + (w/\sqrt{n'}))Exp((-1/2)w^2)dw$$

$$= (1/\sqrt{2\pi})\mu_x\left\{\int_A^\infty Exp((-1/2)w^2dw + \int_{-\infty}^B Exp((-1/2)w^2dw\right\}$$

$$+ (1/\sqrt{2\pi})(1/\sqrt{n'})\left\{\int_A^\infty wExp((-1/2)w^2dw + \int_{-\infty}^B wExp((-1/2)w^2dw\right\}$$

$$= \mu_x\{1 - \Phi(A) + \Phi(B)\} + (1/\sqrt{2\pi n'})\left\{\int_A^\infty wExp((-1/2)w^2dw + \int_{-\infty}^B wExp((-1/2)w^2dw\right\}$$

where $\Phi(.)$ is the cumulative distribution function of N(0,1).

Again

putting $(w^2/2) = t \Rightarrow wdw = dt$

$$I = \mu_x \{1 - \Phi(A) + \Phi(B)\} + (1/\sqrt{2\pi n'}) \left\{ \int_{A^2/2}^{\infty} Exp\ (-t)\, dt + \int_{-\infty}^{B^2/2} Exp\ (-t)\, dt \right\}$$

$$= \mu_x \{1 - \Phi(A) + \Phi(B)\} + (1/\sqrt{2\pi n'}) \{Exp\ (-A^2/2) - Exp\ (-B^2/2)\}$$

$$= \mu_x \{1 - \Phi(A) + \Phi(B)\} + (1/\sqrt{n'}) \{(1/\sqrt{2\pi})\, Exp\ (-A^2/2) - (1/\sqrt{2\pi})\, Exp\ (-B^2/2)\}$$

$$= \mu_x \{1 - \Phi(A) + \Phi(B)\} + (1/\sqrt{n'}) \{\varphi(A) - \varphi(B)\}$$

where $\varphi(.)$ is the density function of N(0,1).

Therefore from (2.9),

$$Bias\ (t_5) = -\rho\mu_x + \rho \left[\mu_x \{1 - \Phi(A) + \Phi(B)\} + (1/\sqrt{n'}) \{\varphi(A) - \varphi(B)\}\right]$$

$$= -\rho\mu_x \{\Phi(A) - \Phi(B)\} + \rho(1/\sqrt{n'}) \{\varphi(A) - \varphi(B)\} \quad \text{.....(2.10)}$$

2.4 Discussion

From Equation (2.10), we know that

$$Bias\ (t_5) = -\rho\mu_x \{\Phi(A) - \Phi(B)\} + \rho(1/\sqrt{n'}) \{\varphi(A) - \varphi(B)\}$$

As partial checks we have, the following :

When α = 0, i.e when we always accept H_0 , then $Z_\alpha = \infty$

Thus $Bias\,(t_5) = -\rho\mu_x$

When α = 1, then $Z_\alpha = 0$ Thus $Bias(t_5) = 0$

The values of $Bias(t_5)$ can be easily computed for different values of μ_x. In order to get an idea about the behavior of the bias function with respect to μ_x, $Bias(t_5)$ is computed for a set of values of n', α and ρ which are represented in Table 2.1 – 2.4 and Figure 2.1 - 2.4. When the bias of the proposed estimator is computed for different values of the mean of the auxiliary variables μ_x, it can be observed (Table 2.1 and Fig 2.1) that the behavior of the bias is symmetrical about $\mu_x = 0$. Thus it suffices to analyze the behavior of the bias for $\mu_x \geq 0$. It is found in general that $Bias(t_5)$ has minimum value zero at $\mu_x = 0$. As μ_x increases, the $Bias(t_5)$ increases to a maximum and then gradually decreases to zero. The Figures (2.1 - 2.4) clearly show that when the mean of the auxiliary variable is close to the hypothetical value, then bias is very close to 0. Also as μ_x moves away from the hypothetical value the bias increases, but after attaining maximum again gradually reduces to zero. This establishes the utility of the present study that the use of partial information and preliminary test reduces the bias of the proposed estimator.

Further, when the parameters α and ρ are fixed (Table 2.2 and Fig 2.2), then the bias is inversely proportional to the square root of the size of the preliminary sample n'. Therefore it can be concluded that with the increase in the preliminary sample size, the bias decreases. However, the bias is not affected by n, the size of the stratified random sample.

2.5 Bias of the CRPTE computed numerically

The above analytical method used for computing the bias of the proposed estimator involves the evaluation of mathematical expectation of the random variables and consequently results in the computation of integrals within certain limits. This may sometimes become very cumbersome, hence an alternative method for the evaluation of the bias is sought with the help of numerical techniques.

During the pre-computer era, significant amount of energy were expended on the technique to find solution, rather than on definition and interpretation of the problem. Mathematical solutions are usually derived for some problems using analytical or exact methods. These solutions were often useful and provide insight into the behavior of some systems. However, analytical solutions can be derived only for a limited class of problems. These problems include those that can be approximated with linear models and those with simple geometry and low dimensionality. Consequently, analytical solutions are often of limited practical value because most real problems are non-linear and involve complex shapes and processes. Numerical methods are extremely powerful problem solving tools. They are capable of handling large systems of equations and complicated geometries that are common in many mathematical and physical phenomena and that are often impossible to solve analytically. Through the use of numerical methods one can successively approximate both simple and complex solutions to evaluate the roots of equations, solving system of equations, evaluate differentiation and integration numerically, finding solutions to ordinary and partial differential equations and many other

mathematical approximations with great precision and accuracy (Chapra and Canale,1989).

Today's high speed computers provide an alternative for such complicated calculations. Using numerical techniques and computers to obtained solutions directly, one can approach these calculations without recourse to simplifying assumptions or time-intensive techniques. Although analytical solutions are extremely valuable both for problem solving and for providing insight, numerical methods represent alternatives that greatly enlarge one's own capabilities to confront and solve problems. As a result more time is available to the use of creative skills. Thus, more emphasis can be placed on problem formulation and interpretation of the solution.

The function to be differentiated or integrated usually will be typically a continuous function such as a polynomial, an exponential, or a trigonometric function or some other complicated function that is difficult to differentiate or integrate analytically. Sometimes one can come across a tabulated function where values of x and f(x) are given at a number of discrete points as is often the case with experimental or field data. In such instances, analytical solutions are difficult to obtain and therefore numerical analysis can be employed.

As we may see in chapter 2, the evaluation of bias of CRPTE involves computations of definite integrals. For this we may use numerical techniques. The most common approach for numerical integrations is the Newton-Cotes formulae which are based on replacing a complicated function with a simple polynomial that is easy to integrate. Three of the most widely used

Newton- cotes formulae are the trapezoidal rule, Simpson's 1/3 rule, and Simpson's 3/8 rule. The error in approximating an integral by Simpson's 1/3 rule is

$$\left|\{(b-a)^5/2880\}f^4(\xi)\right|$$

where $f(.)$ is the function to be integrated, a and b are the limits of integration and ξ is some number between *a* and *b*.

The error is (asymptotically) proportional to $(b-a)^5$. Simpson's rule gains an extra order because the points at which the integrand is evaluated are distributed symmetrically in the interval [*a*, *b*]. It may be noted that Simpson's rule provides exact results for any polynomial of degree three or less, since the error term involves the fourth derivative of *f*. The approximated formulae for numerical integration by Simpson's 1/3 rule is given as follows.

$$\int_{x_0}^{x_k} f(x)dx = (h/3)\left[f(x_0)+4\left\{\sum_{i=1,3,5,..}^{k-1} f(x_i)\right\}+2\left\{\sum_{i=2,4,6,..}^{k-2} f(x_i)\right\}+f(x_k)\right]$$

where $(x_i - x_{i-1}) = h$ (Jain, Iyengar and Jain, 2007)

The evaluation of integrals by numerical techniques involves the computation of the numerical values of the function $f(x)$ at different points $a = x_0, x_1, x_2, \ldots\ldots, x_k = b$. These functional values are then substituted in Simpson's rule to get an approximate value of the integral of $f(x)$. When the number of divisions or partitions of the range (b-a) increases, say when k >35

then the behavior of the bias function (Fig 2.5) becomes smooth and converges. Thus for all numerical integration in the present work, the number of divisions k is fixed at 50.

As mentioned above, the increase in the number of divisions is a necessity, and as a result computation of the integral by Simpson's rule becomes tedious and manual exercise is practically impossible. However the availability of high sped computation facilities makes it possible to evaluate the integral in a much faster and easier way. Thus by using numerical techniques and computers to obtain the bias one can approach these calculations without recourse to simplifying assumptions or time-intensive techniques.

In the present study alternative to analytical methods, attempt is also made to evaluate the bias of the suggested estimator t_5 numerically as follows;

The Bias of the proposed estimator is defined as

$$Bias(t_5) = E(t_5) - \mu_y$$

where E(.) is the mathematical expectation

$$Bias(t_5) = E\left\{(\bar{y}_{st} - \rho\bar{x}_{st}).....if \ \left|\bar{x}_{n'}\right| \le Z_\alpha / \sqrt{n'}\right\}$$
$$+ E\left\{(\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})).....if \ \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'}\right\} - \mu_y$$

$$= E(\bar{y}_{st} - \rho\bar{x}_{st}) + \left\{E(\rho\bar{x}_{n'}).....if \ \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'}\right\} - \mu_y$$

It is given that

$\sum W_h \mu_{x_h} = \mu_x$ and $\sum W_h \mu_{y_h} = \mu_y$ (Cochran, 1977)

Thus

$$Bias\ (t_5) = \mu_y - \rho\mu_x + E\{\rho\bar{x}_{n'} / \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'}\} - \mu_y$$

$$= -\rho\mu_x + E\{\rho\bar{x}_{n'} / \bar{x}_{n'} > Z_\alpha / \sqrt{n'}\} + E\{\rho\bar{x}_{n'} / \bar{x}_{n'} < -Z_\alpha / \sqrt{n'}\} \quad \text{.....(2.11)}$$

Let,

$$I = E\{\bar{x}_{n'} / \bar{x}_{n'} > Z_\alpha / \sqrt{n'}\} + E\{\bar{x}_{n'} / \bar{x}_{n'} < -Z_\alpha / \sqrt{n'}\}$$

$$= \int_{Z_\alpha / \sqrt{n'}}^{\infty} \bar{x}_{n'} f(\bar{x}_{n'}) d\bar{x}_{n'} + \int_{-\infty}^{-Z_\alpha / \sqrt{n'}} \bar{x}_{n'} f(\bar{x}_{n'}) d\bar{x}_{n'}$$

where $f(\bar{x}_{n'})$ is the probability density function of $\bar{x}_{n'} \sim N(\mu_x, 1/n')$ under the assumption $\sigma_x^2 = 1$

$$I = (\sqrt{n'} / \sqrt{2\pi}) \int_{z_\alpha / \sqrt{n'}}^{\infty} \bar{x}_{n'} Exp\left\{(-1/2)((\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}))^2\right\} d\bar{x}_{n'}$$

$$+ (\sqrt{n'} / \sqrt{2\pi}) \int_{-\infty}^{-Z\alpha / \sqrt{n'}} \bar{x}_{n'} Exp\left\{(-1/2)((\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}))^2\right\} d\bar{x}_{n'}$$

Putting $w = (\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}) \Rightarrow dw = \sqrt{n'} d\bar{x}_{n'}$ we have

When $\bar{x}_{n'} = Z_\alpha/(1/\sqrt{n'})$ then $w = \sqrt{n'}(Z_\alpha / \sqrt{n'} - \mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$ and

When $\bar{x}_{n'} = -Z_\alpha/(1/\sqrt{n'})$ then $w = \sqrt{n'}(-Z_\alpha / \sqrt{n'} - \mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

$$I = (1/\sqrt{2\pi}) \int_{A}^{\infty} (\mu_x + (w/\sqrt{n'})) Exp((-1/2)w^2) dw$$

$$+ (1/\sqrt{2\pi}) \int_{-\infty}^{B} (\mu_x + (w/\sqrt{n'})) Exp((-1/2)w^2) dw$$

Substituting the values of I in (2.11), it follows that

$$Bias\ (t_5) = -\rho\mu_x + \rho I$$

$$Bias\ (t_5) = \rho(I - \mu_x) \qquad \text{.....(2.12)}$$

where

$$I = (1/\sqrt{2\pi})\left\{\int_A^{\infty}\{\mu_x + (w/\sqrt{n'})\}\,Exp\,(-0.5w^2)dw\right\}$$

$$+ (1/\sqrt{2\pi})\left\{\int_{-\infty}^{B}\{\mu_x + (w/\sqrt{n'})\}\,Exp\,(-0.5w^2)dw\right\}$$

$$I = I_1 + I_2 \qquad \text{.....(2.13)}$$

The programs written on Fortran 77 (Rajaraman, 1997) were used in the numerical evaluation of the above integrals I_1 and I_2. (Appendix 1 and 2)

The integrals I_1 and I_2 involve the integrand function $g(w) = \{\mu_x + (w/\sqrt{n'})\}Exp(-0.5w^2)$ and in both, the limits of integration has infinity at one end. Such integration is tedious when evaluated by numerical techniques. The function $g(w)$ is plotted graphically for various values of w (Fig 2.6). The graph also show that the function $g(w)$ tapers to zero for $w > 3$ and $w < -3$. As we know that integration is the process of finding the area under the curve, bounded by the rectangular axis and the two ordinates corresponding to the limits of the integral, so the area under the curve $g(w)$ for the entire integrating limits defined in both I_1 and I_2 is approximately equal to that when the limits of

integration is confined to between A to 3 for I_1 and between -3 to B for I_2 . The values of I_1 and I_2 are given in Table 2.5.

2.6 Discussion

The values of $Bias(t_5)$ with respect to different values of μ_x are computed by the use of numerical techniques by substituting the output values of I, depicted in table 2.5 (a), (b), (c) in equation (2.12). In order to get an idea about the behavior of the bias function with respect to μ_x, $Bias(t_5)$ is computed for a set of values of α and ρ which are depicted in Table 2.6 – 2.7 and Figure 2.7 - 2.8. It is found that $Bias(t_5)$ is zero or minimum at $\mu_x = 0$. As μ_x increases, the $Bias(t_5)$ increases to a maximum and then gradually decreases to zero. The figure shows that when the mean of the auxiliary variable obtained by partial information is close to the hypothetical value, then the bias is very close to 0.

Han(1973) and Das and Bez(1995) in their paper constructed their estimators using analytical techniques exclusively. With the advent of modern and high speed digital computers, handling of complex statistical calculations can be done with ease in a short period of time. Also the advancement in field of Numerical techniques has simplified the method of solving complex mathematical analysis like the determination of roots of equations, solving systems of linear algebraic equations, differentiations and integrations, ordinary differential equations and partial differential equations.

In the present work, an attempt is being made to construct a preliminary test estimator in double sampling through stratification of the population. The combined linear regression estimator as suggested by Cochran (1997) is being used to construct a preliminary test estimator in double sampling for the present study. Bias is calculated analytically and the results are also plotted graphically. An attempt is also made in this chapter to evaluate the bias of the above estimators using numerical techniques and compared with that obtained analytically. Fig 2.9 show that the bias obtained by numerical methods depict a pattern similar to that obtained by analytical methods for increasing values of μ_x. The differences in the values of bias between analytical and numerical methods of computation are minimal.

Table 2.1 Behaviour of Bias(t_5) computed analytically with respect to μ_x for n' = 200, ρ = 0.8, α = 0.01.

μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
Bias	0.0	0.059	0.042	0.006	0.0	0.0	0.0	0.0	0.0	0.0	0.0
μ_x	**0**	**-0.1**	**-0.2**	**-0.3**	**-0.4**	**-0.5**	**-0.6**	**-0.7**	**-0.8**	**-0.9**	**-1.0**
Bias	0.0	0.059	0.042	0.006	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Table 2.2 Behaviour of Bias(t_5) computed analytically with respect to μ_x for different values of n' and for ρ = 0.8, α = 0.01

n' \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
120	0	0.069	0.077	0.039	0.006	0.0003	0	0	0	0	0
200	0	0.055	0.043	0.006	0.0001	0	0	0	0	0	0
400	0	0.044	0.007	0	0	0	0	0	0	0	0

Table 2.3 Behaviour of Bias(t_5) computed analytically with respect to μ_x for different values of α and for n' = 200 , ρ = 0.8

α \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.01	0	0.059	0.042	0.006	0	0	0	0	0	0	0
0.05	0	0.037	0.015	0.001	0	0	0	0	0	0	0
0.25	0	0.011	0.002	0	0	0	0	0	0	0	0

Table 2.4 Behaviour of Bias(t_5) computed analytically with respect to μ_x for different values of ρ and for n' = 200 , α = 0.01

ρ \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.7	0	0.033	0.013	0.001	0	0	0	0	0	0	0
0.8	0	0.037	0.015	0.001	0	0	0	0	0	0	0
0.9	0	0.042	0.017	0.001	0	0	0	0	0	0	0

Table 2.5 Numerically computed values of I with n' = 200 and ρ = 0.8 for (a) α = 0.01 (b) α = 0.05 (c) α = 0.25

(a)

μ_x	I_1	I_2	I
0	0.0010	-0.0010	0
0.1	0.02648	-6.44E-06	0.018939
0.2	0.14694	-6.21E-09	0.10346
0.3	0.29256	0	0.25633
0.4	0.3997	0	0.39278
0.5	0.499505	0	0.499505
0.6	0.59987	0	0.59987
0.7	0.69985	0	0.69985
0.8	0.79983	0	0.79983
0.9	0.89981	0	0.89981
1	0.99979	0	0.99979

(b)

μ_x	I_1	I_2	I
0	0.0041	-0.0041	0
0.1	0.0536	-5.81E-05	0.04417
0.2	0.1808	-1.28E-07	0.153693
0.3	0.2987	0	0.28734
0.4	0.39876	0	0.39876
0.5	0.49986	0	0.49986
0.6	0.59987	0	0.59987
0.7	0.69985	0	0.69985
0.8	0.79983	0	0.79983
0.9	0.89981	0	0.89981
1	0.99979	0	0.99979

(c)

μ_x	I_1	I_2	I
0	0.0144	-0.0144	0
0.1	0.0873	-0.0005	0.0821
0.2	0.19745	-3.26E-06	0.191048
0.3	0.299	0	0.2987
0.4	0.39986	0	0.39986
0.5	0.49989	0	0.49989
0.6	0.59987	0	0.59987
0.7	0.69985	0	0.69985
0.8	0.79983	0	0.79983
0.9	0.89981	0	0.89981
1	0.99979	0	0.99979

Table 2.6 Behaviour of Bias(t_5) computed numerically with respect to μ_x for different values of α and for $n' = 200$, $\rho = 0.8$

α \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.01	0	0.059	0.042	0.01	0	0	0	0	0	0	0
0.05	0	0.037	0.015	0	0	0	0	0	0	0	0
0.25	0	0.011	0.002	0	0	0	0	0	0	0	0

Table 2.7 Behaviour of Bias(t_5) computed numerically with respect to μ_x for different values of ρ and for $n' = 200$, $\alpha = 0.01$

ρ \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.7	0	0.033	0.013	0	0	0	0	0	0	0	0
0.8	0	0.037	0.015	0	0	0	0	0	0	0	0
0.9	0	0.05	0.042	0.01	0	0	0	0	0	0	0

Figue 2.1 Behaviour of Bias(t_5) computed analytically with respect to μ_x for n' = 200, ρ = 0.8, α = 0.01.

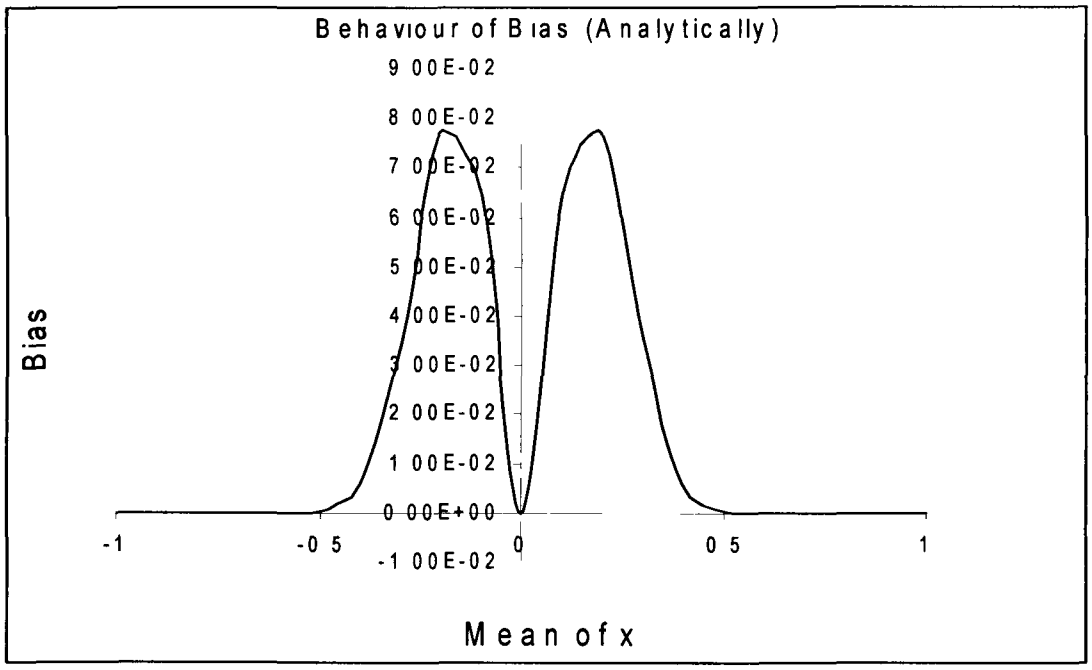


Figure 2.2 Behaviour of Bias(t_5) computed analytically with respect to μ_x for different values of n' and for ρ = 0.8, α = 0.01

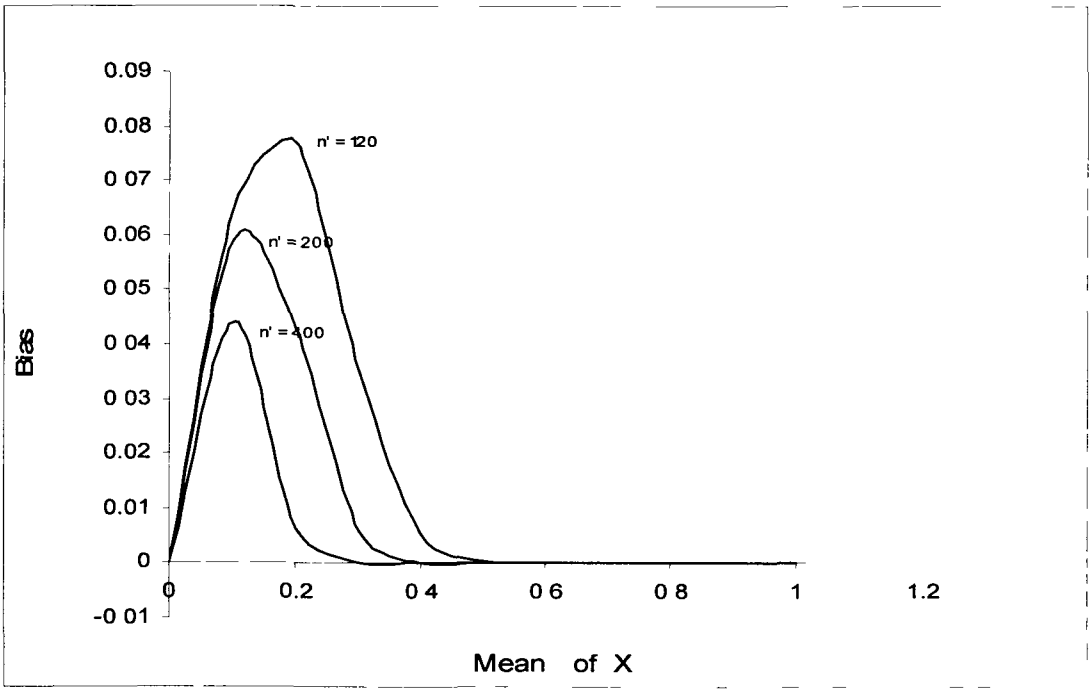


Figure 2.3 Behaviour of Bias(t_5) computed analytically with respect to μ_x for different values of α and for n' = 200 , ρ = 0.8

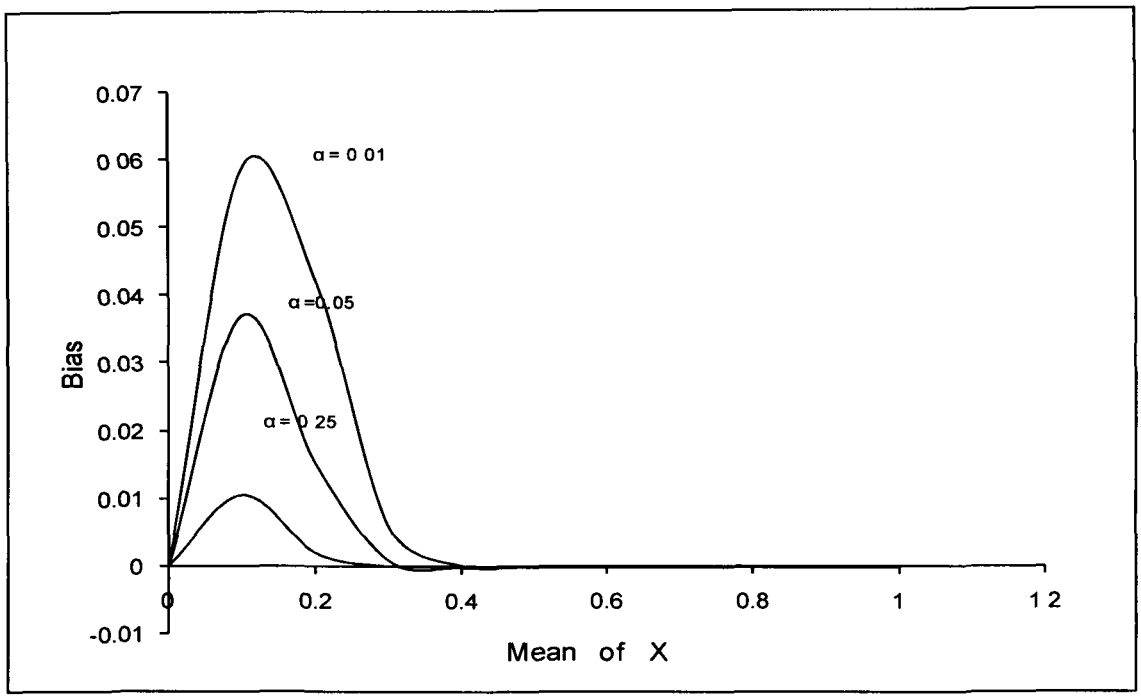


Figure 2.4 Behaviour of Bias(t_5) computed analytically with respect to μ_x for different values of ρ and for n' = 200 , α = 0.01

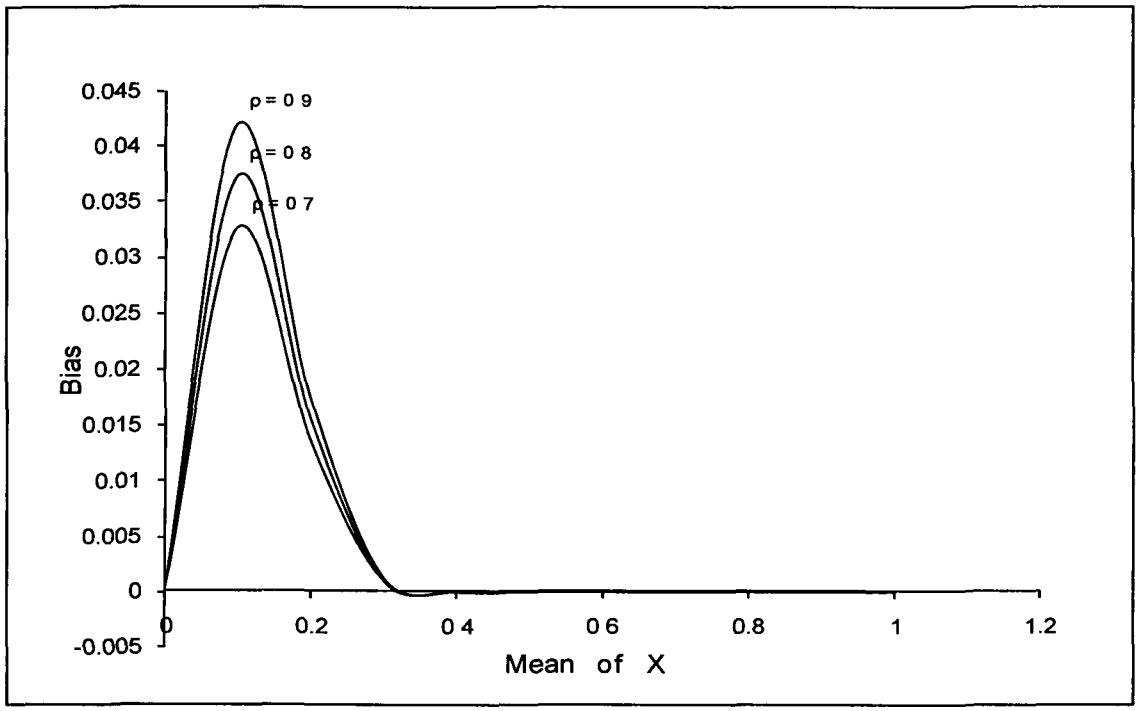


Figure 2.5 Behaviour of Bias(t_5) with respect to μ_x for different refinements k, of the interval of integration and for $\alpha = 0.01$, $\rho = 0.8$, $n' = 200$.

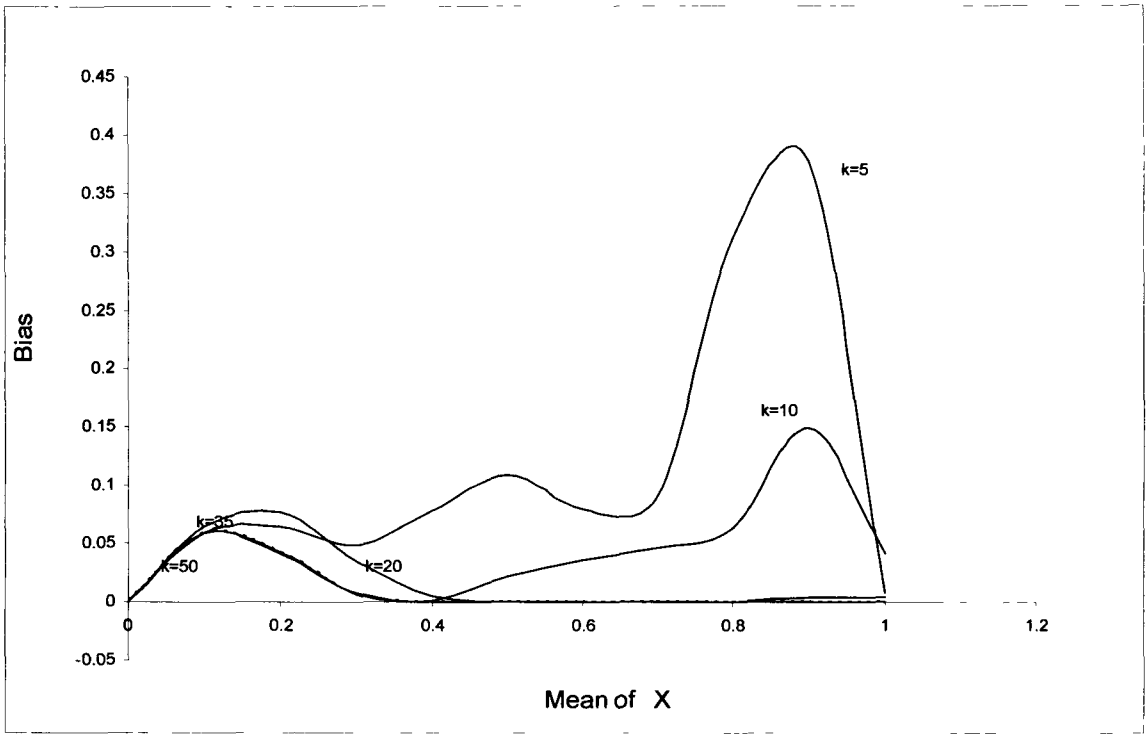


Figure 2.6 Behaviour of the function g(w) with respect to w for different values of μ_x and for $n' = 200$

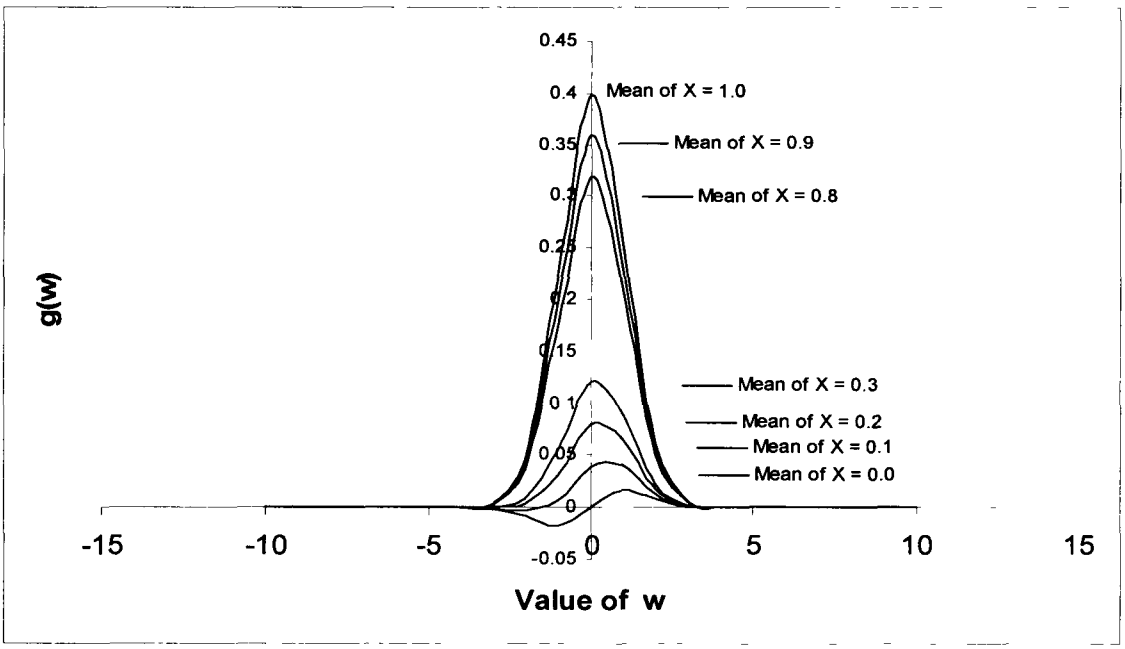


Figure 2.7 Behaviour of Bias(t_5) computed numerically with respect to μ_x for different values of α and for n' = 200, ρ = 0.8

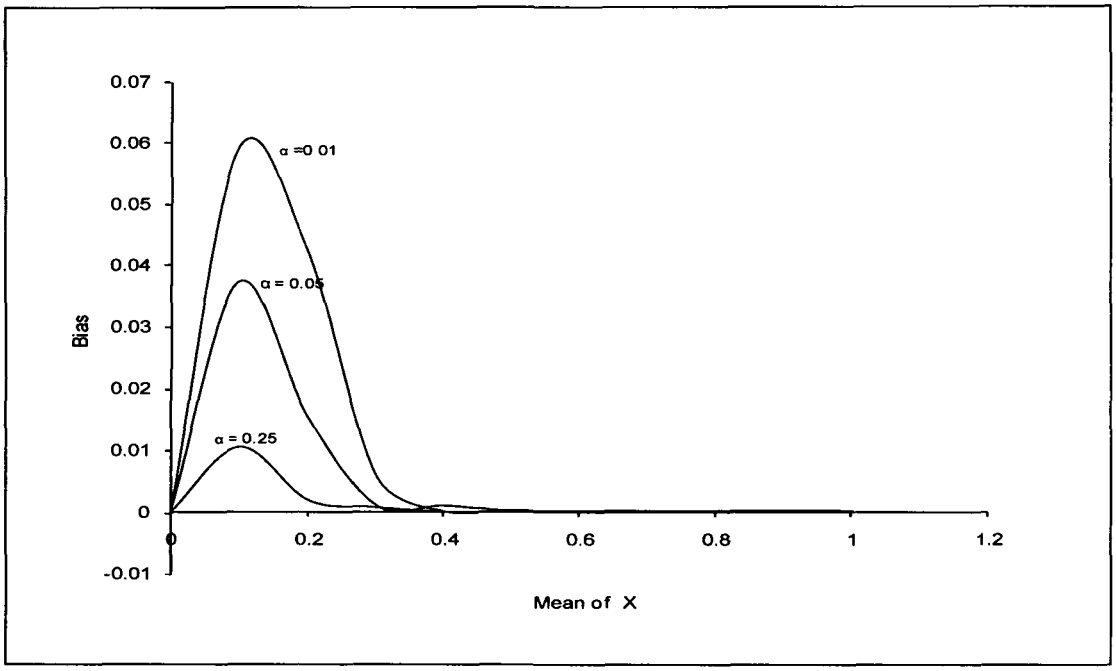


Figure 2.8 Behaviour of Bias(t_5) computed numerically with respect to μ_x for different values of ρ and for n' = 200, α = 0.01

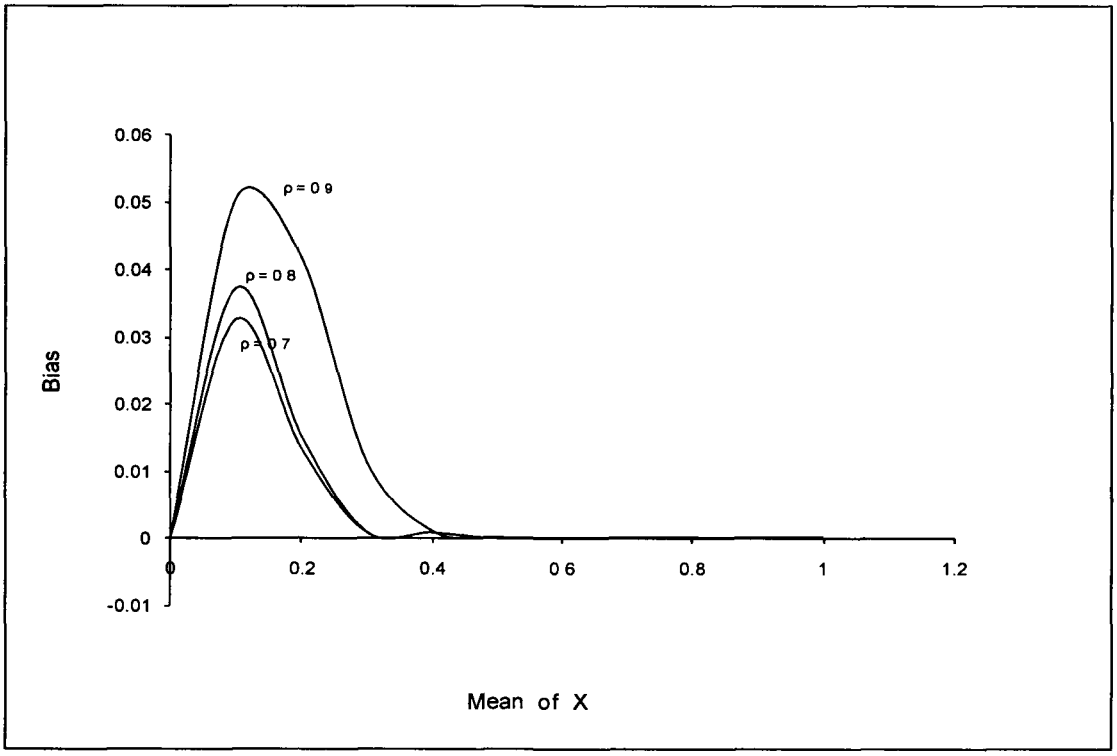
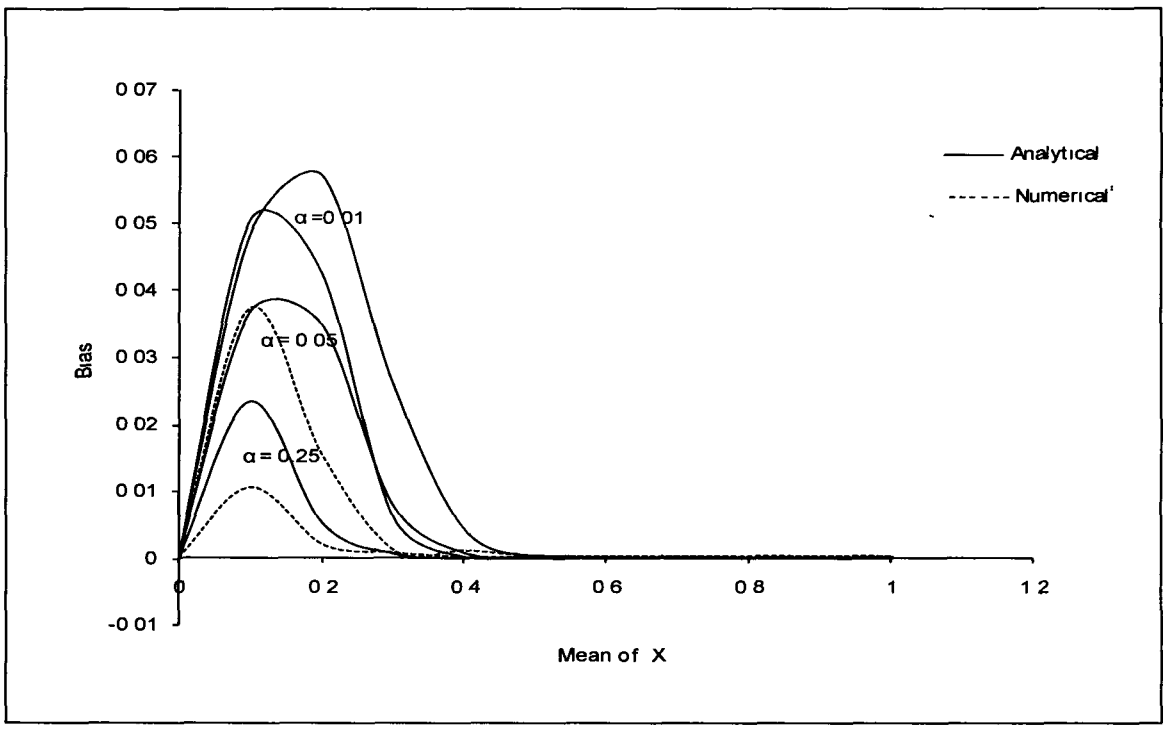


Figure 2.9 Comparative behaviour of Bias(t_5) with respect to μ_x for different values of α and for $\rho = 0.8$, $n' = 200$



Chapter 3

Mean square error function of the CRPTE in double sampling with partial information on the auxiliary variable.

3.1 Introduction

The precision or a measure of the closeness of the sample estimates to the census count taken under identical conditions is judged in sampling theory by the variances of the estimators concerned. Here reliance is placed on the fact that with a small variance, the probability of large deviation from the census count will be small. The general principle is to use estimators which will give the highest concentration of the sample estimates (in the sense of probability) around the valued aimed for. With unbiased estimators the method used for judging the degree of concentration is the variance of the estimators.

It may happen sometimes that the degree of concentration of the sample around the valued aimed at is higher for the distribution of a biased estimator than for an unbiased one. In such a situation the biased estimator is preferable to the unbiased one. However in order to compare a biased estimator with an unbiased estimator, or two estimator with different amounts of bias, variance is not a satisfactory criterion, since it measures deviation from the expected value of the estimator, which is not the same as the population value. A useful criterion is the mean square error (MSE) of the estimator, measured from the population value that is being estimated.

3.2 Mean square error function of CRPTE

To obtain the MSE of t_5, we proceed as follows;

The proposed CRPTE in double sampling is given by

$$t_5 = \begin{cases} (\bar{y}_{st} - \rho \bar{x}_{st}) & if \quad \left|\bar{x}_{n'}\right| \le Z_\alpha / \sqrt{n'} \\ (\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) & if \quad \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'} \end{cases}$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

and the strata means given by

$$\bar{x}_h = \sum_{i=1}^{n_h} x_{hi} / n_h \quad \text{and} \quad \bar{y}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$$

We have assumed that the auxiliary variable X and the study variable Y are jointly normally distributed with parameters given by $(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$. The marginal distributions which is the distribution of the study variable Y and the auxiliary variable X will also follow normal distribution Y ~ $N(\mu_y, \sigma_y^2)$ and X ~ $N(\mu_x, \sigma_x^2)$. The strata population (X_h, Y_h) being carved out from the parent population are also jointly assumed to follow the bivariate normal distribution with parameters written as $\left(\mu_{x_h}, \mu_{y_h}, \sigma_{x_h}^2, \sigma_{y_h}^2, \rho\right)$. The regression estimator depends on weather the covariance matrix is known or not. If known, one may let $\sigma_x^2 = \sigma_y^2 = 1$ without loss of generality (WLOG).

Since the population is assumed to follow normal distribution, the preliminary sample obtained to collect information on the auxiliary variable for the estimation of $\bar{x}_{n'}$, is also assumed to follow normal distribution and therefore $\bar{x}_{n'} \sim N(\mu_x, \sigma_x^2 / n')$ and under the assumption $\sigma_x^2 = \sigma_y^2 = 1$, $\bar{x}_{n'} \sim N(\mu_x, 1/n')$. Further marginal distributions of X_h and Y_h are also normal given as $X_h \sim N(\mu_{x_h}, \sigma_{x_h}^2)$ and $Y_h \sim N(\mu_{y_h}, \sigma_{y_h}^2)$.

The derivation of $MSE(t_5)$ involves conditional expectations, the condition being the acceptance or rejection of the hypothesis considered in the

preliminary test. Further the expectations can be obtained from the integrals involving probability density functions which are assumed to be normal. To obtain the MSE of t_5, we proceed as follows;

$$MSE(t_5) = \text{var}(t_5) + \{Bias(t_5)\}^2$$

$$= E(t_5^2) - \{E(t_5)\}^2 + \{Bias\ (t_5)\}^2 \qquad \text{.....(3.1)}$$

Now,

$$E\ (t_5^{\,2}) = E\ (t_5^{\,2} \quad \text{if} \quad |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'})$$
$$+ E\ (t_5^{\,2} \ \text{if} \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'})$$

$$= \{E\ (\bar{y}_{st} - \rho \bar{x}_{st})^2 \quad \text{if} \quad |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'}\}$$
$$+ [E\{\bar{y}_{st} + \rho\ (\bar{x}_{n'} - \bar{x}_{st})\}^2 \quad \text{if} \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}]$$

$$= \{E(\bar{y}_{st} - \rho \bar{x}_{st})^2 \quad \text{if} \quad |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'}\}$$
$$+ \left[E\{\bar{y}_{st}^{\,2} + 2\rho \bar{y}_{st}(\bar{x}_{n'} - \bar{x}_{st}) + \rho^2(\bar{x}_{n'} - \bar{x}_{st})^2\} \ \text{if} \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right]$$

$$= E(\bar{y}_{st} - \rho \bar{x}_{st})^2$$
$$+ E\{(\rho^2 \bar{x}_{n'}^{\,2} - 2\rho^2 \bar{x}_{n'} \bar{x}_{st} + 2\rho \bar{x}_{n'} \bar{y}_{st}) \quad \text{if} \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$$

i.e

$$E(t_5^2) = E(\bar{y}_{st}^{\,2}) - 2\rho E(\bar{x}_{st} \bar{y}_{st}) + \rho^2 E(\bar{x}_{st}^{\,2})$$
$$+ E\{(\rho^2 \bar{x}_{n'}^{\,2} - 2\rho^2 \bar{x}_{n'} \bar{x}_{st} + 2\rho \bar{x}_{n'} \bar{y}_{st}) \ \text{if} \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$$

Thus,

$$E(t_5^{\,2}) = Var\ (\bar{y}_{st}) + \{E(\bar{y}_{st})\}^2 + \rho^2 \left[(Var\ (\bar{x}_{st}) + \{E(\bar{x}_{st})\}^2\right]$$
$$- 2\rho^2 \sqrt{Var\ (\bar{x}_{st})} \sqrt{Var\ (\bar{y}_{st})} - 2\rho E(\bar{x}_{st}) E(\bar{y}_{st})$$
$$+ E\{(\rho^2 \bar{x}_{n'}^{\,2} - 2\rho^2 \bar{x}_{n'} \bar{x}_{st} + 2\rho \bar{x}_{n'} \bar{y}_{st}) \ \text{if} \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$$
$$\text{.....(3.2)}$$

To evaluate the MSE of t_5, we consider that the joint distribution of $(\bar{x}_{n'}, \bar{x}_{st}, \bar{y}_{st})$ is a multivariate normal with mean (μ_x, μ_x, μ_y) and covariance matrix given by

$$\sum = \begin{pmatrix} 1/n' & 1/n' & \rho/n' \\ 1/n' & 1/n & \rho/n \\ \rho/n' & \rho/n & 1/n \end{pmatrix}$$

(under the assumptions considered in chapter 2)

3.2.1 Evaluation of $E\{(\bar{x}_{n'}^2) \quad if \quad |\bar{x}_{n'}| > Z_\alpha / n'\}$

Let $I_1 = \{E(\bar{x}_{n'}^2) \quad if \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$

$$= \{E(\bar{x}_{n'}^2) \quad if \quad \bar{x}_{n'} > Z_\alpha / \sqrt{n'}\} + \{E(\bar{x}_{n'}^2) \quad if \quad \bar{x}_{n'} < -Z_\alpha / \sqrt{n'}\}$$

$$= \int_{Z_\alpha / \sqrt{n'}}^{\infty} \bar{x}_{n'}^2 f(\bar{x}_{n'}) d\bar{x}_{n'} + \int_{-\infty}^{-Z_\alpha / \sqrt{n'}} \bar{x}_{n'}^2 f(\bar{x}_{n'}) d\bar{x}_{n'}$$

Since $\bar{x}_{n'} \sim N(\mu_x, 1/n')$ we get

$$f(\bar{x}_{n'}) = \sqrt{\frac{n'}{2\pi}} Exp\left\{-\frac{1}{2}\left((\bar{x}_{n'} - \mu_x)(1/\sqrt{n'})\right)^2\right\}$$

Putting $w = (\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}) \Rightarrow dw = \sqrt{n'} d\bar{x}_{n'}$, we have

When $\bar{x}_{n'} = Z_\alpha / \sqrt{n'}$ then $w = \sqrt{n'}(Z_\alpha / \sqrt{n'} - \mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$ and

When $\bar{x}_{n'} = -Z_\alpha / \sqrt{n'}$ then $w = \sqrt{n'}(-Z_\alpha / \sqrt{n'} - \mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

Hence I_1 becomes

$$= (1/\sqrt{2\pi})\left[\int_{A}^{\infty}\{\mu_x + (w/\sqrt{n'})\}^2 Exp\left[-(1/2)w^2\right]dw + \int_{-\infty}^{B}\{\mu_x + (w/\sqrt{n'})\}^2 Exp\left[-(1/2)w^2\right]dw\right]$$

$$= (\mu_x{}^2 + 1/n')\{(1 - \Phi(A) + \Phi(B)\}$$
$$+ (2\mu_x/\sqrt{n'})\{\varphi(A) - \varphi(B)\} + (1/n')\{A\varphi(A) - B\varphi(B)\}$$

.....(3.3)

where $\Phi(.)$ is the cumulative distribution of N(0,1) and $\varphi(.)$ is the density function.

(Detail derivation is given in Appendix 3)

3.2.2 Evaluation of $E(\bar{x}_{n'}\bar{x}_{st} \;\; if \;\; |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'})$

Let $I_2 = \{E(\bar{x}_{n'}\bar{x}_{st}) \;/\; |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}\} = \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\{E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}})\} \;\; if \;\; |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}\right]_{t_1=t_2=0}$

Here $E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}})$ is the moment generating function of $(\bar{x}_{st}, \bar{x}_{n'})$.

Again let

$$I_2' = \{E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}}) \;\; if \;\; |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}\}$$

$$= \left[E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}}) \;\; if \;\; \bar{x}_{n'} > Z_\alpha/\sqrt{n'}\right] + \left[E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}}) \;\; if \;\; \bar{x}_{n'} < -Z_\alpha/\sqrt{n'}\right]$$

$$= \int_{\bar{x}_{st}=-\infty}^{\infty} \int_{\bar{x}_{n'}=Z_\alpha/\sqrt{n'}}^{\infty} e^{t_1 \bar{x}_{st} + t_2 \bar{x}_{n'}} f(\bar{x}_{st}, \bar{x}_{n'}) d\bar{x}_{st} d\bar{x}_{n'}$$

$$+ \int_{\bar{x}_{st}=-\infty}^{\infty} \int_{\bar{x}_{n'}=-\infty}^{-Z_\alpha/\sqrt{n'}} e^{t_1 \bar{x}_{st} + t_2 \bar{x}_{n'}} f(\bar{x}_{st}, \bar{x}_{n'}) d\bar{x}_{st} d\bar{x}_{n'}$$

where $f(\bar{x}_{st}, \bar{x}_{n'})$ is the bivariate normal probability density function of the pair $(\bar{x}_{st}, \bar{x}_{n'})$ with mean (μ_x, μ_x) and the variance covariance matrix given by

$$\sum_1 = \begin{bmatrix} 1/n & 1/n' \\ 1/n' & 1/n' \end{bmatrix}$$ Under the assumption that $\sigma_{x_h}^2 = \sigma_{v_h}^2 = 1$, $\sigma_x^2 = \sigma_y^2 = 1$ (WLOG)

For a bivariate normal density we are given that

$$f(x,y) = \{1/(2\pi|\Sigma|^{1/2})\} Exp\left[(-1/2)\{(x-\mu_x)\ (y-\mu_y)\}(\Sigma^{-1})\begin{pmatrix}(x-\mu_x)\\(y-\mu_y)\end{pmatrix}\right]$$

Here $$\sum_1^{-1} = n'/(n'-n)\begin{pmatrix} n & -n \\ -n & n' \end{pmatrix}$$

Thus,

$$f(\bar{x}_{st},\bar{x}_{n'}) = \frac{1}{2\pi\{(1/nn')-(1/n'^2)\}^{1/2}} Exp\left[-\frac{1}{2}\left\{n(\bar{x}_{st}-\mu_x)^2 - 2\sqrt{\frac{n}{n'}}\frac{(\bar{x}_{st}-\mu_x)}{1/\sqrt{n}}\frac{(\bar{x}_{n'}-\mu_x)}{1/\sqrt{n'}} + n'(\bar{x}_{n'}-\mu_x)^2\right\}\right]$$

Letting $(\bar{x}_{st}-\mu_x)/(1/\sqrt{n}) = x_1$ and $(\bar{x}_{n'}-\mu_x)/(1/\sqrt{n'}) = x_2$,

we get $d\bar{x}_{st} = dx_1/\sqrt{n}$ and $d\bar{x}_{n'} = dx_2/\sqrt{n'}$

When $\bar{x}_{n'} = Z_\alpha/\sqrt{n'}$ then $x_2 = \sqrt{n'}(Z_\alpha/\sqrt{n'} - \mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$ and
when $\bar{x}_{n'} = -Z_\alpha/\sqrt{n'}$ then $x_2 = \sqrt{n'}(-Z_\alpha/\sqrt{n'} - \mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

Hence,

$$I_2' = \left[\{\sqrt{n'}/(2\pi\sqrt{n'-n})\}\int_{-\infty}^{\infty}\int_{x_2=A}^{\infty} Exp\left[t_1\{\mu_x + (x_1/\sqrt{n})\} + t_2\{\mu_x + (x_2/\sqrt{n'})\}\right] * Exp\left[-\{n'/2(n'-n)\}\{x_1^2 - 2\sqrt{n/n'}x_1x_2 + x_2^2\}\right]dx_1dx_2\right.$$

$$\left. + \{\sqrt{n'}/(2\pi\sqrt{n'-n})\}\int_{-\infty}^{\infty}\int_{-\infty}^{x_2=B} Exp\left[t_1\{\mu_x + (x_1/\sqrt{n})\} + t_2\{\mu_x + (x_2/\sqrt{n'})\}\right] Exp\left[-\{n'/2(n'-n)\}\{x_1^2 - 2\sqrt{n/n'}x_1x_2 + x_2^2\}\right]dx_1dx_2\right] \quad \text{(Appendix 4)}$$

Now we can rewrite

$$\{x_1^{\,2} - 2\sqrt{n/n'}x_1x_2 + x_2^{\,2}\} - (2(n'-n)/n')((t_1x_1/\sqrt{n}) + (t_2x_2/\sqrt{n'}))$$
$$= \{(x_1 - (\sqrt{n/n'})x_2) - (1-n/n')t_1/\sqrt{n}\}^2$$
$$+ (1-n/n')\{x_2 - (\sqrt{n/n'})t_1/\sqrt{n} - t_2/\sqrt{n'}\}^2 - t_1^2/n - t_2^2/n' - 2(\sqrt{n/n'})t_1t_2(1/\sqrt{nn'})$$

Substituting

$$(x_1 - (\sqrt{n/n'})x_2) - (1-n/n')(t_1/\sqrt{n}) = \{(n'-n)/n'\}^{1/2}u \qquad \text{and}$$

$$x_2 - (\sqrt{n/n'})(t_1/\sqrt{n}) - (t_2/\sqrt{n'}) = v$$

Then in this transformation, the Jacobian is given by

$$J = \frac{\partial(u,v)}{\partial(x_1,x_2)} = \sqrt{1 - (\tfrac{n}{n'})} \qquad \text{and} \qquad dudv = |J|dx_1dx_2$$

When

$$x_2 = A \quad , \quad v = A - (t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = A - \{(t_1+t_2)/\sqrt{n'}\} = A'$$

$$x_2 = B \quad , \quad v = B - (t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = B - \{(t_1+t_2)/\sqrt{n'}\} = B'$$

Thus,

$$I_2' = \{\sqrt{n'}/(2\pi\sqrt{n'-n})\}Exp(t_1\mu_x + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{v=A'}^{\infty} Exp[(-(n'/2(n'-n))[((n'-n)/n')u^2 + (1-n/n')\{v^2 - (t_1^2/n) - (t_2^2/n') - (2t_1t_2/n')\}]\{(\sqrt{n'-n})/n'\}dudv + \int_{-\infty}^{\infty}\int_{-\infty}^{v=B'} Exp[(-(n'/2(n'-n))[((n'-n)/n')u^2 + (1-n/n')\{v^2 - (t_1^2/n) - (t_2^2/n') - (2t_1t_2/n')\}]\{(\sqrt{n'-n})/n'\}dudv\right]$$

$$= Exp\{t_1\mu_x + t_2\mu_x + (1/2)\{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\} \left\{(1/2\pi)\int_{u=\infty}^{\infty}\int_{v=A'}^{\infty} Exp(-(1/2)(u^2+v^2))dudv + (1/2\pi)\int_{u-\infty}^{\infty}\int_{v=-\infty}^{B'} Exp(-(1/2)(u^2+v^2))dudv\right\}$$

$$= Exp\{t_1\mu_x + t_2\mu_x + (1/2)\{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\}\{(1 - \Phi(A') + \Phi(B')\}$$

where $\Phi(.)$ is the cumulative distribution function of N(0,1).

Now,

$$I_2 = E(\bar{x}_{n'}, \bar{x}_{n''}) / \ |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}$$

$$= \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{E(e^{t_1\bar{x}_{n'}+t_2\bar{x}_{n''}}) / \ |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}\right\}\right]_{t_1=t_2=0}$$

$$= \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{Exp(t_1\mu_x + t_2\mu_x + (1/2)\left\{(t_1^2/n)+(t_2^2/n')+(2t_1t_2/n')\{1-\Phi(A')+\Phi(B')\}\right\}\right\}\right]_{t_1=t_2=0}$$

Differentiating under an integral sign is given by the formula

$$\xi(y) = \int_{g(y)}^{h(y)} f(x, y)\,dx \quad , \quad \text{then}$$

$$\xi'(y) = \int_{g(y)}^{h(y)} f_y(x,y) + h'(y)f(h(y),y) - g'(y)f(g(y),y)$$

Thus,

$$I_2 = \mu_x^2(1-\Phi(A)+\Phi(B)) + (\mu_x/\sqrt{n'})(\varphi(A)-\varphi(B)) + (1/n')(1-\Phi(A)+\Phi(B))$$
$$+ (\mu_x/\sqrt{n'})(\varphi(A)-\varphi(B)) + (1/\sqrt{2\pi n'})(e^{-A^2/2}(A/\sqrt{n'}) - e^{-B^2/2}(B/\sqrt{n'})) \qquad \text{(see Appendix 4)}$$

$$= \mu_x^2(1-\Phi(A)+\Phi(B)) + (2\mu_x/\sqrt{n'})(\varphi(A)-\varphi(B))$$
$$+ (1/n')(1-\Phi(A)+\Phi(B)) + (1/n')(A\varphi(A) - B\varphi(B))$$

Hence,

$$I_2 = E(\bar{x}_{st}, \bar{x}_{n'} \text{ if } |\bar{x}_{n'}| > z_\alpha/\sqrt{n'})$$

$$= (\mu_x{}^2 + 1/n')\{(1-\Phi(A)+\Phi(B)\}$$
$$+ (2\mu_x/\sqrt{n'})\{\varphi(A)-\varphi(B)\} + (1/n')\{A\varphi(A) - B\varphi(B)\}$$

.....(3.4)

where $\Phi(.)$ is the cumulative distribution of N(0,1) and $\varphi(.)$ is the density function.
(Detail derivation given Appendix 4)

3.2.3 Evaluation of $E(\bar{y}_{st}\,\bar{x}_{n'}\ \ if\ \ \left|\bar{x}_{n'}\right| > Z_{\alpha}/\sqrt{n'})$

Let $I_3 = \left\{E(\bar{y}_{st}\,\bar{x}_{n'})\ \ /\ \ \left|\bar{x}_{n'}\right| > Z_{\alpha}/\sqrt{n'}\right\}$

$$= \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})\right\}\ \ if\ \ \left|\bar{x}_{n'}\right| > Z_{\alpha}/\sqrt{n'}\right]_{t_1=t_2=0}$$

Here $E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})$ is the moment generating function $(\bar{y}_{st}, \bar{x}_{n'})$.

Again let

$$I_3' = \left\{E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})if\left|\bar{x}_{n'}\right| > Z_{\alpha}/\sqrt{n'}\right\}$$

$$= \left[E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})\ \ if\ \ \bar{x}_{n'} > Z_{\alpha}/\sqrt{n'}\right] + \left[E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})\ if\ \ \bar{x}_{n'} < -Z_{\alpha}/\sqrt{n'}\right]$$

$$= \int_{\bar{y}_{st}=-\infty}^{\infty} \int_{\bar{x}_{n'}=Z_\alpha/\sqrt{n'}}^{\infty} e^{t_1 \bar{y}_{st} + t_2 \bar{x}_{n'}} f(\bar{y}_{st}, \bar{x}_{n'}) d\bar{y}_{st} d\bar{x}_{n'}$$

$$+ \int_{\bar{y}_{st}=-\infty}^{\infty} \int_{\bar{x}_{n'}=-\infty}^{-Z_\alpha/\sqrt{n'}} e^{t_1 \bar{y}_{st} + t_2 \bar{x}_{n'}} f(\bar{y}_{st}, \bar{x}_{n'}) d\bar{y}_{st} d\bar{x}_{n'}$$

where $f(\bar{y}_{st}, \bar{x}_{n'})$ is the bivariate normal probability density function of the pair $(\bar{y}_{st}, \bar{x}_{n'})$ with mean (μ_y, μ_x) and the variance covariance matrix given by

$$\sum_2 = \begin{bmatrix} 1/n & \rho/n' \\ \rho/n' & 1/n' \end{bmatrix}$$ Under the assumption that $\sigma_{x_n}^2 = \sigma_{y_n}^2 = 1, \sigma_x^2 = \sigma_y^2 = 1$ (WLOG)

For a bivariate normal density we are given that

$$f(x,y) = \{1/(2\pi|\Sigma|^{1/2})\} Exp\left[(-1/2)\{((x-\mu_x)\ (y-\mu_y))(\Sigma^{-1})\begin{pmatrix}(x-\mu_x)\\(y-\mu_y)\end{pmatrix}\right]$$

Here, $$\sum_2^{-1} = n'/(n'-\rho^2 n)\begin{pmatrix} n & -n\rho \\ -n\rho & n' \end{pmatrix}$$

Thus,

$$f(\bar{y}_{st},\bar{x}_{n'})=\frac{1}{2\pi\left\{(1/nn')-(\rho^2/n'^2)\right\}^{1/2}}Exp\left[-\frac{n'}{2(n'-n\rho^2)}\left\{n(\bar{y}_{st}-\mu_y)^2-2\rho\sqrt{\frac{n}{n'}}\frac{(\bar{y}_{st}-\mu_y)}{1/\sqrt{n}}\frac{(\bar{x}_{n'}-\mu_x)}{1/\sqrt{n'}}+n'(\bar{x}_{n'}-\mu_x)^2\right\}\right]$$

Letting $(\bar{y}_{st}-\mu_y)/(1/\sqrt{n})=z_1$ and $(\bar{x}_{n'}-\mu_x)/(1/\sqrt{n'})=z_2$, we get $d\bar{y}_{st}=dz_1/\sqrt{n}$ and $d\bar{x}_{n'}=dz_2/\sqrt{n'}$

When $\bar{x}_{n'}=Z_\alpha/\sqrt{n'}$ then $z_2=\sqrt{n'}(Z_\alpha/\sqrt{n'}-\mu_x)=Z_\alpha-\sqrt{n'}\mu_x=A$ and

when $\bar{x}_{n'}=-Z_\alpha/\sqrt{n'}$ then $z_2=\sqrt{n'}(-Z_\alpha/\sqrt{n'}-\mu_x)=-Z_\alpha-\sqrt{n'}\mu_x=B$

Thus,

$$I_3'=\left[\{\sqrt{n'}/(2\pi\sqrt{n'-\rho^2n})\}\int_{-\infty}^{\infty}\int_{z_2=A}^{\infty}Exp\left[t_1\{\mu_y+(z_1/\sqrt{n})\}+t_2\{\mu_x+(z_2/\sqrt{n'})\}\right]\right.$$
$$Exp\left[-\{n'/2(n'-\rho^2n)\}\{z_1^{\,2}-2\rho\sqrt{n/n'}z_1z_2+z_2^{\,2}\}\right]dz_1dz_2$$

$$\left.+\{\sqrt{n'}/(2\pi\sqrt{n'-\rho^2n})\}\int_{-\infty}^{\infty}\int_{-\infty}^{z_2=B}Exp\left[t_1\{\mu_y+(z_1/\sqrt{n})\}+t_2\{\mu_x+(z_2/\sqrt{n'})\}\right]\right.$$
$$\left.Exp\left[-\{n'/2(n'-\rho^2n)\}\{z_1^{\,2}-2\rho\sqrt{n/n'}z_1z_2+z_2^{\,2}\}\right]dz_1dz_2\right]$$

Now we can rewrite

$$\{z_1^{\,2} - 2\rho\sqrt{n/n'}z_1z_2 + z_2^{\,2}\} - \left\{2(n'-n\rho^2)/n')((t_1z_1/\sqrt{n}) + (t_2z_2/\sqrt{n'})\right\}$$
$$= \left\{(z_1 - \rho(\sqrt{n/n'})z_2) - (1-n\rho^2/n')t_1/\sqrt{n}\right\}^2$$
$$+ (1-n\rho^2/n')\left[\{z_2 - \rho t_1/\sqrt{n'} - t_2/\sqrt{n'}\}^2 - t_1^2/n - t_2^2/n' - 2\rho t_1t_2/n'\right]$$

Substituting

$$(z_1 - (\rho\sqrt{n/n'})z_2) - (1-n\rho^2/n')(t_1/\sqrt{n}) = \{(n'-n\rho^2)/n'\}^{1/2}u \quad \text{and}$$

$$z_2 - \rho(t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = v$$

Then in this transformation, the Jacobian is given by

$$J = \frac{\partial(u,v)}{\partial(x_1,x_2)} = \sqrt{1 - (\frac{\rho^2 n}{n'})} \qquad \text{and} \quad dudv = |J|dz_1dz_2$$

When

$$z_2 = A \;,\; v = A - (\rho t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = A - \{(\rho t_1 + t_2)/\sqrt{n'}\} = A'$$

$$z_2 = B \;,\; v = B - (\rho t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = B - \{(\rho t_1 + t_2)/\sqrt{n'}\} = B'$$

Hence,

$$I_3' = \{\sqrt{n'}/(2\pi\sqrt{n'-\rho^2 n})\}Exp(t_1\mu_y + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{v=A'}^{\infty} Exp\{-(n'/2(n'-\rho^2 n))[((n'-\rho^2 n)/n')u^2 + \{(n'-n\rho^2)/n'\}[v^2 - \{(t_1^2/n)+(t_2^2/n')+(2t_1t_2/n')\}]\sqrt{(n'-\rho^2 n)/n'}\}dudv + \int_{-\infty}^{\infty}\int_{-\infty}^{v=B'} Exp\{-(n'/2(n'-\rho^2 n))[((n'-\rho^2 n)/n')u^2 + \{(n'-n\rho^2)/n'\}[v^2 - \{(t_1^2/n)+(t_2^2/n')+(2t_1t_2/n')\}]\sqrt{(n'-\rho^2 n)/n'}\}dudv\right]$$

$$= Exp\{t_1\mu_y + t_2\mu_x + (1/2)\{(t_1^2/n) + (t_2^2/n') + (2\rho t_1 t_2/n')\} \left\{(1/2\pi)\int_{u=\infty}^{\infty}\int_{v=A'}^{\infty} Exp(-(1/2)(u^2+v^2))dudv + (1/2\pi)\int_{u-\infty}^{\infty}\int_{v=-\infty}^{B'} Exp(-(1/2)(u^2+v^2))dudv\right\}$$

$$= Exp\{t_1\mu_y + t_2\mu_x + (1/2)\{(t_1^2/n) + (t_2^2/n') + (2\rho t_1 t_2/n')\}\{(1 - \Phi(A') + \Phi(B')\}$$

where $\Phi(.)$ is the cumulative distribution of N(0,1) .

Now,

$$I_3 = \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{E(e^{t_1 \bar{y}_{st} + t_2 \bar{x}_{n'}})\right\} \quad if \quad \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'}\right]_{t_1 = t_2 = 0}$$

$$= \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{Exp\,(t_1\mu_y + t_2\mu_x + (1/2)\left\{(t_1^2/n) + (t_2^2/n') + (2\rho t_1 t_2/n')\{1 - \Phi(A') + \Phi(B')\}\right\}\right\}\right]_{t_1 = t_2 = 0}$$

Differentiation under an integral sign is given by the formula

$$\xi(y) = \int_{g(y)}^{h(y)} f(x, y)\,dx \quad , \quad \text{then}$$

$$\xi'(y) = \int_{g(y)}^{h(y)} f_y(x,y) + h'(y) f(h(y),y) - g'(y) f(g(y),y)$$

Differentiating I_3' under an integral sign, we get

$$I_3 = \mu_x\mu_y(1-\Phi(A)+\Phi(B)) + (\rho\mu_x/\sqrt{n'})(\varphi(A)-\varphi(B)) + (\rho/n')(1-\Phi(A)+\Phi(B))$$
$$+ \mu_y(\varphi(A)-\varphi(B))/\sqrt{n'} + (\rho/\sqrt{2\pi n'})(e^{-A^2/2}(A/\sqrt{n'}) - e^{-B^2/2}(B/\sqrt{n'}))$$

$$= (\mu_x\mu_y + \rho/n')(1-\Phi(A)+\Phi(B)) + (1/\sqrt{n'})(\mu_y + \rho\mu_x)(\varphi(A)-\varphi(B))$$
$$+ (\rho/n')(A\varphi(A) - B\varphi(B))$$

Hence,

$$E(\bar{y}_{st}, \bar{x}_{n'} / |\bar{x}_{n'}| > z_\alpha/\sqrt{n'})$$

$$= (\mu_x\mu_y + \rho/n')\{(1-\Phi(A)+\Phi(B)\}$$
$$+ (1/\sqrt{n'})(\mu_y + \rho\mu_x)\{\varphi(A)-\varphi(B)\} + (\rho/n')\{A\varphi(A) - B\varphi(B)\} \quad \text{.....(3.5)}$$

where $\Phi(.)$ is the cumulative distribution of N(0,1) and $\varphi(.)$ is the density function.

(Detail derivation given in Appendix 5)

3.2.4 Evaluation of MSE(t_5)

Now, substituting eqn (3.3),(3.4) and (3.5) in eqn (3.2), we get

$$E(t_5^2) = Var(\bar{y}_{st}) + \{E(\bar{y}_{st})\}^2 + \rho^2\left[(Var(\bar{x}_{st}) + \{E(\bar{x}_{st})\}^2\right]$$
$$- 2\rho^2\sqrt{Var(\bar{x}_{st})}\sqrt{Var(\bar{y}_{st})} - 2\rho E(\bar{x}_{st})E(\bar{y}_{st})$$
$$+ \{(\mu_x^2 + 1/n)(1 - \Phi(A) + \Phi(B)) + (2\mu_x/\sqrt{n'})(\varphi(A) - \varphi(B)$$
$$+ (1/n')(A\ \varphi(A) - B\ \varphi(B))\}\ (\rho^2 - 2\rho^2)$$
$$+ 2\rho(\mu_x\mu_y + \rho/n')(1 - \Phi(A) + \Phi(B))$$
$$+ 2\rho(1/\sqrt{n'})(\mu_y + \rho\mu_x)(\varphi(A) - \varphi(B) + 2(\rho^2/n')\{A\ \varphi(A) - B\ \varphi(B)\}$$

$$= Var(\bar{y}_{st}) + \{E(\bar{y}_{st})\}^2 + \rho^2\left[(Var(\bar{x}_{st}) + \{E(\bar{x}_{st})\}^2\right]$$
$$- 2\rho^2\sqrt{Var(\bar{x}_{st})}\sqrt{Var(\bar{y}_{st})} - 2\rho E(\bar{x}_{st})E(\bar{y}_{st})$$
$$+ \{-(\mu_x^2 + 1/n)\rho^2 + 2\rho(\mu_x\mu_y + \rho/n')\}(1 - \Phi(A) + \Phi(B))$$
$$+ \{-\rho^2(2\mu_x/\sqrt{n'}) + (2\rho/\sqrt{n'})(\mu_y + \rho\mu_x)\}(\varphi(A) - \varphi(B)$$
$$+ \{(-\rho^2/n') + (2\rho^2/n')\}(A\ \varphi(A) - B\ \varphi(B))\}$$

$$E(t_5^2) = Var\ (\bar{y}_{st}) + \{E(\bar{y}_{st})\}^2 + \rho^2\left[(Var\ (\bar{x}_{st}) + \{E(\bar{x}_{st})\}^2\right]$$
$$- 2\rho^2\sqrt{Var\ (\bar{x}_{st})}\sqrt{Var\ (\bar{y}_{st})} - 2\rho E(\bar{x}_{st})E(\bar{y}_{st})$$
$$+ \{(2\ \mu_x\mu_y - \rho\mu_x^2)\rho + (\rho^2/n')\}(1 - \Phi(A) + \Phi(B))$$
$$+ \{(2\rho/\sqrt{n'})\mu_y(\varphi(A) - \varphi(B)) + (\rho^2/n')(A\ \varphi(A) - B\ \varphi(B))$$

The variance of $\bar{x}_{st}$ and $\bar{y}_{st}$ being given by

$$Var(\bar{x}_{st}) = \sum W_h^2 \sigma_{x_h}^2 / n_h \qquad Var(\bar{y}_{st}) = \sum W_h^2 \sigma_{y_h}^2 / n_h$$ (Cochran 1977)

which under the stated assumptions becomes

$$Var(\bar{x}_{st}) = \sum W_h^2 / n_h \qquad Var(\bar{y}_{st}) = \sum W_h^2 / n_h$$ (under the assumption $\sigma_{x_h}^2 = \sigma_{y_h}^2 = 1$)

Therefore,

$$E(t_5^2) = \sum_h W_h^2 / n_h + \mu_y^2 + \rho^2\left\{\sum_h W_h^2 / n_h + \mu_x^2\right\} - 2\rho^2\sum_h W_h^2 / n_h - 2\rho\mu_x\mu_y$$
$$+ \{(2\ \mu_x\mu_y - \rho\mu_x^2)\rho + (\rho^2/n')\}\{1 - \Phi(A) + \Phi(B))$$
$$+ \{(2\rho/\sqrt{n'})\mu_y(\varphi(A) - \varphi(B)) + (\rho^2/n')(A\ \varphi(A) - B\ \varphi(B))$$

when the samples are selected with proportional allocation then the stratum weight is given by $W_h = (N_h/N) = (n_h/n)$

Thus $\sum_h W_h^2 / n_h = \sum_h W_h^2 / nW_h = (1/n)\sum_h W_h = (1/n)$ (as $\sum_h W_h = 1$)

Hence,

$$E(t_5^2) = (1-\rho^2)/n + \mu_y^2 + \rho\mu_x^2 - 2\rho\mu_x\mu_y$$
$$+ \{(2\rho\mu_x\mu_y - \rho^2\mu_x^2 + (\rho^2/n')\}$$
$$- \{2\rho\mu_x\mu_y - \rho^2\mu_x^2 + (\rho^2/n')\}\{\Phi(A) - \Phi(B)\}$$
$$+ \{(2\rho\mu_y/\sqrt{n'})(\varphi(A) - \varphi(B) + (\rho^2/n')(A\ \varphi(A) - B\ \varphi(B))$$

$$= (1-\rho^2)/n + \mu_y^2 + \rho^2/n'$$
$$- \{2\rho\mu_x\mu_y - \rho^2\mu_x^2 + (\rho^2/n')\}\{\Phi(A) - \Phi(B)\}$$
$$+ \{(2\rho\mu_y/\sqrt{n'})(\varphi(A) - \varphi(B) + (\rho^2/n')(A\ \varphi(A) - B\ \varphi(B)) \quad \text{.....(3.6)}$$

Mean square error of t_5 is given by

$$MSE(t_5) = E(t_5^2) - \{E(t_5)\}^2 + \{E(t_5) - \mu_y\}^2 \qquad \text{from (3.1)}$$

$$= E(t_5^2) - 2E(t_5)\mu_y + \mu_y^2 \qquad \text{.....(3.7)}$$

and from Chapter 2

$$E(t_5) = \mu_y - \rho\mu_x\{\Phi(A) - \Phi(B)\} + \rho(1/\sqrt{n'})\{\varphi(A) - \varphi(B)\}\ldots \qquad \ldots(3.8)$$

Therefore, substituting (3.6) and (3.8) in (3.7)

$$\begin{aligned} MSE\ (t_5) &= (1-\rho^2)/n + \mu_y{}^2 + \rho^2/n' \\ &\quad - \{2\rho\mu_x\mu_y - \rho^2\mu_x{}^2 + (\rho^2/n')\}\{\Phi(A) - \Phi(B)\} \\ &\quad + \{(2\rho\mu_y/\sqrt{n'})(\varphi(A) - \varphi(B) + (\rho^2/n')(A\ \varphi(A) - B\ \varphi(B)) \\ &\quad - 2\{\mu_y - \rho\mu_x(\Phi(A) - \Phi(B)) + (\rho/\sqrt{n'})(\varphi(A) - \varphi(B))\}\ \mu_y + \mu_y{}^2 \end{aligned}$$

Thus,

$$\begin{aligned} MSE(t_5) = \{(1-\rho^2)/n + \rho^2/n'\} + (\rho^2/n')\{A\varphi(A) - B\varphi(B)\} \\ - \rho^2(1/n' - \mu_x{}^2)\{\Phi(A) - \Phi(B)\} \end{aligned} \qquad \ldots\ldots(3.9)$$

The Mean square error function of t_5 given by equation (3.9) shows that its behavior can be observed for different values of μ_x. As MSE(t_5) is symmetric about $\mu_x = 0$, hence we need to consider the behavior only for $\mu_x \geq 0$.This above behavior can be studied for selected values of the level of significance α and correlation coefficient ρ and by fixing some hypothetical values for n and n'.

3.3 Discussion

The above Mean squared error of t_5 is given as

$$MSE(t_5) = \{(1-\rho^2)/n + \rho^2/n'\} + (\rho^2/n')\{A\varphi(A) - B\varphi(B)\} - \rho^2(1/n' - \mu_x{}^2)\{\Phi(A) - \Phi(B)\}$$

$$= g_1 + h_1, \quad \text{where}$$

$g_1 = \{(1-\rho^2)/n + \rho^2/n'\}$ and

$$h_1 = (\rho^2/n')\{A\varphi(A) - B\varphi(B)\} - \rho^2(1/n' - \mu_x{}^2)\{\Phi(A) - \Phi(B)\}$$

We note that g_1 is the MSE of t_4 , the combined linear regression estimator in double sampling (Appendix 6) , when information on μ_x is not known.

The values of $MSE(t_5)$ can be easily computed for different values of μ_x. In order to get an idea about the behavior of the mean square error function with respect to μ_x, $MSE(t_5)$ is computed for a set of values of n, n', α and ρ which are given in Table 3.1 – 3.2 and Figure 3.1 – 3.2. It is found in general that $MSE(t_5)$ is minimum at $\mu_x = 0$. As μ_x is increases, the $MSE(t_5)$ increases to a maximum and then gradually decreases and then becomes constant . The figures clearly show that when the mean of the auxiliary variable is close to the

hypothetical value, then the $MSE(t_5)$ is minimum. Also as μ_x moves away from the hypothetical value the $MSE(t_5)$ increases, but after attaining maximum again gradually decreases and then becomes constant. This establishes the utility of the present study that the use of partial information and preliminary test of the auxiliary variable reduces the $MSE(t_5)$ of the proposed estimator.

3.4 Mean square error function of CRPTE computed numerically

It is seen that the analytical method of determining the MSE involves evaluating the mathematical expectation of the random variables like $E(\bar{x}_{n'}^2)$, $E(\bar{x}_{n'}\bar{x}_{st})$ and $E(\bar{x}_{n'}\bar{y}_{st})$. The derivation of these expectation is done using moment generating function and also involves the application of single and double integration technique. In the evaluation, using bivariate frequency distributions, a tedious substitution of change of variables is necessary to simplify the integral. The above expectation is finally obtained by differentiating under the integral sign.

The above analytical derivation in the evaluation of MSE is tedious. An alternative method is sought with the help of numerical techniques. The mathematical expectations of the random variables like $E(\bar{x}_{n'}^2)$, $E(\bar{x}_{n'}\bar{x}_{st})$ and $E(\bar{x}_{n'}\bar{y}_{st})$ can be evaluated without the use of moment generating function and also avoid the complex substitution. By the use of numerical technique the steps of differentiating under the integral sign can also be avoided. The double integral involves in the evaluation of the above mentioned expectations can be

calculated by using a numerical integration technique for a function in two variables $f(x,y)$, given by simpson 1/3rd rule.

With the advent and rapid development of high speed digital computers and the increasing desire for accurate and faster solution to applied problems, it has become possible to evaluate complex mathematical formulation numerically in a fewer number of steps in a short duration of time. In this study, computers can be used for evaluating the above numerical integrals with the help of programs written in Fortran 77.

By the definition of t_5

$$t_5 = \begin{cases} (\bar{y}_{st} - \rho \bar{x}_{st}) & if \quad |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'} \\ (\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) & if \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n} \end{cases}$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

and the stratum means given by

$$\bar{x}_h = \sum_{i=1}^{n_h} x_{hi} / n_h \quad \text{and} \quad \bar{y}_h = \sum_{i=1}^{n_h} y_{hi} / n_h$$

As before to obtain MSE of t_5, we notice that

$$MSE\ (t_5) = E(t_5^{\,2}) - \{E(t_5)\}^2 + \{Bias\ (t_5)\}^2$$

$$\Rightarrow MSE\ (t_5) = E(t_5^{\,2}) - 2E(t_5)\mu_y + \mu_y^2$$

Now,

$$E(t_5^2) = E(t_5^2 \quad \text{if} \quad |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'}) + E(t_5^2 \ \text{if} \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'})$$

After simplification we get,

$$E(t_5^2) = Var(\bar{y}_{st}) + \{E(\bar{y}_{st})\}^2 + \rho^2 \left[(Var(\bar{x}_{st}) + \{E(\bar{x}_{st})\}^2 \right] - 2\rho^2 \sqrt{Var(\bar{x}_{st})} \sqrt{Var(\bar{y}_{st})} - 2\rho E(\bar{x}_{st}) E(\bar{y}_{st}) + E\{(\rho^2 \bar{x}_{n'}^2 - 2\rho^2 \bar{x}_{n'} \bar{x}_{st} + 2\rho \bar{x}_{n'} \bar{y}_{st}) \ \text{if} \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$$

Now, substituting

$I_1 = E(\bar{x}_{n'}^2 \ \ if \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'})$, $I_2 = \{E(\bar{x}_{n'} \bar{x}_{st}) \ / \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$ and

$I_3 = \{E(\bar{x}_{n'} \bar{y}_{st}) \ / \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$ in $E(t_5^2)$ and thereafter $E(t_5^2)$ in the $MSE(t_5)$, we get

$$MSE(t_5) = (1/n) + \mu_y^2 + \rho^2\{(1/n) + \mu_x^2\} - (2\rho^2 / n) - 2\rho\mu_x\mu_y + [\rho^2 I_1 - 2\rho^2 I_2 + 2\rho I_3] - 2\{\rho(I - \mu_x) + \mu_y\}\mu_y + \mu_y^2$$

where $I = E(\bar{x}_{n'} \ \ if \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'})$

Thus,

$$MSE(t_5) = (1/n) + \rho^2\{(1/n) + \mu_x^2\} - (2\rho^2 / n) + [\rho^2 I_1 - 2\rho^2 I_2 + 2\rho I_3] - 2\rho I \mu_y$$

The mean square error function derived above involves the population parameter μ_y, the mean of the study variable which is needed to be estimated. Thus, in order to eliminate this parameter, the analytical result of I_3 from eqn. (3.5) and I from section 2.3 are substituted in the above MSE function to obtain the required result given as follows;

$$MSE\ (t_5) = (1/n) + \rho^2\{(1/n) + \mu_x{}^2\} - (2\rho^2/n) + \rho^2 I_1 - 2\rho^2 I_2$$
$$+ 2\rho\left\{(\mu_x\mu_y + (\rho/n))(1 - \Phi(A) + \Phi(B))\right\}$$
$$+ \left\{(1/\sqrt{n'})(\mu_y + \rho\mu_x)(\varphi(A) - \varphi(B))\right\}$$
$$+ (\rho/n')(A\varphi(A) - B\varphi(B))$$
$$- 2\rho\mu_y\left\{\mu_x(1 - \Phi(A) + \Phi(B)) + (1/\sqrt{n'})(\varphi(A) - \varphi(B))\right\}$$

$$= (1/n) + \rho^2\{(1/n) + \mu_x{}^2\} - (2\rho^2/n) + \rho^2 I_1 - 2\rho^2 I_2$$
$$+ (2\rho^2/n')(1 - \Phi(A) + \Phi(B))$$
$$+ (2\rho^2\mu_x/\sqrt{n'})(\varphi(A) - \varphi(B))$$
$$+ (2\rho^2/n')(A\varphi(A) - B\varphi(B))$$

.....(3.10)

In the above MSE function $\Phi(.)$, the cumulative distribution function of N(0,1) and $\varphi(.)$ its density function, are obtained from normal tables.

3.4.1 Numerical computation of I_1

Now $I_1 = E(\bar{x}_{n'}{}^2 \ \ if\ |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'})$

$$= \{E(\bar{x}_{n'}{}^2)\ \ if\ \ \bar{x}_{n'} > Z_\alpha/\sqrt{n'}\} + \{E(\bar{x}_{n'}{}^2)\ \ if\ \ \bar{x}_{n'} < -Z_\alpha/\sqrt{n'}\}$$

$$= \int_{Z_\alpha/\sqrt{n'}}^{\infty} \bar{x}_{n'}{}^2 f(\bar{x}_{n'})d\bar{x}_{n'} + \int_{-\infty}^{-Z_\alpha/\sqrt{n'}} \bar{x}_{n'}{}^2 f(\bar{x}_{n'})d\bar{x}_{n'}$$

Since $\bar{x}_{n'} \sim N(\mu_x, 1/n')$ we get

$$f(\bar{x}_{n'}) = \sqrt{\frac{n'}{2\pi}}Exp\left\{-\frac{1}{2}\left((\bar{x}_{n'} - \mu_x)(1/\sqrt{n'})\right)^2\right\}$$

Thus the integrals becomes

$$I_1 = (\sqrt{n'}/\sqrt{2\pi}) \int_{Z_\alpha/\sqrt{n'}}^{\infty} \bar{x}_{n'}{}^2 Exp\left[-(1/2)\{(\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'})\}^2\right] d\bar{x}_{n'}$$

$$+ (\sqrt{n'}/\sqrt{2\pi}) \int_{-\infty}^{-Z_\alpha/\sqrt{n'}} \bar{x}_{n'}{}^2 Exp\left[-(1/2)\{(\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'})\}^2\right] d\bar{x}_{n'}$$

Putting $w = (\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}) \Rightarrow dw = \sqrt{n'} d\bar{x}_{n'}$, we have

When $\bar{x}_{n'} = Z_\alpha/\sqrt{n'}$ then $w = \sqrt{n'}(Z_\alpha/\sqrt{n'} - \mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$
and when $\bar{x}_{n'} = -Z_\alpha/\sqrt{n'}$ then $w = \sqrt{n'}(-Z_\alpha/\sqrt{n'} - \mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

Therefore,

$$I_1 = (1/\sqrt{2\pi})\left[\int_A^{\infty} \{\mu_x + (w/\sqrt{n'})\}^2 Exp\left[-(1/2)w^2\right] dw + \int_{-\infty}^{B} \{\mu_x + (w/\sqrt{n'})\}^2 Exp\left[-(1/2)w^2\right] dw\right]$$

or $I_1 = I_{11} + I_{12}$ .....(3.11)

The integrals I_{11} and I_{12} are evaluated by the use of Fortran 77 and the output values are given in Table 3.3. (Appendix 7 and 8)

3.4.2 Numerical computation of I_2

$$I_2 = \left\{E(\bar{x}_{n'}\bar{x}_{st}) \quad / \quad |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}\right\}$$
$$= \{E(\bar{x}_{n'}\bar{x}_{st})/\bar{x}_{n'} > Z_\alpha/\sqrt{n'}\} + \{E(\bar{x}_{n'}\bar{x}_{st})/\bar{x}_{n'} < -Z_\alpha/\sqrt{n'}\}$$

$$= \int_{-\infty}^{\infty}\int_{\bar{x}_{n'}=Z_\alpha/\sqrt{n'}}^{\infty}\bar{x}_{st}\,\bar{x}_{n'}\,f(\bar{x}_{st},\bar{x}_{n'})\,d\bar{x}_{st}\,d\bar{x}_{n'} + \int_{-\infty}^{\infty}\int_{\bar{x}_{n'}=-\infty}^{-Z_\alpha/\sqrt{n'}}\bar{x}_{st}\,\bar{x}_{n'}\,f(\bar{x}_{st},\bar{x}_{n'})\,d\bar{x}_{st}\,d\bar{x}_{n'}$$

where

$$f(\bar{x}_{st},\bar{x}_{n'}) = \frac{1}{2\pi\{(1/nn') - (1/n'^2)\}^{1/2}}\,Exp\left[-\frac{1}{2}\left\{n(\bar{x}_{st}-\mu_x)^2 - 2\sqrt{\frac{n}{n'}}\,\frac{(\bar{x}_{st}-\mu_x)}{(1/\sqrt{n})}\,\frac{(\bar{x}_{n'}-\mu_x)}{(1/\sqrt{n'})} + n'(\bar{x}_{n'}-\mu_x)^2\right\}\right]$$

(under the assumptions considered in the present study)

Letting $(\bar{x}_{st}-\mu_x)/(1/\sqrt{n}) = x_1$ and $(\bar{x}_{n'}-\mu_x)/(1/\sqrt{n'}) = x_2$,

we get $d\bar{x}_{st} = dx_1/\sqrt{n}$ and $d\bar{x}_{n'} = dx_2/\sqrt{n'}$

When $\bar{x}_{n'} = Z_\alpha/\sqrt{n'}$ then $x_2 = \sqrt{n'}(Z_\alpha/\sqrt{n'} - \mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$

and when $\bar{x}_{n'} = -Z_\alpha/\sqrt{n'}$, then $x_2 = \sqrt{n'}(-Z_\alpha/\sqrt{n'} - \mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

Therefore,

$$I_2 = \left[\{\sqrt{n'}/(2\pi\sqrt{n'-n})\}\int_{-\infty}^{\infty}\int_{x_2=A}^{\infty}\{\mu_x + (x_1/\sqrt{n})\}\{\mu_x + (x_2/\sqrt{n'})\}\right.$$
$$Exp\left[-\{n'/2(n'-n)\}\{x_1^2 - 2\sqrt{n/n'}x_1x_2 + x_2^2\}\right]dx_1dx_2$$

$$\left. + \{\sqrt{n'}/(2\pi\sqrt{n'-n})\}\int_{-\infty}^{\infty}\int_{x_2=-\infty}^{B}\{\mu_x + (x_1/\sqrt{n})\}\{\mu_x + (x_2/\sqrt{n'})\}\right.$$
$$\left.Exp\left[-\{n'/2(n'-n)\}\{x_1^2 - 2\sqrt{n/n'}x_1x_2 + x_2^2\}\right]dx_1dx_2\right]$$

$$\text{or } I_2 = I_{21} + I_{22} \qquad \text{.....(3.12)}$$

The integrals I_{21} and I_{22} are evaluated by the use of Fortran 77 and the output values are given in Table 3.4.(Appendix 9 and 10)

3.5 Discussion

For a set of values of n, n', ρ the integrals I_1 and I_2 are computed for different values of the level of significance α. The integrals are evaluated by Simpson's 1/3rd rule through programs written in Fortran 77 in a Linux operating system and the outputs of the program for the values of I_1 and I_2 are given in Tables 3.3 and 3.4. The values of the cumulative function $\Phi(.)$ and the corresponding density function $\varphi(.)$ are also evaluated for the same set of values of n, n', and α. All the set of values are compiled in Excel work sheet and finally the $MSE(t_5)$ is evaluated by substituting all the corresponding set of values of the integrals I_1 , I_2 , $\Phi(.)$ and $\varphi(.)$ for various values of level of significant α and ρ in (3.10) and are given in Tables 3.5 and 3.6.

The $MSE(t_5)$ is plotted for different values of level of significance α (Figure 3.3) and also for different values of correlation coefficient ρ (Figure 3.4) respectively. It is found in general that $MSE(t_5)$ is minimum at $\mu_x = 0$. As μ_x increases, the $MSE(t_5)$ increases to a maximum and then gradually decreases and thereafter becomes constant with further increase in the value of μ_x. The figures clearly show that when the mean of the auxiliary variable is close to the

hypothetical value, then the MSE is minimum. Also as μ_x moves away from the hypothetical value, the MSE increases, but after attaining maximum again gradually reduces and then becomes constant. Figure 3.5 and 3.6 show that the mean square error obtained by numerical methods depict a pattern similar to that obtained by analytical methods for increasing values of μ_x. The differences in the values of MSE between analytical and numerical methods of computation are minimal.

Table 3.1 Behavior of MSE(t_5) computed analytically with respect to μ_x for different values of α and for n = 100, n' = 200 , ρ = 0.8

α \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.01	0.004	0.01	0.015	0.009	0.007	0.007	0.007	0.007	0.007	0.007	0.007
0.05	0.004	0.01	0.01	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
0.25	0.006	0.008	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007

Table 3.2 Behaviour of MSE(t_5) computed analytically with respect to μ_x for different values of ρ and for n = 100, n' = 200 , α = 0.05

ρ \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.7	0.0062	0.0109	0.0161	0.012	0.0095	0.0092	0.0092	0.0092	0.0092	0.0092	0.0092
0.8	0.0051	0.0112	0.018	0.0126	0.0094	0.0089	0.0089	0.0089	0.0089	0.0089	0.0089
0.9	0.0038	0.0115	0.0201	0.0133	0.0093	0.0087	0.0086	0.0086	0.0086	0.0086	0.0086

Table 3.3 Numerically computed values of I_1 with n = 100 , n' = 200 and $\rho = 0.8$ for (a) $\alpha = 0.01$ (b) $\alpha = 0.05$ (c) $\alpha = 0.25$.

(a)

I_{11}	I_{12}	I_1
0.000348	0.000348	0.000696
0.00532	7.93E-06	0.0053279
0.03154	5.92E-08	0.0315401
0.08991	1.40E-10	0.08991
0.16684	1.14E-13	0.16684
0.25819	2.62E-17	0.25819
0.36825	1.85E-21	0.36825
0.49823	4.03E-26	0.49823
0.648203	2.70E-31	0.648203
0.81817	5.55E-37	0.81817
1.0081	3.52E-43	1.0081

(b)

I_{11}	I_{12}	I_1
0.001162	0.001162	0.002324
0.010538	4.75E-05	0.010586
0.042008	6.61E-07	0.042009
0.096468	2.98E-09	0.096468
0.168142	4.20E-12	0.168142
0.258275	1.88E-15	0.258275
0.368259	2.58E-19	0.368259
0.498232	1.08E-23	0.498232
0.648202	1.38E-28	0.648202
0.818168	5.43E-34	0.818168
1.00813	6.45E-30	1.00813

(c)

I_{11}	I_{12}	I_1
0.00299	0.00299	0.00598
0.016174	2.37E-04	0.016411
0.047559	6.96E-06	0.047566
0.098308	6.94E-08	0.098308
0.16829	2.24E-10	0.16829
0.25828	2.29E-13	0.25828
0.368259	7.33E-17	0.368259
0.498233	7.23E-21	0.498233
0.648202	2.18E-25	0.648202
0.818169	2.02E-30	0.818169
1.00813	5.71E-36	1.00813

Table 3.4 Numerically computed values of I_2 with n = 100 , n' = 200 and ρ = 0.8 for (a) α = 0.01. (b) α = 0.05. (c) α = 0.25.

(a)

I_{21}	I_{22}	I_2
0.000317	0.000317	0.000634
0.005044	0.000007	0.005051
0.030573	0	0.030573
0.088609	0	0.088609
0.165882	0	0.165882
0.257301	0	0.257301
0.367307	0	0.367307
0.497205	0	0.497205
0.647086	0	0.647086
0.816951	0	0.816951
1.0068	0	1.0068

(b)

I_{21}	I_{22}	I_2
0.001058	0.001058	0.002116
0.010032	0.000042	0.010074
0.040982	0.000001	0.040983
0.095477	0	0.095477
0.167265	0	0.167265
0.25739	0	0.25739
0.367309	0	0.367309
0.497205	0	0.497205
0.647086	0	0.647086
0.816951	0	0.816951
1.0068	0	1.0068

(c)

I_{21}	I_{22}	I_2
0.002727	0.002727	0.005454
0.015482	0.000204	0.015686
0.046709	0.000017	0.046726
0.097393	0	0.097393
0.167462	0	0.167462
0.257396	0	0.257396
0.367309	0	0.367309
0.497205	0	0.497205
0.647086	0	0.647086
0.816951	0	0.816951
1.0068	0	1.0068

Table 3.5 Behaviour of MSE(t_5) computed numerically with respect to μ_x for different values of α and for $n = 100,\ n' = 200,\ \rho = 0.8$

α \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.01	0	0.01	0.012	0.015	0.011	0.007	0.01	0.01	0.01	0.01	0.01
0.05	0	0.01	0.012	0.013	0.008	0.006	0.01	0.01	0.01	0.01	0.01
0.25	0	0	0.011	0.008	0.006	0.006	0.01	0.01	0.01	0.01	0.01

Table 3.6 Behaviour of MSE(t_5) computed numerically with respect to μ_x for different values of ρ and for $n = 100,\ n' = 200,\ \alpha = 0.05$

ρ \ μ_x	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	**1.0**
0.7	0	0.01	0.011	0.012	0.008	0.007	0.01	0.01	0.01	0.01	0.01
0.8	0	0.01	0.012	0.013	0.008	0.006	0.01	0.01	0.01	0.01	0.01
0.9	0	0.01	0.012	0.014	0.007	0.005	0	0.01	0.01	0.01	0.01

Figure 3.1 Behaviour of MSE(t_5) computed analytically with respect to μ_x for different values of α and for n = 100, n' = 200 , ρ = 0.8

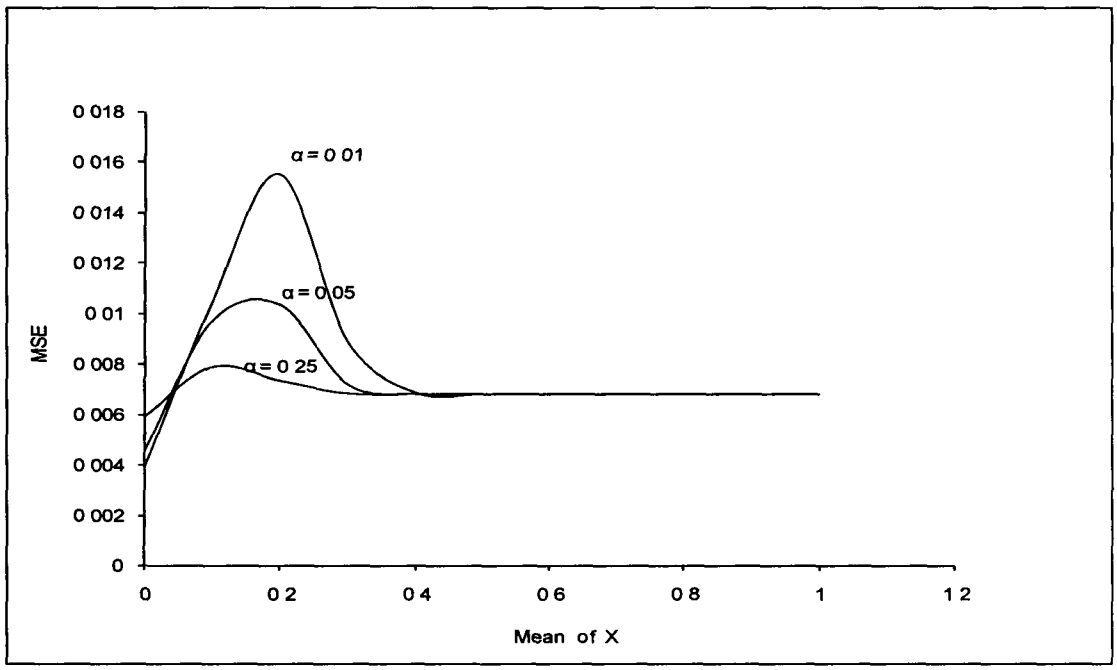


Figure 3.2 Behaviour of MSE(t_5) computed analytically with respect to μ_x for different values of ρ and for n = 100, n' = 200 , α = 0.05

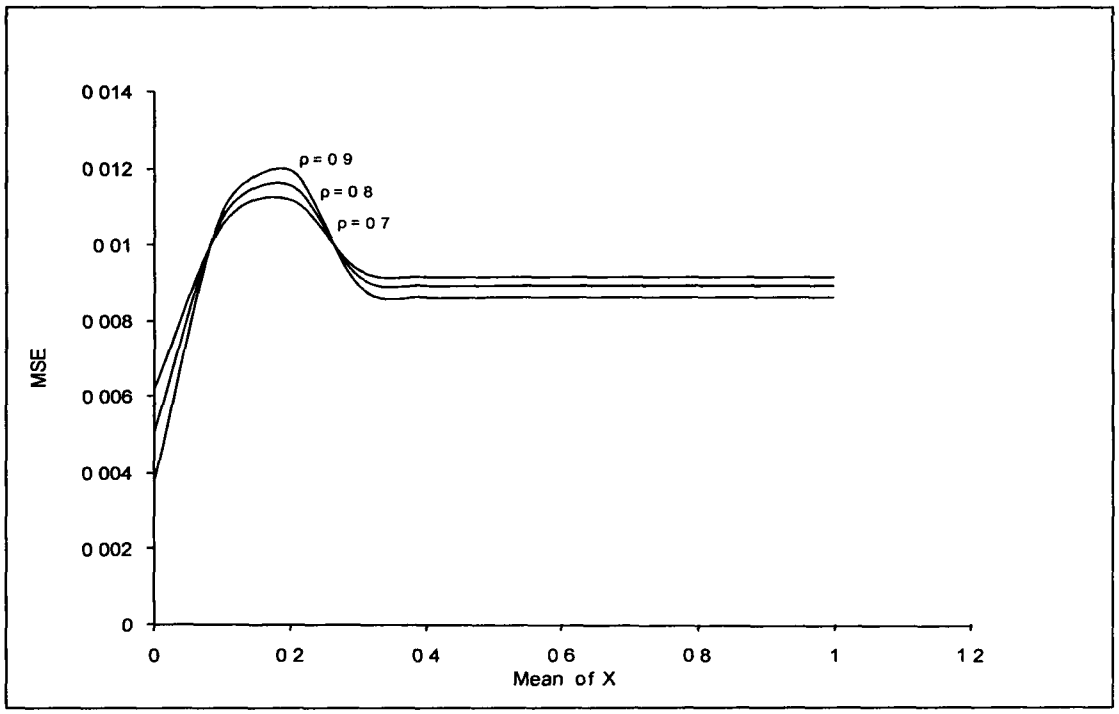


Figure 3.3 Behaviour of the MSE(t_5) computed numerically with respect to μ_x for different values of α and for n = 100 , n' = 200, ρ = 0.8

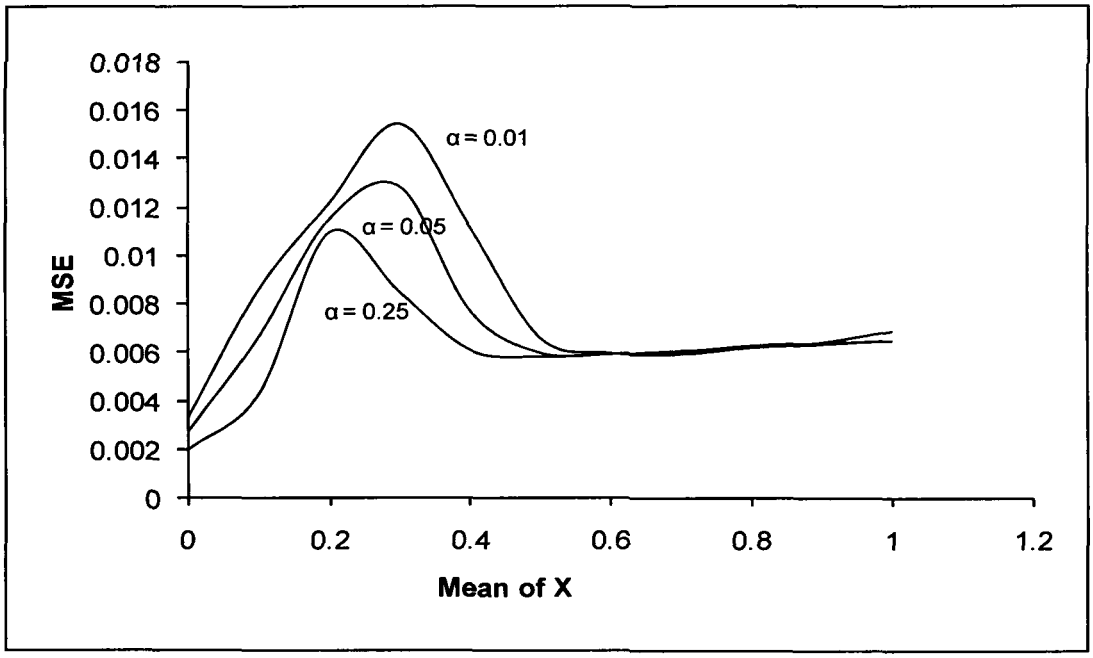


Figure 3.4 Behaviour of the MSE(t_5) computed numerically with respect to μ_x for different values of ρ and for n = 100 , n' = 200, α = 0.05

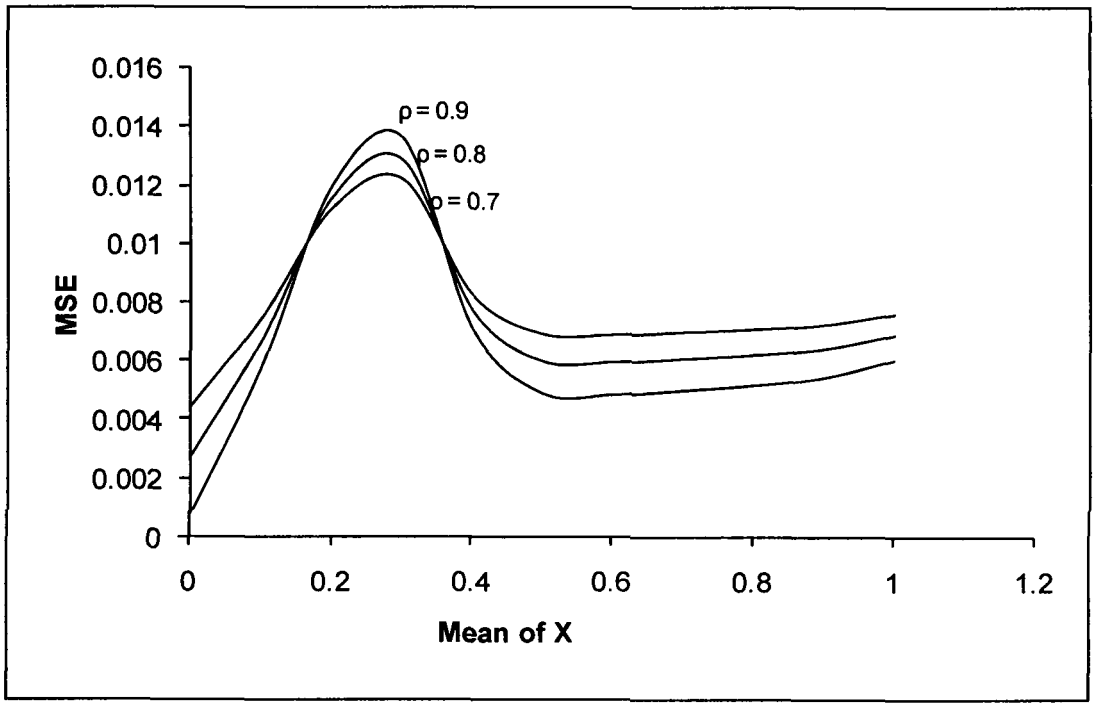


Figure 3.5 Comparative behaviour of the MSE(t_5) with respect to μ_x for different values of α and for $\rho = 0.8$, $n = 100$, $n' = 200$

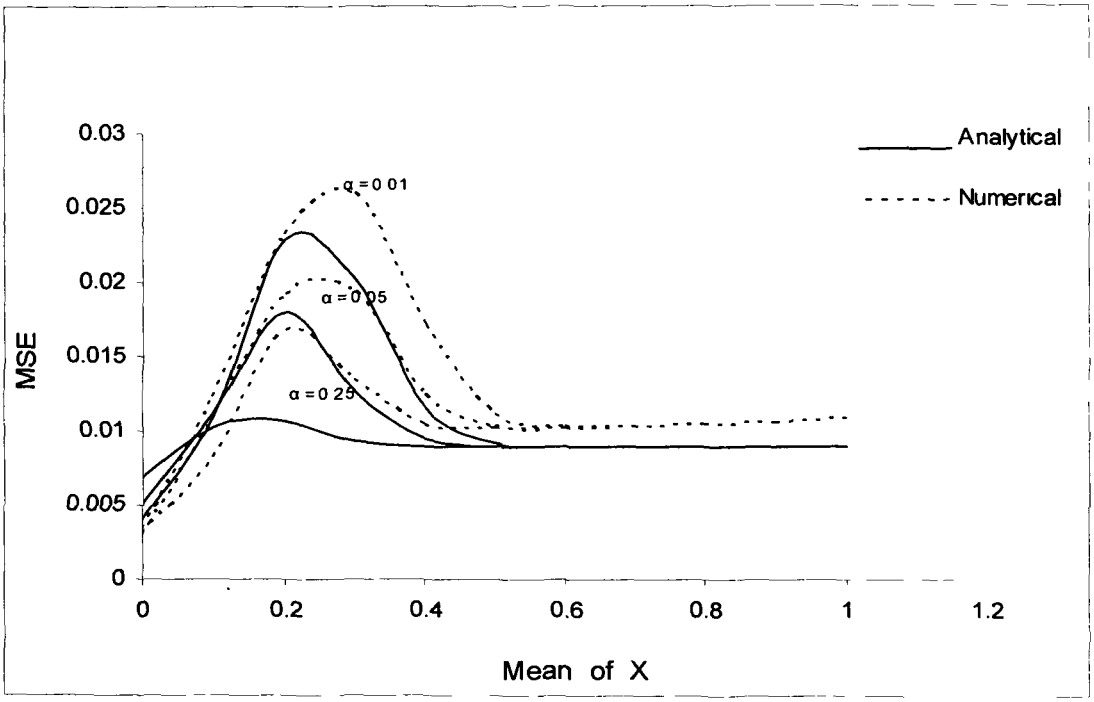
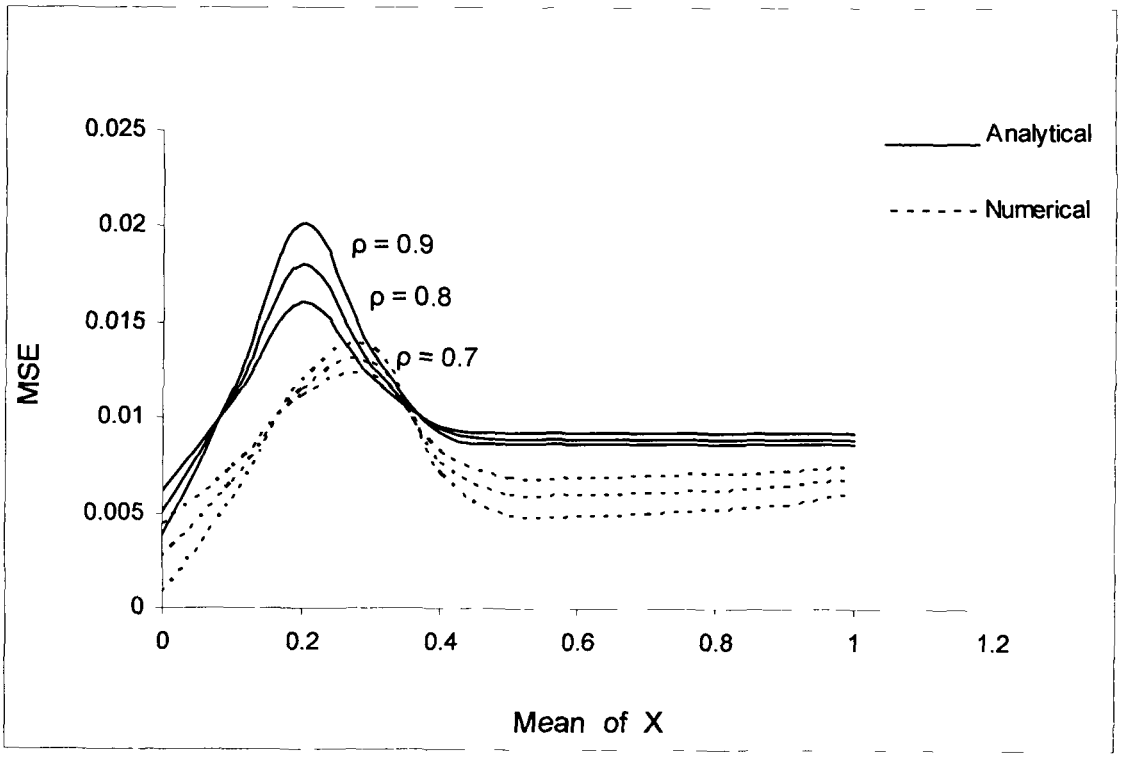


Figure 3.6 Comparative behaviour of the MSE(t_5) with respect to μ_x for different values of ρ and for $\alpha = 0.05$, $n = 100$, $n' = 200$



Chapter 4

Relative efficiency and optimum allocation of the CRPTE in double sampling with partial information on the auxiliary variable.

4.1 Introduction

The study of the mean square error of an estimator will not be completed unless it is compared with other estimators. Without the use of real data, precision between estimators can be compared by analytical derivation and conclusion can be drawn from an inequality showing mean square expression of the two estimators on both sides. Another useful criteria for the comparison of the precision of any estimator is the relative efficiency defined as the ratio of the variance or mean square error of one estimator to that of the mean square error of the proposed estimator. If the relative efficiency is greater than 1, it can be concluded that the proposed estimator is more efficient in comparison to the other estimator.

In the present chapter, the mean square error of the proposed combined regression preliminary test estimator(CRPTE) in double sampling is compared with other estimator and conclusion is drawn through the relative efficiency. The relative efficiency of the proposed estimator is also simulated for different sample sizes, correlation and level of significance and the values are plotted for depicting its behaviour corresponding to various values of the mean of the auxiliary variable.

4.2 Relative efficiency of the CRPTE

The Combined regression estimator under double sampling is given by

$$t_4 = \bar{y}_{st} + b_{yx}(\bar{x}_{n'} - \bar{x}_{st})$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

The $MSE(t_4)$ can be derived as follows,

Let

$$\bar{y}_{st} = \mu_y + \varepsilon_1, \quad \bar{x}_{st} = \mu_x + \varepsilon_2, \quad \bar{x}_{n'} = \mu_x + \varepsilon_3,$$

$$b_{yx} = s_{yx} / s_x^2, \quad s_{yx} = S_{yx} + \varepsilon_4, \quad s_x^2 = S_x^2 + \varepsilon_5$$

where ε_i's are errors such that $E(\varepsilon_i) = 0$ for i = 1,2,3,4 and 5 and

S_x^2, S_y^2 and S_{yx} are population variances and covariance respectively.

Now,

$$MSE\ (t_4) = E\{t_4 - E(t_4)\}^2$$

$$= E\{\bar{y}_{st} + b_{yx}(\bar{x}_{n'} - \bar{x}_{st}) - \mu_y\}^2$$

$$= E\{(\mu_y + \varepsilon_1) + (s_{yx} / s_x^2)(\varepsilon_3 - \varepsilon_2) - \mu_y\}^2$$

$$= E[\varepsilon_1 + \{(S_{yx} + \varepsilon_4)/(S_x^2 + \varepsilon_5)\}(\varepsilon_3 - \varepsilon_2)]^2$$

$$= E[\varepsilon_1 + (S_{yx} / S_x^2)\{1 + (\varepsilon_4 / S_{yx})\}\{1 + (\varepsilon_5 / S_x^2)\}^{-1}(\varepsilon_3 - \varepsilon_2)]^2$$

Assuming

$\left|\varepsilon_5 / S_x^2\right| < 1$ we get,

$$\text{MSE}(t_4) = E[\varepsilon_1 + (S_{yx} / S_x^2)\{1 + (\varepsilon_4 / S_{yx})\}\{1 - (\varepsilon_5 / S_x^2) + ...\}\ (\varepsilon_3 - \varepsilon_2)]^2$$

$$= E[\varepsilon_1 + (S_{yx} / S_x^2)\{1 + (\varepsilon_4 / S_{yx})\}\{\varepsilon_3 - (\varepsilon_3\varepsilon_5 / S_x^2) - \varepsilon_2 + (\varepsilon_2\varepsilon_5 / S_x^2)\}]^2$$

$$= E[\varepsilon_1 + (S_{yx} / S_x^2)\{\varepsilon_3 - (\varepsilon_3\varepsilon_5 / S_x^2) - \varepsilon_2 + (\varepsilon_2\varepsilon_5 / S_x^2) + (\varepsilon_4\varepsilon_3 / S_{yx}) - (\varepsilon_2\varepsilon_4 / S_{yx})]^2$$

which upto to second order of approximation, gives

$$MSE\ (t_4) = E[\varepsilon_1 + (S_{yx} / S_x^2)(\varepsilon_3 - \varepsilon_2)]^2$$

$$= E[\varepsilon_1^2 + (S_{yx}^2 / S_x^4)(\varepsilon_3^2 - 2\varepsilon_3\varepsilon_2 + \varepsilon_2^2) + 2(S_{yx} / S_x^2)(\varepsilon_1\varepsilon_3 - \varepsilon_1\varepsilon_2)]$$

$$= Var\ (\bar{y}_{st}) + (S_{yx}^2 / S_x^4)Var\ (\bar{x}_{st}) + (S_{yx}^2 / S_x^4)Var\ (\bar{x}_{n'})$$
$$- 2(S_{yx}^2 / S_x^4)Cov\ (\bar{x}_{n'}, \bar{x}_{st}) + 2(S_{yx} / S_x^2)\{Cov\ (\bar{x}_{n'}, \bar{y}_{st}) - Cov\ (\bar{y}_{st}, \bar{x}_{st})\}$$

Now, the variance of $\bar{x}_{st}$ and $\bar{y}_{st}$ is given by

$$Var(\bar{y}_{st}) = \sum (W_h^2 S_{y_h}^2 / n_h)(1 - f_h) \quad Var(\bar{x}_{st}) = \sum (W_h^2 S_{x_h}^2 / n_h)(1 - f_h)$$
$$= \sum W_h^2 S_{y_h}^2 / n_h \qquad = \sum (W_h^2 S_{x_h}^2 / n_h)$$

(Cochran, 1977)

(Under the consideration that the sampling fraction n_h / N_h are negligible.)

$S_{y_h}^2$ and $S_{x_h}^2$ are variance of Y and X in the h[th] stratum.

Also the variance of $\bar{x}_{n'}$ is given by

$$Var(\bar{x}_{n'}) = \{(1/n') - (1/N)\}S_x^2$$
$$= S_x^2 / n'$$

(Cochran, 1977)

(for large population size N, 1/N is negligible)

The pair (X, Y) is considered to be a bivariate random variable with mean (μ_x, μ_y) and covariance matrix $\sum_{(X,Y)}$ in which the variances are denoted by S_x^2 and S_y^2 and the correlation coefficient by ρ. The regression estimator depends on weather the covariance matrix is known or not. If $\sum_{(X,Y)}$ is known, we may let $S_x^2 = S_y^2 = 1$ without the loss of generality. The strata population (X_h, Y_h) can also be considered as a bivariate random variable for every h, with mean (μ_{x_h}, μ_{y_h}). If the covariance matrix $\sum_{(X_h, Y_h)}$ of the pair (X_h, Y_h) is known, we can let $S_{x_h}^2 = S_{y_h}^2 = 1$(WLOG).

When the samples are selected with proportional allocation then the strata weights are given by

$$W_h = (N_h / N) = (n_h / n) \qquad \text{(Cochran, 1977)}$$

Thus $\sum_h W_h^2 / n_h = \sum_h W_h^2 / nW_h = (1/n)\sum_h W_h = (1/n)$ (as $\sum_h W_h = 1$)

Hence the covariance matrix of $(\bar{x}_{n'}, \bar{x}_{st}, \bar{y}_{st})$ reduces to

$$\sum = \begin{pmatrix} Var(\bar{x}_{n'}) & Cov(\bar{x}_{n'}, \bar{x}_{st}) & Cov(\bar{x}_{n'}, \bar{y}_{st}) \\ Cov(\bar{x}_{st}, \bar{x}_{n'}) & Var(\bar{x}_{st}) & Cov(\bar{x}_{st}, \bar{y}_{st}) \\ Cov(\bar{y}_{st}, \bar{x}_{n'}) & Cov(\bar{y}_{st}, \bar{x}_{st}) & Var(\bar{y}_{st}) \end{pmatrix}$$

i.e, $$\sum = \begin{pmatrix} 1/n' & 1/n' & \rho/n' \\ 1/n' & 1/n & \rho/n \\ \rho/n' & \rho/n & 1/n \end{pmatrix}$$

Hence,

$$MSE(t_4) = (1/n) + (\rho^2/n) + (\rho^2/n') - 2(\rho^2/n') + 2\rho\{(\rho/n') - (\rho/n)\}$$

$$= (1/n)(1 - \rho^2) + (1/n')\rho^2 \qquad \text{.....(4.1)}$$

Now, MSE(t_5) is given by (3.9) as follows

$$MSE(t_5) = \{(1-\rho^2)/n + \rho^2/n'\} + (\rho^2/n')\{A\varphi(A) - B\varphi(B)\} - \rho^2(1/n' - \mu_x^{\,2})\{\Phi(A) - \Phi(B)\}$$

or $MSE(t_5) = g_1 + h_1$, where

$$g_1 = \{(1-\rho^2)/n + \rho^2/n'\}$$

$$h_1 = (\rho^2/n')\{A\varphi(A) - B\varphi(B)\} - \rho^2(1/n' - \mu_x^{\,2})\{\Phi(A) - \Phi(B)\}$$

Therefore, the relative efficiency of t_5 to t_4 is given by

$$e_1 = [MSE\ (t_4)\ ] / [MSE\ (t_5)\ ] = g_1/(g_1 + h_1)$$

The behavior of relative efficiency can be observed for different values of μ_x. As e_1 is symmetric about $\mu_x = 0$, we need to consider e_1 only for $\mu_x \geq 0$. The above behavior can be studied for selected values of the level of significance α and correlation coefficient ρ respectively and by fixing some hypothetical values for n and n' (Table 4.1, Table 4.2 ; Fig 4.1, 4.2).

4.3 Discussion

In order to get an idea about the behavior of the relative efficiency function of t_5 with respect to t_4 for different values of μ_x, e_1 can be computed for a set of values of n, n', α and ρ. Table 4.1, 4.2 and Figure 4.1, 4.2 show that in general that e_1 has a maximum at $\mu_x = 0$. As μ_x increases, e_1 decreases to a minimum and then again increases to unity and remains constant thereafter. It is found that e_1 is very close to 1 at $\mu_x = 1$.

The Figures clearly show that when the mean of the auxiliary variable is close to the hypothetical value, then relative efficiency is maximum. Also as μ_x moves away from the hypothetical value the relative efficiency decreases, but after attaining minimum again increases to unity and remains constant thereafter.

Relative efficiency is high when μ_x is close to the hypothetical value, i.e, when the hypothesis considered in the study is likely to be accepted and as a result the mean of the partial information of X is used in the proposed estimator. Thus the proposed estimator reduces to the usual combined regression estimator which as we know has efficiency higher than the combined regression

estimator under double sampling. On the other hand when μ_x is far away from the hypothetical value of the mean of the auxiliary variable, then the hypothesis is likely to be rejected. In this case the mean of X from the preliminary sample is used for the proposed estimator, so that the estimator t_5 reduces to the combined regression estimator under double sampling which is same as t_4 and hence e_1 is equals to unity for values of μ_x far away form the hypothetical value of X. For the intermediate values of μ_x, i.e, under situations when one is not certain as whether to accept or reject the hypothesis, then the CRPTE has a lesser efficiency as compared to the combined regression estimator under double sampling. This establishes the utility of the present study of preliminary test estimator using reliable partial information on the auxiliary variable.

4.4 Optimum allocation

In planning of a sample survey, a stage is always reached at which a decision must be made about the size of the sample. This decision is important. Too large a sample involves utilization of more time and resources and too small a sample diminishes the precision of the results. Thus an optimum size of the sample is required so as to balance precision and cost involved in the survey. The optimum allocation of sample sizes are attained either by minimizing precision against a given cost or minimizing cost against given precision.

In the sampling scheme of the proposed estimator, the samples are extracted by double sampling in which the first sample is a stratified simple

random sample of size n in which the pair (x_{hi}, y_{hi}) values are measured from n_h units drawn from each stratum and consequently estimating the pair $(\bar{x}_{st}, \bar{y}_{st})$, with $n = \sum_h n_h$. The second sample is a larger simple random sample of size $n'(=n+m)$ is obtained by supplementing m more independent observations on X where only x_i is measured and evaluates $\bar{x}_{n'}$ which is utilized to estimate μ_x.

In order to obtain optimum allocation of sample sizes for the suggested estimator, let us consider simple linear cost function C given by

$$C = c'n' + cn \qquad \text{.....(4.2)}$$

where c is the cost per unit of observing the variable Y and c' is the cost per unit of observing the variable X, assuming that the cost per unit is the same for all strata.

The CRPTE in double sampling constructed in the present study is given by

$$t_5 = \begin{cases} (\bar{y}_{st} - \rho\bar{x}_{st}) & if \quad |\bar{x}_{n'}| \le Z_\alpha / \sqrt{n'} \\ (\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) & if \quad |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'} \end{cases}$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

In the present study the values of the sample sizes n and n' will be obtained by minimizing the $MSE(t_5)$ for a specific cost C*. In order to evaluate the optimum MSE for the estimator t_5 under the above mentioned constraint, we proceed as follows;

The mean square error of the proposed estimator is given by (3.9) as

$$MSE(t_5) = \{(1-\rho^2)/n + \rho^2/n'\} + (\rho^2/n')\{A\varphi(A) - B\varphi(B)\} - \rho^2(1/n' - \mu_x{}^2)\{\Phi(A) - \Phi(B)\}$$

In general the values of μ_x are unknown, the experimenter has partial information about it. When $\mu_x = 0$, the mean square error of t_5 is least and the relative efficiency is largest. Thus it would be reasonable to let $\mu_x = 0$ in $MSE(t_5)$ and obtain the values of n and n' under the optimum situation.

When $\mu_x = 0$, $A = Z_\alpha$ and $B = -Z_\alpha$, which further implies that

$$\Phi(A) = 1 - (\alpha/2) \quad \text{and} \quad \Phi(B) = \alpha/2$$

Hence,

$$\begin{aligned} MSE(t_5) &= \{(1-\rho^2)/n + \rho^2/n'\} \\ &\quad + \{\rho^2/n'(Z_\alpha\varphi(Z_\alpha) - (-Z_\alpha)\varphi(-Z_\alpha)\} - (\rho^2/n')\{1 - \alpha/2 - \alpha/2\} \\ &= (1-\rho^2)/n + \rho^2\{\alpha + 2Z_\alpha\varphi(Z_\alpha)\}/n' \\ &= (K/n) + (K'/n') \end{aligned} \qquad \text{.....(4.3)}$$

where $K = 1 - \rho^2$ $\qquad K' = \rho^2\{\alpha + 2Z_\alpha\varphi(Z_\alpha)\}$

We minimize the MSE(t_5) subject to a specific cost constraint C*, given by

$$C^* = c'n' + cn \qquad \text{.....(4.4)}$$

Lagrange's multipliers method is used to minimize of the MSE(t_5) in eqn(4.3) subject to the cost constraint C* in eqn(4.4). The Langrange's equation can be written as follows

$$L(n',n,\lambda)=\left[K/n+K'/n'\right]-\lambda(C^*-c'n'-cn) \quad \text{.....(4.5)}$$

where λ is a constant or the Langrange multiplier.

The necessary conditions for a minimum of $MSE\ (t_s)=(K/n)+(K'/n')$ subject to $C^*-c'n'-cn=0$ are given by

$$\frac{\partial L}{\partial n'}=0, \qquad \frac{\partial \mathrm{L}}{\partial n}=0 \quad \text{and} \quad \frac{\partial \mathrm{L}}{\partial \lambda}=0$$

Hence,

$$-(K'/n'^2)+\lambda c'=0$$

$$-(K/n^2)+\lambda c=0$$

and

$$C^*-c'n'-cn=0$$

Thus,

$$(K'/n'^2)=c'\lambda \quad \text{and} \quad (K/n^2)=c\lambda$$

$$\Rightarrow (n'^2c'/K')=(n^2c/K)=1/\lambda$$

$$\Rightarrow (n'^2/K'c)=(n^2/Kc')=(1/\lambda cc')=L\,(say\ )$$

$$\Rightarrow n'=\sqrt{L}\sqrt{K'c} \quad \text{and} \quad n=\sqrt{L}\sqrt{Kc'} \quad \text{.....(4.6)}$$

Substituting the value of n and n' in the cost function C*, we get

$$C^*=c'n'+cn$$

$$C^*=(c'\sqrt{K'c}+c\sqrt{Kc'})\sqrt{L}$$

$$\sqrt{L}=C^*/(c'\sqrt{K'c}+c\sqrt{Kc'})$$

$$\sqrt{L}=C^*/\sqrt{cc'}\{\sqrt{(Kc}+\sqrt{K'c'}\} \quad \text{.....(4.7)}$$

Thus the optimum value of $MSE(t_5)$ is obtained by substituting the value of n and n' from eqn (4.6) and eqn (4.7) in eqn(4.3)

$$i..e,\quad M_{opt}(t_5) = \left[K/\sqrt{Kc'} + K'/\sqrt{K'c}\right](1/\sqrt{L})$$

$$= \left[\sqrt{K/c'} + \sqrt{K'/c}\right](1\sqrt{L})$$

$$= \frac{\left[\sqrt{(Kc} + \sqrt{K'c'}\right]}{\sqrt{cc'}} \frac{\sqrt{cc'}\left[\sqrt{(Kc} + \sqrt{K'c'}\right]}{C^*}$$

Thus

$$M_{opt}(t_5) = \left[\sqrt{Kc} + \sqrt{c'K'}\right]^2 / C^* \qquad \text{.....(4.8)}$$

which gives the required optimum mean square error of the proposed estimator.

4.5 Comparison of the CRPTE with combined regression estimator under optimum condition

The Combined regression estimator under double sampling is given by

$$t_4 = \bar{y}_{st} + b_{yx}(\bar{x}_{n'} - \bar{x}_{st})$$

The Mean square error of t_4 is given by

$MSE(t_4) = (1/n)(1-\rho^2) + (1/n')\rho^2 = (K/n) + (K''/n')$ (given in eqn 4.1)

where $K = (1-\rho^2)$ and $K'' = \rho^2$

and the cost function given by eqn (4.2)

Again by proceeding as in Section 4.4 by applying Lagrange's multipliers method , we minimize $MSE(t_4)$ for a specific cost C* and we get

$$M_{opt}(t_4) = (\sqrt{Kc} + \sqrt{K''c'})^2 / C^* \qquad \text{.....(4.9)}$$

Analytically it can be observed that $\alpha + 2Z_\alpha \varphi(Z_\alpha)$ is a decreasing function of Z_α with a maximum equal to unity at Z_α =0. Therefore we can conclude that $M_{opt}(t_5) \leq M_{opt}(t_4)$ with equality holding for Z_α = 0 in which case the two estimators concide.

4.6 Discussion

Thus we have proved that under optimum conditions, mean square error of the CRPTE in double sampling with an auxiliary variable is smaller than the mean square error of combined regression estimator under double sampling. Therefore under the stated assumptions, the proposed estimator is more efficient than the combined regression estimator under double sampling.

Table 4.1 Behaviour of relative efficiency of t_5 to t_4 with respect to μ_x for different values of α and for $n' = 200$, $n = 100$, $\rho = 0.8$

α \ μ_x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.01	1.76	0.66	0.44	0.76	0.99	1	1	1	1	1	1
0.05	1.51	0.7	0.66	0.95	1	1	1	1	1	1	1
0.25	1.15	0.86	0.93	1	1	1	1	1	1	1	1

Table 4.2 Behaviour of relative efficiency of t_5 to t_4 with respect to μ_x for different values of ρ and for $n' = 200$, $n = 100$, $\alpha = 0.05$

ρ \ μ_x	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
0.7	1.31	0.77	0.73	0.96	1	1	1	1	1	1	1
0.8	1.51	0.7	0.66	0.95	1	1	1	1	1	1	1
0.9	1.96	0.62	0.57	0.92	1	1	1	1	1	1	1

Figure 4.1 Behaviour of relative efficiency of t_5 to t_4 with respect to μ_x for different values of α and for $n' = 200$, $n = 100$, $\rho = 0.8$

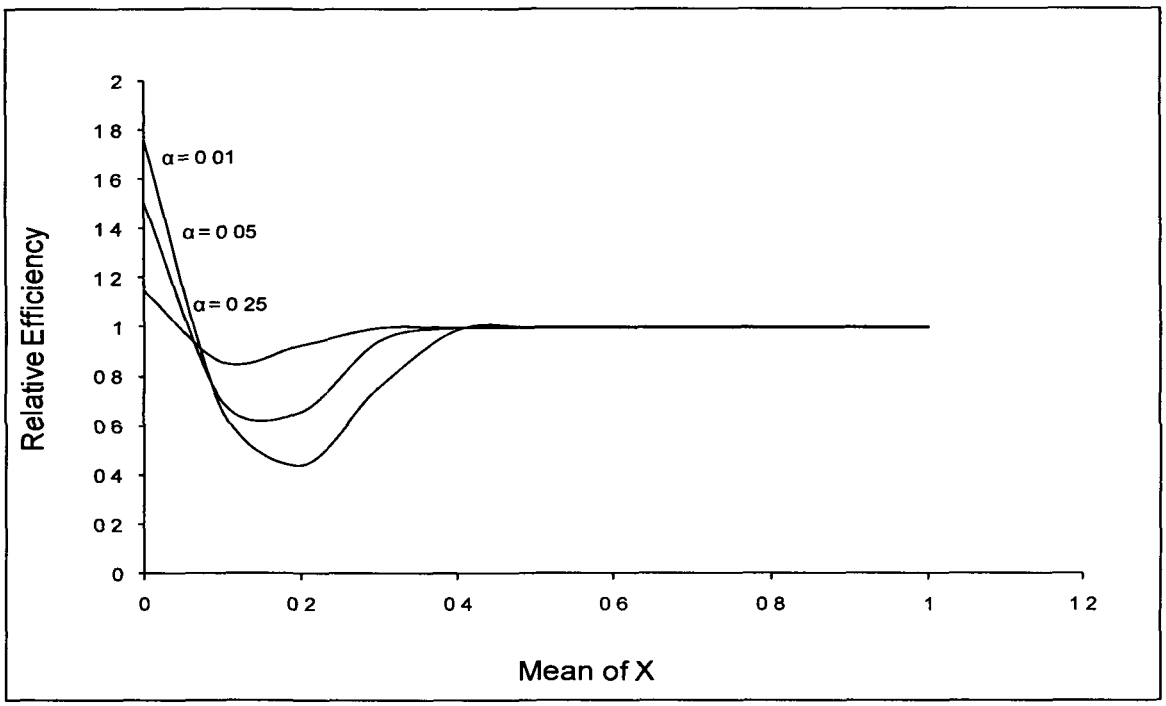
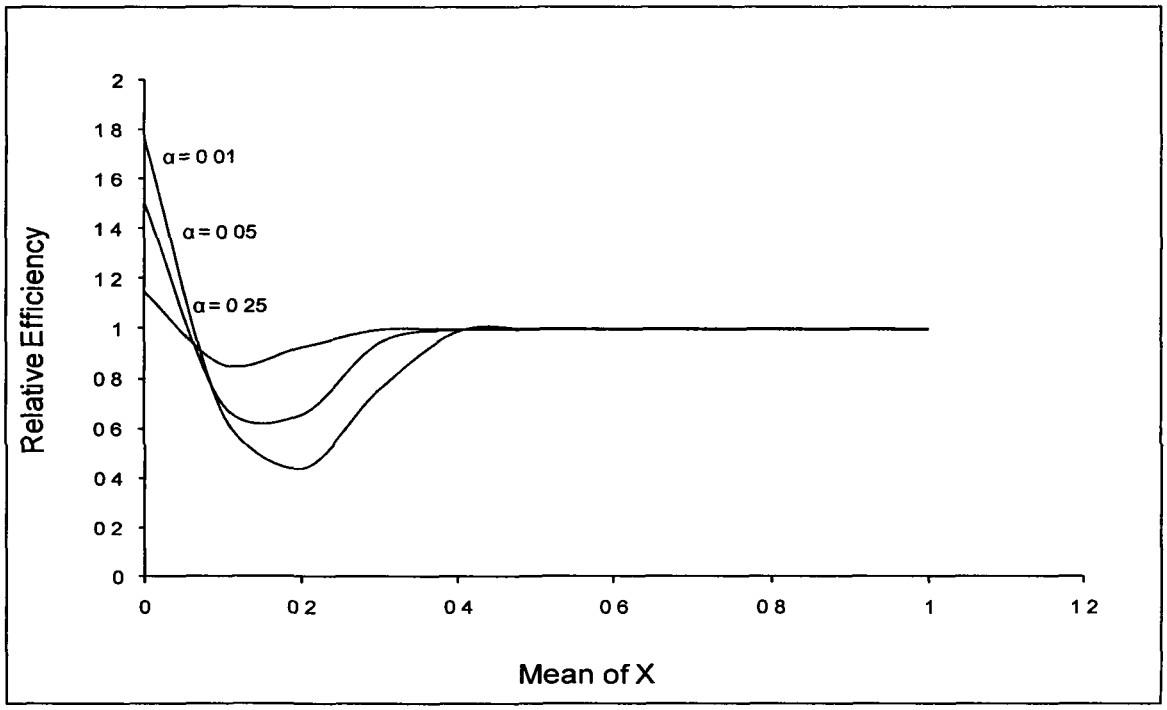


Figure 4.2 Behaviour of relative efficiency of t_5 to t_4 with respect to μ_x for different values of ρ and for $n' = 200$, $n = 100$, $\alpha = 0.05$



Chapter 5

Empirical studies based on the CRPTE in double sampling with partial information on the auxiliary variable.

5.1 Introduction

In the previous chapters, a new estimator is proposed using partial information on auxiliary variable in stratified random sampling. The Bias function of the proposed estimator was derived both analytically and numerically. In order to judge the precision of the estimator, the mean square error function is derived. The behavior of the bias and MSE is studied with respect to different numerical values of the mean of the auxiliary variable and other corresponding parameters. Optimum allocation of sample sizes is also studied and the MSE of the proposed estimator is compared under the optimum condition.

In the present chapter, efforts are made to study the performance of the proposed estimator through empirical data. Here, two sets of data were used : one a real life data and another a simulated data, generated with the help of statistical software STATA 8.0(2003).

5.2 An empirical study through real life data

Real life data were obtained from Reproductive Child Health report (2000). The data provides district wise demographic indicators for the Empowered Action Groups (EAG) States. From the data two distinct characteristics of the population were identified, namely complete child immunization and female literacy rate. A correlation coefficient was computed between the two characteristics and found that the characters are positively correlated, with correlation coefficient as high as 0.637. For the empirical analysis of the combined regression preliminary test estimator(CRPTE) in

double sampling using the real data set, it is assumed that a more literate mother is more liable to get her children complete immunization. Hence the percent of complete child immunization is taken to be the dependent variable (Y) and female literacy rate is considered as an auxiliary variable (X).

In the present analysis, the primary sample unit is set at district level and altogether there is a total of N =191 districts of the entire EAG states. After eliminating those districts in which data were not available partly or wholly the total size of the population reduces to N = 158. Each state is governed by a distinct political and cultural system, which can have far reaching consequences on the economic, social status of the population within each state. This in turn can have an impact on the demographic characteristics of each state differently. So the data, on the selected variables can be homogeneous within each state and heterogeneous between states. Hence for the purpose of the present study of the EAG states, each state is considered as a stratum.

In the present study, the estimator was constructed under the assumption that both the study variable and the auxiliary variable follow normal distribution. The real life data considered to study are given by complete child immunization(Y) and female literacy rate(X) in terms of percentages being obtained as a result of counting and categorizing the attributes into two outcomes viz. literate or not literate and immunized of not immunized. Thus X and Y can be considered to behave like the binomial variables. As it is known that arcsine transformation convert a binomial random variable into one that is nearly normal, this transformation is used to convert the present data into the

one that is nearly normal. The Arcsine transformed data on Y and X are tested for normality using the Shapiro-Wilk test statistic carried on in SYSTAT 12.0 on both sets of variable Y and X and the results are shown in Table 5.1. The Shapiro-Wilk test is a standard test for normality used when the sample size is between 3 and 5000. The p-value given by this test is an indication of how good the fit is, the smaller the p-value is, the worse is the fit. Generally, p-values of the order of 0.05 or 0.01 are considered small enough to declare the fit poor. Table 5.1 reveals that both sets of data on X and Y follow normal distribution.

As discussed above, in order to utilize the data to see the performance of the proposed estimator under study, the seven EAG states are considered as seven strata and the corresponding strata sizes are shown in Table 5.2. Proportional allocation of sample sizes is adopted with n = 30. The formula for proportional allocation is given by $W_h = (N_h / N) = (n_h / n)$. The selection of samples within each stratum is done by simple random sampling. Also, it is considered that the pair (X, Y) follows a bivariate normal distribution with mean (μ_x, μ_y) and known covariance matrix $\sum$ in which the variances are denoted by σ_x^2 and σ_y^2 and the correlation coefficient by ρ. The value of $\sigma_x = 0.1347$ and $\sigma_y = 0.2237$ are computed from the population consisting of seven EAG states.

To use the regression estimator it is usually assumed that the population mean μ_x of the auxiliary variable is known. When μ_x is unknown, alternative information about the mean of the auxiliary variable is also obtained by the use

of double sampling procedures. A stratified simple random sample of size n = 30 in which the pair (x_{hi}, y_{hi}) values are measured from n_h units drawn from each stratum and consequently leading to the estimation of the pair $(\bar{x}_{st}, \bar{y}_{st})$, with $n = \sum_h n_h$ as sample size. An additional sample of size m = 40, is selected by simple random sampling in which only x_i values are measured. Here $n' = n + m$ and the preliminary sample is a simple random sample of size $n' = 70$, which evaluates $\bar{x}_{n'}$.

Further suppose that we have partial information about μ_x. In the present case the partial information on X can be obtained from Census of India (1991) where the literacy rate is computed for the EAG states and given as 26.4% which on using the arcsine transformation gives μ_0 = 0.5396. In the present study the preliminary sample is utilized to test the hypothesis

$$H_0 : \mu_x = \mu_0 \quad \text{against} \quad H_1 : \mu_x \neq \mu_0$$

If the hypothesis H_0 is accepted then μ_0 obtained from the partial information will be used in the proposed estimator t_5; if H_0 is rejected, the sample mean $\bar{x}_{n'}$ based on the preliminary sample is used.

The mean of the auxiliary variable is considered to be partially known as μ_0. We can assume that μ_0 = 0 without the loss of generality. The covariance matrix Σ is also considered to be known, hence without loss of generality it can be assumed that $\sigma_x^2 = 1$ and $\sigma_y^2 = 1$. In order to compare μ_0 with the sample mean $\bar{x}_{n'}$ in the testing of hypothesis, the sample values of the auxiliary variable X is transformed both in origin and scale maintaining the characteristics

of X as $(x_t - \mu_0)/\sigma_X$, and the variable Y is transform to y_t/σ_y .The mean of X obtained from preliminary sample in double sampling is $\bar{x}_{n'} = 1.438$.

If the hypothesis H_0 is accepted then $\mu_0 = 0.0$ obtained from the partial information will be used in the regression estimator t_5 . If H_0 is rejected, the sample mean $\bar{x}_{n'} = 1.438$ based on the preliminary sample is used.

The combined regression preliminary test estimator in double sampling is given by

$$t_5 = \begin{cases} (\bar{y}_{st} - \rho\bar{x}_{st}) & if \quad |\bar{x}_{n'}| \le Z_\alpha/\sqrt{n'} \\ (\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) & if \quad |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'} \end{cases}$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$ and ρ being the correlation coefficient of X and Y. The strata means in terms of the transformed values in X and Y is shown in Table 5.3.

Table 5.3 shows that $\bar{x}_{st}$=1.497 , $\bar{y}_{st}$=3.347 and $\bar{x}_{st}$=1.438. In terms of the original unit $\bar{x}_{st} = 45.59\%$, $\bar{y}_{st} = 46.31\ \%$ and $\bar{x}_{n'} = 44.8\ \%$.

The criteria for the acceptance and rejection of the hypothesis H_0 is based on the mean of the auxiliary variable obtained from the preliminary sample in double sampling. Table 5.4 shows that $|\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}$ for different values of the level of significant α. Hence in the present case the hypothesis is rejected and consequently $\bar{x}_{n'}$ is utilized in the proposed estimator.

Thus,

$$t_5 = \{\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st}) \quad if \quad |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'}\}$$
$$= 3.30$$

and in terms of the original units

$$t_5 = 45.29\ \%$$

Thus the estimate of the percentage of complete child immunization for the EAG states by the use of the combined regression preliminary test estimator in double sampling 45.29 %

In order to judge the precision of the estimator, the MSE(t_5) is computed for various values of the level of significance. The mean square error of the CRPTE in double sampling is given by

$$MSE(t_5) = g_1 + h_1$$

where

$$g_1 = (1-\rho^2)/n + \rho^2/n' \quad \text{and} \quad h_1 = \{\rho^2/n'(A\varphi(A) - B\varphi(B)) - \rho^2(1/n' - \mu_x^{\ 2})(\Phi(A) - \Phi(B))\}$$

5.3 Dissussion

The values of $MSE(t_5)$ are shown in Table 5.5 for different values of α. When a reliable partial information on the mean of the auxiliary variable is not available as in the present case , the hypothesis is rejected, as a result of which $\bar{x}_{n'}$ is utilized in the estimation of μ_x . Hence the proposed estimator t_5 reduces to the usual combined regression estimator under double sampling i.e. t_4 . In the computation of the $MSE(t_5)$ the contribution of h_1 is highly negligible, hence the values of the mean square error is contributed mostly by g_1. That is why $MSE(t_5)$ does not change with change in the values of α. Also empirically in this case the $MSE(t_5)$ is same as MSE of the the usual combined regression

estimator under double sampling. That is to say that in this case the two estimators t_5 and t_4 are equally efficient.

5.4 An empirical study through simulated data

The significance of simulated data is the fact that the data can be manipulated to suit the conditions of a model under study. Unlike real life data where the experimenter has no control over the parameters, simulated data provides the necessary outputs in which one has complete control over the parameters. Simulated data provides alternative means of replicating realistic situations under controlled environment.

In the present study an attempt is made to compute t_5 by using simulated data set and also evaluate the mean square error of the proposed estimator to compare with other estimator through relative efficiency.

In order to obtain simulated data following the assumptions of the present study, a statistical software STATA 8.0(2003) is utilized. In the process of generation of bivariate normally distributed random variables, input parameters of the means, variances and correlation coefficient $(\mu_{x_h}, \mu_{y_h}, \sigma_{x_h}, \sigma_{y_h}, \rho)$ are given for every stratum h (h= 1,2,3,4) with strata population sizes taken as N_1=35, N_2=40, N_3=50 and N_4=45. The stratification was done according to the mean of the value of the study variable Y and the stratum correlation coefficients are assumed to be constant and equal to the population correlation coefficient.

Next, the strata population are combined into a single population for which the total size is found to be N = 170. The correlation coefficient of the combined population is computed and found to be ρ = 0.868.The data on Y and X are tested for normality using the Shapiro Wilk test statistic carried on in SYSTAT 12.0 and the results is shown in table 5.6. Result reveal that both sets of data on X and Y follow normal distribution.

Proportional allocation of sample sizes is adopted with n = 35(Table 5.7). The selection of samples within each stratum is done by simple random sampling. Also, it is considered that the pair (X, Y) follows a bivariate normal distribution with mean (μ_x, μ_y) and known covariance matrix $\sum$ in which the variances are denoted by σ_x^2 and σ_y^2 and the correlation coefficient by ρ.

As before, in order to use the regression estimator it is usually assumed that the population mean μ_x of the auxiliary variable is known. When μ_x is unknown, alternative information about the mean of the auxiliary variable is also obtained by the use of double sampling procedures. A stratified simple random sample of size n = 35 in which the pair (x_{hi}, y_{hi}) values are measured from n_h units drawn from each stratum and consequently leading to the estimation of the pair $(\bar{x}_{st}, \bar{y}_{st})$, with $n = \sum_h n_h$ as sample size. An additional sample of size m = 40, is selected by simple random sampling in which only x_i values are measured. Here $n' = n + m$ and the preliminary sample is a simple random sample of size $n' = 75$.

. In the present study, it is considered that there exists partial information about the mean of the auxiliary variable and let us assume that the partial information so obtained given by $\mu_0 = 75.0$.

Since the mean of the auxiliary variable is considered to be partially known as μ_0. We can assumed that $\mu_0 = 0$ without the loss of generality. The covariance matrix Σ is also considered to be known, hence without loss of generality it can be assumed that $\sigma_x^2 = 1$ and $\sigma_y^2 = 1$. In order to test the hypothesis $H_0 : \mu_x = 0$ against $H_1 : \mu_x \neq 0$, the sample values of the auxiliary variable X is transformed both in origin and scale maintaining the characteristics of X as $(x_i - \mu_0)/\sigma_X$, and the variable Y is transformed to y_i / σ_Y with value of $\sigma_X = 17.85$ and $\sigma_Y = 26.47$. The mean of X obtained from preliminary sample in double sampling is $\bar{x}_{n'} = -0.06024$.

If the hypothesis H_0 is accepted then $\mu_0 = 0.0$ obtained from the partial information will be used in the regression estimator t_5 and if H_0 is rejected, the sample mean $\bar{x}_n$ based on the preliminary sample is used in t_5 .The stratum means in terms of the transformed values in X and Y are shown in the Table 5.8. The table reveals that

$$\bar{x}_{st} = -0.1095 \quad , \quad \bar{y}_{st} = 5.022 \quad \text{and} \quad \bar{x}_{n'} = -0.06024$$

In terms of the previous origin and scale

$$\bar{x}_{st} = 72.62 \qquad \bar{y}_{st} = 132.93 \quad \text{and} \quad \bar{x}_{n'} = 73.69$$

The CRPTE in double sampling is given by

$$t_5 = \begin{cases} (\bar{y}_{st} - \rho \bar{x}_{st}) & if \quad \left|\bar{x}_{n'}\right| \le Z_\alpha / \sqrt{n'} \\ (\bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st})) & if \quad \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'} \end{cases}$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

ρ being the correlation coefficient of X and Y

The criteria for the acceptance and rejection of the hypothesis H_0 is based on the mean of the auxiliary variable obtained from the preliminary sample in double sampling. Table 5.9 shows that $\left|\bar{x}_{n'}\right| \le Z_\alpha / \sqrt{n'}$ for different level of significance α. Hence in the present case the hypothesis is accepted and consequently $\mu_0 = 0.0$ is utilized in the estimation of μ_x.

Thus,

$$t_5 = (\bar{y}_{st} - \rho \bar{x}_{st}) \quad if \quad \left|\bar{x}_{n'}\right| \le (Z_\alpha / \sqrt{n'})$$
$$= 5.12$$

and in terms of the previous origin and scale

$$t_5 = 135.44$$

Thus the estimate μ_y by the use of the CRPTE in double sampling is 135.44.

In order to judge the precision of the estimator, the MSE(t_5) is computed for different level of significance. The Mean square error of the CRPTE in double sampling is given by

$$MSE(t_5) = g_1 + h_1$$

where

$g_1 = (1-\rho^2)/n + \rho^2/n'$ and $h_1 = \{\rho^2/n'(A\varphi(A) - B\varphi(B)) - \rho^2(1/n' - \mu_x^{\ 2})(\Phi(A) - \Phi(B))\}$

The Table 5.10 reveals that the mean square error of t_5 increases with an increase in the value of the levels of significance.

5.5 Relative efficiency of CRPTE compared with the combined regression estimator under double sampling.

The combined regression estimator under double sampling without preliminary test is given by

$$t_4 = \bar{y}_{st} + \rho(\bar{x}_{n'} - \bar{x}_{st}) = 4.874$$

and in terms of the previous origin and scale

$$t_4 = 129.04$$

Thus the estimate μ_y, mean of Y variable by the use of the combined regression estimator in double sampling is 129.04

Also,

$$MSE\ (t_4) = (1/n)(1-\rho^2) + (1/n')\rho^2$$
$$= 0.01709$$

The relative efficiency of t_5 to t_4 is given by

$$e_1 = [MSE\ (t_4)\ ] / [MSE\ (t_5)\ ]$$

Table 5.11 clearly shows that MSE(t_5) is smaller than the MSE(t_4) for different values of level of significance α. Therefore we can say that when a reliable partial information on the mean of the auxiliary variable is available then the efficiency of the proposed estimator increases.

5.6 Discussion

We have seen in section 5.2 and 5.3, where real life data set was used for which a reliable partial information on the mean of the auxiliary variable is not available then the combined regression preliminary test estimator in double sampling is as efficient as the combined regression estimator in double sampling. However, in section 5.4 and 5.5 where a simulated data set was used and considered the availability of reliable partial information, then the proposed estimator becomes more efficient than the usual combined regression estimator in double sampling. Thus we may conclude that under the stated assumptions the combined regression preliminary test estimator(CRPTE) in double sampling is more efficient than the usual combined regression estimator, only when a reliable information about the mean of the auxiliary variable is available.

Table 5.1 Shapiro-Wilk test statistic for testing normality of complete child immunization (Y) and female literacy rate (X) for the EAG states

EAG States	Stratum	Test statistic for X	p – value for X	Test statistic for Y	p – value for Y
Bihar	1	0.97699	0.74116	0.99332	0.99933
Chatisgarh	2	0.9671	0.87681	0.83346	0.08628
Jharkand	3	0.90267	0.1455	0.95378	0.65657
Madhya Pradesh	4	0.98086	0.74789	0.98352	0.83675
Orrisa	5	0.93633	0.07245	0.96564	0.42766
Rajasthan	6	0.95441	0.22161	0.9572	0.26227
Uttaranchal	7	0.92988	0.44674	0.85108	0.05984
COMBINED		0.98303	0.04973	0.97653	0.00854

Table 5.2 Strata sample sizes obtained by proportional allocation for the EAG states

EAG States	Stratum	N_h	$W_h = (N_h/N)$	n_h
Bihar	1	30	0.189873	6
Chatisgarh	2	7	0.044304	1
Jharkand	3	13	0.082278	2
Madhya Pradesh	4	38	0.240506	7
Orrisa	5	30	0.189873	6
Rajasthan	6	30	0.189873	6
Uttaranchal	7	10	0.063291	2
TOTAL		158	1	30

Table 5.3 Computation of the strata means $\bar{x}_{st}$ and $\bar{y}_{st}$ for the EAG states

EAG States	Stratum	W_h	$\bar{x}_h$	$\bar{y}_h$	$W_h\bar{x}_h$	$W_h\bar{y}_h$
Bihar	1	0.189873	0.6143	2.237	0.116639	0.424747
Chatisgarh	2	0.044304	0.6133	3.717	0.027172	0.164677
Jharkand	3	0.082278	2.036	3.75	0.167519	0.308544
Madhya Pradesh	4	0.240506	1.894	3.524	0.455519	0.847544
Orrisa	5	0.189873	1.582	3.965	0.30038	0.752848
Rajasthan	6	0.189873	1.525	2.994	0.289557	0.568481
Uttaranchal	7	0.063291	2.219	4.424	0.140443	0.28
	TOTAL	1			$\bar{x}_{st} = 1.497228$	$\bar{y}_{st} = 3.346842$

Source: Rapid household survey (RHS-RCH project, phase 1,1998).

Table 5.4 Computation for testing H_0 for different values of level of significance α

α	Z_α	$(Z_\alpha / \sqrt{n'})$	$\lvert x_{n'} \rvert$
0.01	2.58	0.30836898	1.438
0.05	1.96	0.23426481	
0.25	1.16	0.13864652	

Table 5.5 MSE(t_5) for different values of level of significance α

α	$MSE(t_5)$
0.01	0.025595
0.05	0.025595
0.25	0.025595

Table 5.6 Shapiro-Wilk test statistic for testing normality of X and Y generated by simulation

Stratum	Test statistic for X	p – value for X	Test statistic for Y	p – value for Y
1	0.962	0.268	0.972	0.506
2	0.986	0.909	0.965	0.247
3	0.983	0.714	0.987	0.854
4	0.967	0.23	0.96	0.122
COMBINED	0.9899	0.2699	0.9903	0.304

Table 5.7 Strata sample sizes obtained by proportional allocation for data generated by simulation

Stratum	N_h	$W_h = (N_h/N)$	n_h
1	35	0.205882	7
2	40	0.235294	8
3	50	0.294118	10
4	45	0.264706	9
Total	**170**	**1**	**35**

Table 5.8 Computation of the strata means $\bar{x}_{st}$ and $\bar{y}_{st}$ for data generated by simulation

Stratum	W_h	$\bar{x}_h$	$\bar{y}_h$	$W_h\bar{x}_h$	$W_h\bar{y}_h$
1	0.205882	1.253061	6.524664	0.257983	1.343313
2	0.235294	-0.06938	5.152176	-0.01632	1.212277
3	0.294118	-0.38026	4.806777	-0.11184	1.413758
4	0.264706	-0.90399	3.976942	-0.23929	1.05272
TOTAL	1			$\bar{x}_{st} = -0.10947$	$\bar{y}_{st} = 5.022068$

Table 5.9 Computation for testing H_0 for different values of level of significance α for data generated by simulation

α	Z_α	$(Z_\alpha / \sqrt{n'})$	$\lvert x_{n'} \rvert$
0.01	2.58	0.297913	0.06024
0.05	1.96	0.226321	
0.25	1.16	0.133945	

Table 5.10 MSE(t_5) for different values of level of significance α for data generated by simulation

α	$MSE\ (t_5)$
0.01	0.008119759
0.05	0.010574744
0.25	0.015758148

Table 5.11 Relative efficiency of t_5 to t_4 for different values of level of significance α for data generated by simulation

α	MSE(t_4)	Relative Efficiency
0.01	0.017090682	2.104
0.05	0.017090682	1.616
0.25	0.017090682	1.0845

APPENDIX 1

Fortran 77 program to evaluate an integral (I_1) using Simpson's 1/3rd rule

```
      F(X)=0.398862*((PI+(X/SQRT(200.0)))*EXP(-0.5*X*X)
      PI = -0.1
      DO 10 I=1,11
      B=10
      N=50
      PI=PI+0.1
      Z=2.58
      A=Z-PI*SQRT(200.0)
      PRINT 2,A
2     FORMAT(////"THE VALUE OF IOWER LIMIT A=",F10.4)
      H=(B-A)/N
      X0=A
      XN=B
      SUM=F(X0)+F(XN)
      DO 20 K=1,N-1
        XK=X0+K*H
        IF (MOD(K,2).EQ.0) THEN
          SUM=SUM+2*F(XK)
        ELSE
          SUM=SUM+4*F(XK)
        ENDIF
20    ENDDO
      SUM=(H/3)*SUM
      PRINT 6
6     FORMAT(//1X,"RESULT:",/1X)
      PRINT*,"THE VALUE OF THE INTEGRAL IS:",SUM
10    ENDDO
      STOP
      END
```

*Note a similar program can be run for a set of values $\alpha = 0.05$, $\alpha = 0.25$

APPENDIX 2

Fortran 77 program to evaluate an integral (I_2) using Simpson's 1/3rd rule

```
      F(X)=0.398862*((PI+(X/SQRT(200.0)))*EXP(-0.5*X*X)
      PI =-0.1
      DO 10 I=1, 11
      A= -15
      N=50
      PI=PI+0.1
      Z=2.58
      B=-Z-PI*SQRT (200.0)
      PRINT 2, B
2     FORMAT(//"THE VALUE OF IOWER LIMIT B=",F10.4)
      H=(B-A)/N
      X0=A
      XN=B
      SUM=F(X0)+F(XN)
      DO 20 K=1,N-1
        XK=X0+K*H
        IF (MOD(K,2).EQ.0) THEN
          SUM=SUM+2*F(XK)
        ELSE
          SUM=SUM+4*F(XK)
        ENDIF
20    ENDDO
      SUM=(H/3)*SUM
      PRINT 6
6     FORMAT(//1X,"RESULT:",/1X)
      PRINT*,"THE VALUE OF THE INTEGRAL IS:",SUM
10    ENDDO
      STOP
      END
```

*Note a similar program can be run for a set of values $\alpha = 0.05$, $\alpha = 0.25$

APPENDIX 3

Evaluation of $E\{(\bar{x}_{n'}{}^2)$ if $|\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\}$

Let $I_1 = E(\bar{x}_{n'}{}^2 \ \ if \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'})$

$$= \{E(\bar{x}_{n'}{}^2) \ \ if \ \ \bar{x}_{n'} > Z_\alpha / \sqrt{n'}\} + \{E(\bar{x}_{n'}{}^2) \ \ if \ \ \bar{x}_{n'} < -Z_\alpha / \sqrt{n'}\}$$

$$= \int_{Z_\alpha / \sqrt{n'}}^{\infty} \bar{x}_{n'}{}^2 f(\bar{x}_{n'}) d\bar{x}_{n'} + \int_{-\infty}^{-Z_\alpha / \sqrt{n'}} \bar{x}_{n'}{}^2 f(\bar{x}_{n'}) d\bar{x}_{n'}$$

Since $\bar{x}_{n'} \sim N(\mu_x, 1/n')$ we get

$$f(\bar{x}_{n'}) = \frac{\sqrt{n'}}{\sqrt{2\pi}} Exp\left\{-\frac{1}{2}((\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}))^2\right\}$$

Thus the integrals becomes

$$I_1 = (\sqrt{n'}/\sqrt{2\pi}) \int_{Z_\alpha / \sqrt{n'}}^{\infty} \bar{x}_{n'}{}^2 Exp\left[-(1/2)\{(\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'})\}^2\right] d\bar{x}_{n'}$$

$$+ (\sqrt{n'}/\sqrt{2\pi}) \int_{-\infty}^{-Z_\alpha / \sqrt{n'}} \bar{x}_{n'}{}^2 Exp\left[-(1/2)\{(\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'})\}^2\right] d\bar{x}_{n'}$$

Putting $w = (\bar{x}_{n'} - \mu_x)/(1/\sqrt{n'}) \Rightarrow dw = \sqrt{n'} d\bar{x}_{n'}$, we have

When $\bar{x}_{n'} = Z_\alpha / \sqrt{n'}$ then $w = \sqrt{n'}(Z_\alpha / \sqrt{n'} - \mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$ and

When $\bar{x}_{n'} = -Z_\alpha / \sqrt{n'}$ then $w = \sqrt{n'}(-Z_\alpha / \sqrt{n'} - \mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

Hence I_1 becomes

$$= (1/\sqrt{2\pi})\left[\int_A^{\infty}\{\mu_x + (w/\sqrt{n'})\}^2 Exp\left[-(1/2)w^2\right]dw + \int_{-\infty}^{B}\{\mu_x + (w/\sqrt{n'})\}^2 Exp\left[-(1/2)w^2\right]dw\right]$$

$$= \mu_x{}^2\left[(1/\sqrt{2\pi})\int_A^{\infty} Exp\,((-1/2)w^2)dw + (1/\sqrt{2\pi})\int_{-\infty}^{B} Exp\,((-1/2)w^2)dw\right]$$

$$+ (2\mu_x/\sqrt{2\pi n'})\left[\int_A^{\infty} wExp\,((-1/2)w^2)dw + \int_{-\infty}^{B} wExp\,((-1/2)w^2)dw\right]$$

$$+ \{1/(n'\sqrt{2\pi}\}\left[\int_A^{\infty} w^2 Exp\,((-1/2)w^2)dw + \int_{-\infty}^{B} w^2 Exp\,((-1/2)w^2)dw\right]$$

$$= \mu_x{}^2\{1 - \Phi(A) + \Phi(B)\} + (2\mu_x/\sqrt{n'})\{\varphi(A) - \varphi(B)\}$$

$$+ (1/(n'\sqrt{2\pi}))\left[\int_A^{\infty} w^2 Exp\,((-1/2)w^2)dw + \int_{-\infty}^{B} w^2 Exp\,((-1/2)w^2)dw\right]$$

(follows from the evaluation of I in Bias)

Integration by parts, we get

$$I_1 = \mu_x{}^2\{1 - \Phi(A) + \Phi(B)\} + (2\mu_x/\sqrt{n'})\{\varphi(A) - \varphi(B)\}$$

$$+ \{1/(n'\sqrt{2\pi})\}\left[\left\{-wExp\,((-1/2)w^2)\right\}_A^{\infty} + \int_A^{\infty} Exp\,((-1/2)w^2)dw\right]$$

$$+ \{1/(n'\sqrt{2\pi})\}\left[\left\{-wExp\,((-1/2)w^2)\right\}_{-\infty}^{B} + \int_{-\infty}^{B} Exp\,((-1/2)w^2)dw\right]$$

$$= \mu_x{}^2\{1 - \Phi(A) + \Phi(B)\} + (2\mu_x / \sqrt{n'})\{\varphi(A) - \varphi(B)\}$$
$$+ (1/n')\{A\varphi(A) + 1 - \Phi(A)\} + (1/n')\{-B\varphi(B) + \Phi(B)\}$$

$$= \{\mu_x{}^2 + (1/n')\}\{1 - \Phi(A) + \Phi(B)\}$$
$$+ (2\mu_x / \sqrt{n'})\{\varphi(A) - \varphi(B)\} + (1/n')\{A\varphi(A) - B\varphi(B)\}$$

Thus,

$$I_1 = E(\bar{x}_{n'}{}^2 / \quad |\bar{x}_{n'}| > z_\alpha / \sqrt{n'})$$

$$= (\mu_x{}^2 + 1/n')\{(1 - \Phi(A) + \Phi(B)\}$$
$$+ (2\mu_x / \sqrt{n'})\{\varphi(A) - \varphi(B)\} + (1/n')\{A\varphi(A) - B\varphi(B)\}$$

APPENDIX 4

Evaluation of $E(\bar{x}_{st}\bar{x}_{n'}$ if $\left|\bar{x}_{n'}\right| > Z_\alpha/\sqrt{n'})$

Let $$I_2 = \left\{E(\bar{x}_{st}\bar{x}_{n'}) \;/\; \left|\bar{x}_{n'}\right| > Z_\alpha/\sqrt{n'}\right\} = \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}})\right\} \;if\; \left|\bar{x}_{n'}\right| > Z_\alpha/\sqrt{n'}\right]_{t_1=t_2=0}$$

Here $E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}})$ is the moment generating function of the pair $(\bar{x}_{st}, \bar{x}_{n'})$. Again let

$$I_2' = \left\{E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}}) \quad if \quad \left|\bar{x}_{n'}\right| > Z_\alpha/\sqrt{n'}\right\}$$

$$= \left[E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}}) \;if\; \bar{x}_{n'} > Z_\alpha/\sqrt{n'}\right] + \left[E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}}) \;if\; \bar{x}_{n'} < -Z_\alpha/\sqrt{n'}\right]$$

$$I_2' = \int_{\bar{x}_{st}=-\infty}^{\infty}\int_{\bar{x}_{n'}=Z_\alpha/\sqrt{n'}}^{\infty} e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}} f(\bar{x}_{st}, \bar{x}_{n'})\, d\bar{x}_{st}\, d\bar{x}_{n'}$$

$$+ \int_{\bar{x}_{st}=-\infty}^{\infty}\int_{\bar{x}_{n'}=-\infty}^{-Z_\alpha/\sqrt{n'}} e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}} f(\bar{x}_{st}, \bar{x}_{n'})\, d\bar{x}_{st}\, d\bar{x}_{n'}$$

where $f(\bar{x}_{st}, \bar{x}_{n'})$ is the bivariate normal probability density function of the pair $(,\bar{x}_{st}, \bar{x}_{n'})$ with mean (μ_x, μ_x) and the variance covariance matrix given by

$$\sum_{1} = \begin{bmatrix} 1/n & 1/n' \\ 1/n' & 1/n' \end{bmatrix}$$ Under the assumption that $\sigma_{x_s}^{\ 2} = \sigma_{y_s}^{\ 2} = 1, \sigma_x^{\ 2} = \sigma_y^{\ 2} = 1$ (WLOG)

For a bivariate normal density we are given that

$$f(x,y) = \{1/(2\pi|\Sigma|^{1/2})\} Exp\left[(-1/2)\{((x-\mu_x)\ (y-\mu_y))(\Sigma^{-1})\begin{pmatrix}(x-\mu_x)\\(y-\mu_y)\end{pmatrix}\right]$$

Here $$\sum_{1}^{-1} = n'/(n'-n)\begin{pmatrix} n & -n \\ -n & n' \end{pmatrix}$$

Thus,

$$f(\bar{x}_{st}, \bar{x}_{n'}) = \frac{1}{2\pi\{(1/nn') - (1/n'^2)\}^{1/2}} Exp\left[-\frac{1}{2}\left\{\begin{bmatrix}\bar{x}_{st}-\mu_x & \bar{x}_{n'}-\mu_x\end{bmatrix} n'/(n'-n)\begin{pmatrix} n & -n \\ -n & n' \end{pmatrix}\begin{bmatrix}\bar{x}_{st}-\mu_x \\ \bar{x}_{n'}-\mu_x\end{bmatrix}\right\}\right]$$

$$= \frac{1}{2\pi\left\{(1/nn') - (1/n'^2)\right\}^{1/2}} Exp\left[-\frac{1}{2}(n'/(n'-n))\left\{n(\bar{x}_{st}-\mu_x)^2 - 2n(\bar{x}_{st}-\mu_x)(\bar{x}_{n'}-\mu_x) + n'(\bar{x}_{n'}-\mu_x)^2\right\}\right]$$

$$= \frac{1}{2\pi\left\{(1/nn') - (1/n'^2)\right\}^{1/2}} Exp\left[-\frac{1}{2}(n'/(n'-n))\left\{n(\bar{x}_{st}-\mu_x)^2 - 2\sqrt{\frac{n}{n'}}\frac{(\bar{x}_{st}-\mu_x)}{1/\sqrt{n}}\frac{(\bar{x}_{n'}-\mu_x)}{1/\sqrt{n'}} + n'(\bar{x}_{n'}-\mu_x)^2\right\}\right]$$

Letting $(\bar{x}_{st}-\mu_x)/(1/\sqrt{n}) = x_1$ and $(\bar{x}_{n'}-\mu_x)/(1/\sqrt{n'}) = x_2$, we get $d\bar{x}_{st} = dx_1/\sqrt{n}$ and $d\bar{x}_{n'} = dx_2/\sqrt{n'}$

When $\bar{x}_{n'} = Z_\alpha/\sqrt{n'}$ then $x_2 = \sqrt{n'}(Z_\alpha/\sqrt{n'} - \mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$ and

when $\bar{x}_{n'} = -Z_\alpha/\sqrt{n'}$ then $x_2 = \sqrt{n'}(-Z_\alpha/\sqrt{n'} - \mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

Hence,

$$I_2' = \left[\{\sqrt{n'}/(2\pi\sqrt{n'-n})\}\int_{-\infty}^{\infty}\int_{x_2=A}^{\infty} Exp\left[t_1\{\mu_x + (x_1/\sqrt{n})\} + t_2\{\mu_x + (x_2/\sqrt{n'})\}\right] * Exp\left[-\{n'/2(n'-n)\}\{x_1^2 - 2\sqrt{n/n'}x_1x_2 + x_2^2\}\right]dx_1dx_2\right.$$

$$\left. + \{\sqrt{n'}/(2\pi\sqrt{n'-n})\}\int_{-\infty}^{\infty}\int_{-\infty}^{x_2=B} Exp\left[t_1\{\mu_x + (x_1/\sqrt{n})\} + t_2\{\mu_x + (x_2/\sqrt{n'})\}\right] Exp\left[-\{n'/2(n'-n)\}\{x_1^2 - 2\sqrt{n/n'}x_1x_2 + x_2^2\}\right]dx_1dx_2\right]$$

$$= \{\sqrt{n'}/(2\pi\sqrt{n'-n})\}Exp(t_1\mu_x + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{x_2=A}^{\infty} Exp[((t_1x_1/\sqrt{n})+(t_2x_2/\sqrt{n'})) - (n'/2(n'-n))\{x_1^{\,2} - 2\sqrt{n/n'}x_1x_2 + x_2^{\,2}\}]dx_1dx_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{x_2=B} Exp[((t_1x_1/\sqrt{n})+(t_2x_2/\sqrt{n'})) - (n'/2(n'-n))\{x_1^{\,2} - 2\sqrt{n/n'}x_1x_2 + x_2^{\,2}\}]dx_1dx_2\right]$$

$$= \sqrt{n'}/(2\pi\sqrt{n'-n})\}Exp(t_1\mu_x + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{x_2=A}^{\infty} Exp[(-n'/2(n'-n))\{(x_1^{\,2} - 2\sqrt{n/n'}x_1x_2 + x_2^{\,2}) - (2(n'-n)/n')(t_1x_1/\sqrt{n})+(t_2x_2/\sqrt{n'})\}]dx_1dx_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{x_2=B} Exp[(-n'/2(n'-n))\{(x_1^{\,2} - 2\sqrt{n/n'}x_1x_2 + x_2^{\,2}) - (2(n'-n)/n')(t_1x_1/\sqrt{n})+(t_2x_2/\sqrt{n'})\}]dx_1dx_2\right]$$

Now we can rewrite

$$\{x_1^{\,2} - 2\sqrt{n/n'}x_1x_2 + x_2^{\,2}\} - (2(n'-n)/n')((t_1x_1/\sqrt{n}) + (t_2x_2/\sqrt{n'}))$$
$$= \{(x_1 - (\sqrt{n/n'})x_2) - (1-n/n')t_1/\sqrt{n}\}^2$$
$$+ (1-n/n')\{x_2 - (\sqrt{n/n'})t_1/\sqrt{n} - t_2/\sqrt{n'}\}^2 - t_1^2/n - t_2^2/n' - 2(\sqrt{n/n'})t_1t_2(1/\sqrt{nn'})$$

Substituting

$$(x_1 - (\sqrt{n/n'})x_2) - (1-n/n')(t_1/\sqrt{n}) = \{(n'-n)/n'\}^{1/2}u \qquad \text{and}$$

$$x_2 - (\sqrt{n/n'})(t_1/\sqrt{n}) - (t_2/\sqrt{n'}) = v$$

Then in this transformation, the Jacobian is given by

$$J = \frac{\partial(u,v)}{\partial(x_1,x_2)} = \sqrt{1 - (\tfrac{n}{n'})} \qquad dudv = |J|dx_1dx_2$$

When

$$x_2 = A \quad , \quad v = A - (t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = A - \{(t_1+t_2)/\sqrt{n'}\} = A'$$

$$x_2 = B \quad , \quad v = B - (t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = B - \{(t_1+t_2)/\sqrt{n'}\} = B'$$

Thus

$$I_2^{'} = \{\sqrt{n'}/(2\pi\sqrt{n'-n})\}Exp(t_1\mu_x + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{v=A'}^{\infty} Exp[(-(n'/2(n'-n))[((n'-n)/n')u^2 + (1-n/n')\{v^2 - (t_1^2/n) - (t_2^2/n') - (2t_1t_2/n')\}]\{(\sqrt{n'-n})/n'\}dudv \right.$$

$$\left. + \int_{-\infty}^{\infty}\int_{-\infty}^{v=B'} Exp[(-(n'/2(n'-n))[((n'-n)/n')u^2 + (1-n/n')\{v^2 - (t_1^2/n) - (t_2^2/n') - (2t_1t_2/n')\}]\{(\sqrt{n'-n})/n'\}dudv \right]$$

$$= \{1/2\pi\}Exp(t_1\mu_x + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{v=A'}^{\infty} Exp[(-(n'/2(n'-n))\{((n'-n)/n')u^2 + (1-n/n')v^2 - (1-n/n')\{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\}]\}dudv \right.$$

$$\left. + \int_{-\infty}^{\infty}\int_{-\infty}^{v=B'} Exp[(-(n'/2(n'-n))\{((n'-n)/n')u^2 + (1-n/n')v^2 - (1-n/n')\{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\}]\}dudv \right]$$

$$I_2^{'} = Exp\{t_1\mu_x + t_2\mu_x + (1/2)\{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\}$$
$$\left\{(1/2\pi)\int_{u=\infty}^{\infty}\int_{v=A'}^{\infty} Exp(-(1/2)(u^2+v^2)dudv + (1/2\pi)\int_{u-\infty}^{\infty}\int_{v=-\infty}^{B'} Exp(-(1/2)(u^2+v^2)dudv\right\}$$

$$= Exp\{t_1\mu_x + t_2\mu_x + (1/2)\{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\}$$
$$\left[(1/\sqrt{2\pi})\int_{u=-\infty}^{\infty} Exp\{-(1/2)(u^2)\}du\left\{(1/\sqrt{2\pi})\int_{v=A'}^{\infty} Exp\{(-(1/2)(v^2)\}dv\right\} +\right.$$
$$\left.(1/\sqrt{2\pi})\int_{u=-\infty}^{\infty} Exp\{-(1/2)(u^2)\}du\left\{(1/\sqrt{2\pi})\int_{v=-\infty}^{B'} Exp\{(-(1/2)(v^2)\}dv\right\}\right]$$

$$I_2^{'} = Exp\{t_1\mu_x + t_2\mu_x + (1/2)\{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\}\ \{(1 - \Phi(A') + \Phi(B')\}$$

Now,

$$I_2 = \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{E(e^{t_1\bar{x}_{st}+t_2\bar{x}_{n'}})\right\} \quad if \quad \left|\bar{x}_{n'}\right| > Z_\alpha/\sqrt{n'}\right]_{t_1=t_2=0}$$

$$= \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{Exp(t_1\mu_x + t_2\mu_x + (1/2)\left\{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\{1-\Phi(A')+\Phi(B')\}\right\}\right\}\right]_{t_1=t_2=0}$$

Differentiating under an integral sign is given by the formula

$$\xi(y) = \int_{g(y)}^{h(y)} f(x, y)\,dx \quad , \quad \text{then}$$

$$\xi'(y) = \int_{g(y)}^{h(y)} f_y(x,y) + h'(y)f(h(y),y) - g'(y)f(g(y),y)$$

Differentiation under an integral sign, we get

$$I_2 = (\partial / \partial t_1)\Big[\Big\{\mu_x e^{\mu_x t_1 + \mu_x t_2}\,(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2/n')})(1 - \Phi(A') + \Phi(B'))\Big\} +$$
$$\Big\{e^{\mu_x t_1 + \mu_x t_2}\,(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2/n')})((t_1 + t_2)/n')(1 - \Phi(A') + \Phi(B'))\Big\} +$$
$$\Big\{e^{\mu_x t_1 + \mu_x t_2}\,(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2/n')})(1/\sqrt{2\pi n'})(e^{(-1/2)(A - (t_1+t_2)/\sqrt{n'})^2} - e^{(-1/2)(B - (t_1+t_2)/\sqrt{n'})^2}\Big\}\Big]$$

$$= \Big[\Big\{\mu_x^2 e^{\mu_x t_1 + \mu_x t_2}\,(e^{1/2(t_1^2/n + t_2^2/n' + 2t_1 t_2/n')})(1 - \Phi(A') + \Phi(B'))\Big\}$$
$$+ \Big\{\mu_x e^{\mu_x t_1 + \mu_x t_2}\,(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2/n')})((t_1 + t_2)/n')(1 - \Phi(A') + \Phi(B'))\Big\}$$

$$+ \Big\{\mu_x e^{\mu_x t_1 + \mu_x t_2}\,(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2/n')})$$
$$(1/\sqrt{2\pi n'}(e^{(-1/2)(A - (t_1+t_2)/\sqrt{n'})^2} - e^{(-1/2)(B - (t_1+t_2)/\sqrt{n'})^2})\Big\}$$

$$+\left\{\mu_x e^{\mu_x t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2/n')}((t_1+t_2)/n')(1-\Phi(A')+\Phi(B'))\right\}\right.$$

$$+\left\{e^{\mu_x t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2/n')}((t_1+t_2)/n')^2\right.\right.$$

$$\left.\left.+\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2/n')}(1/n')\right)(1-\Phi(A')+\Phi(B'))\right\}\right.$$

$$+\left\{e^{\mu_x t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2/n')}((t_1+t_2)/n')\right.\right.$$

$$\left.\left.(1/\sqrt{2\pi n'}\left(e^{(-1/2)(A-(t_1+t_2)\sqrt{n'})^2}-e^{(-1/2)(B-(t_1+t_2)/\sqrt{n'})^2}\right)\right\}\right.$$

$$+\left\{\mu_x e^{\mu_x t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2/n')}\right.\right.$$
$$\left.\left.(1/\sqrt{2\pi n'}\left(e^{(-1/2)(A-(t_1+t_2))^2}-e^{(-1/2)(B-(t_1+t_2))^2}\right)\right\}\right.$$

$$\left\{e^{\mu_x t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2/n')}((t_1+t_2)/n')\right.\right.$$
$$\left.\left.(e^{(-1/2)(A-(t_1+t_2))^2}-e^{(-1/2)(B-(t_1+t_2))^2}\right)\right\}$$

$$\left\{e^{\mu_x t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2/n')}\right.\right.$$
$$(1/\sqrt{2\pi n'})(e^{(-1/2)(A-(t_1+t_2)/\sqrt{n'})^2}((-1/2)2(A-(t_1+t_2)/\sqrt{n'})(-1/\sqrt{n'})$$
$$\left.\left.-e^{(-1/2)(B-(t_1+t_2)/\sqrt{n'})^2}((-1/2)2(B-(t_1+t_2)/\sqrt{n'})(-1/\sqrt{n'})\right\}\right.$$

When $t_1=t_2=0$, then

$$I_2=\mu_x^2(1-\Phi(A)+\Phi(B))+(\mu_x/\sqrt{n'})(\varphi(A)-\varphi(B))+(1/n')(1-\Phi(A)+\Phi(B))$$
$$+(\mu_x/\sqrt{n'})(\varphi(A)-\varphi(B))+(1/\sqrt{2\pi n'})(e^{-A^2/2}(A/\sqrt{n'})-e^{-B^2/2}(B/\sqrt{n'}))$$

$$=\mu_x^2(1-\Phi(A)+\Phi(B))+(2\mu_x/\sqrt{n'})(\varphi(A)-\varphi(B))$$
$$+(1/n')(1-\Phi(A)+\Phi(B))+(1/n')(A\varphi(A)-B\varphi(B))$$

Hence,

$$I_2 = E(\bar{x}_{st}, \bar{x}_{n'} \text{ if } \left|\bar{x}_{n'}\right| > Z_\alpha / \sqrt{n'})$$

$$= (\mu_x{}^2 + 1/n')\{(1 - \Phi(A) + \Phi(B)\}$$
$$+ (2\mu_x / \sqrt{n'})\{\varphi(A) - \varphi(B)\} + (1/n')\{A\varphi(A) - B\varphi(B)\}$$

APPENDIX 5

Evaluation of $E\{(\bar{y}_{st}, \bar{x}_{n'})\ if\ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'})\}$

Let $I_3 = \left\{E(\bar{y}_{st}\bar{x}_{n'})\ \ /\ \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right\}$

$$= \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})\right\}\ \ if\ \ |\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right]_{t_1=t_2=0}$$

Here $E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})$ is the moment generating function of the pair $(\bar{y}_{st}, \bar{x}_{n'})$.

Again let

$$I_3' = \left\{E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})if|\bar{x}_{n'}| > Z_\alpha / \sqrt{n'}\right\}$$

$$= \left[E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})\ if\ \ \bar{x}_{n'} > Z_\alpha / \sqrt{n'}\right] + \left[E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})\, if\ \ \bar{x}_{n'} < -Z_\alpha / \sqrt{n'}\right]$$

$$I_3' = \int_{\bar{x}_{st}=-\infty}^{\infty} \int_{\bar{x}_{n'}=Z_\alpha/\sqrt{n'}}^{\infty} e^{t_1 \bar{y}_{st} + t_2 \bar{x}_{n'}} f(\bar{y}_{st}, \bar{x}_{n'}) d\bar{y}_{st} d\bar{x}_{n'}$$

$$+ \int_{\bar{x}_{st}=-\infty}^{\infty} \int_{\bar{x}_{n'}=-\infty}^{-Z_\alpha/\sqrt{n'}} e^{t_1 \bar{y}_{st} + t_2 \bar{x}_{n'}} f(\bar{y}_{st}, \bar{x}_{n'}) d\bar{y}_{st} d\bar{x}_{n'}$$

where $f(\bar{y}_{st}, \bar{x}_{n'})$ is the bivariate normal probability density function of the pair $(\bar{y}_{st}, \bar{x}_{n'})$ with mean (μ_y, μ_x) and the variance covariance matrix given by $\sum_2 = \begin{bmatrix} 1/n & \rho/n' \\ \rho/n' & 1/n' \end{bmatrix}$

Under the assumption that $\sigma_{x_i}^2 = \sigma_{y_i}^2 = 1, \sigma_x^2 = \sigma_y^2 = 1$ (WLOG)

For a bivariate normal density we are given that

$$f(x,y) = \{1/(2\pi|\Sigma|^{1/2})\} Exp\left[(-1/2)\{((x-\mu_x)\ (y-\mu_y))(\Sigma^{-1})\begin{pmatrix}(x-\mu_x)\\(y-\mu_y)\end{pmatrix}\right]$$

Here, $$\sum_2^{-1} = n'/(n'-\rho^2 n)\begin{pmatrix} n & -n\rho \\ -n\rho & n' \end{pmatrix}$$

Thus, $$f(\bar{y}_{st},\bar{x}_{n'}) = \frac{1}{2\pi\{(1/nn')-(\rho^2/n'^2)\}^{1/2}} Exp\left[-\frac{1}{2}\frac{n'}{n'-n\rho^2}\left\{n(\bar{y}_{st}-\mu_x)^2 - 2\rho\sqrt{\frac{n}{n'}}\frac{(\bar{y}_{st}-\mu_x)}{1/\sqrt{n}}\frac{(\bar{x}_{n'}-\mu_x)}{1/\sqrt{n'}} + n'(\bar{x}_{n'}-\mu_x)^2\right\}\right]$$

Letting $(\bar{y}_{st}-\mu_y)/(1/\sqrt{n}) = z_1$ and $(\bar{x}_{n'}-\mu_x)/(1/\sqrt{n'}) = z_2$, we get $d\bar{y}_{st} = dz_1/\sqrt{n}$ and $d\bar{x}_{n'} = dz_2/\sqrt{n'}$

When $\bar{x}_{n'} = Z_\alpha/\sqrt{n'}$ then $z_2 = \sqrt{n'}(Z_\alpha/\sqrt{n'}-\mu_x) = Z_\alpha - \sqrt{n'}\mu_x = A$ and

When $\bar{x}_{n'} = -Z_\alpha/\sqrt{n'}$ then $z_2 = \sqrt{n'}(-Z_\alpha/\sqrt{n'}-\mu_x) = -Z_\alpha - \sqrt{n'}\mu_x = B$

Thus,

$$I_3' = \{\sqrt{n'}/(2\pi\sqrt{n'-\rho^2 n})\}\left[\int_{-\infty}^{\infty}\int_{z_2=A}^{\infty} Exp\left[t_1\{\mu_y+(z_1/\sqrt{n})\}+t_2\{\mu_x+(z_2/\sqrt{n'})\}\right] Exp\left[-\{n'/2(n'-\rho^2 n)\}\{z_1^2 - 2\rho\sqrt{n/n'}z_1z_2 + z_2^2\}\right]dz_1dz_2\right.$$

$$\left. + \int_{-\infty}^{\infty}\int_{-\infty}^{z_2=B} Exp\left[t_1\{\mu_y+(z_1/\sqrt{n})\}+t_2\{\mu_x+(z_2/\sqrt{n'})\}\right] Exp\left[-\{n'/2(n'-\rho^2 n)\}\{z_1^2 - 2\rho\sqrt{n/n'}z_1z_2 + z_2^2\}\right]dz_1dz_2\right]$$

$$= \{\sqrt{n'}/(2\pi\sqrt{n'-\rho^2 n})\} Exp(t_1\mu_y + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{z_2=A}^{\infty} Exp[((t_1 z_1/\sqrt{n}) + (t_2 z_2/\sqrt{n'}) - \{n'/2(n'-\rho^2 n)\}\{z_1^{\,2} - 2\sqrt{n/n'}z_1 z_2 + z_2^{\,2}\}]dz_1 dz_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{z_2=B} [Exp((t_1 z_1/\sqrt{n}) + (t_2 z_2/\sqrt{n'}) - \{n'/2(n'-\rho^2 n)\}\{z_1^{\,2} - 2\sqrt{n/n'}z_1 z_2 + z_2^{\,2}\}]dz_1 dz_2\right]$$

$$= \{\sqrt{n'}/(2\pi\sqrt{n'-\rho^2 n})\} Exp\,(t_1\mu_y + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{z_2=A}^{\infty} Exp\,[(-(n'/2(n'-n\rho^2))\{(z_1^{\,2} - 2\rho\sqrt{n/n'}z_1 z_2 + z_2) - (2(n'-n\rho^2)/n')((t_1 z_1/\sqrt{n}) + (t_2 z_2/\sqrt{n'}))\}]\,dz_1 dz_2 + \int_{-\infty}^{\infty}\int_{-\infty}^{z_2=B} Exp\,[(-(n'/2(n'-n\rho^2))\{(z_1^{\,2} - 2\rho\sqrt{n/n'}z_1 z_2 + z_2^{\,2}) \cdot - (2(n'-n\rho^2)/n')((t_1 z_1/\sqrt{n}) + (t_2 z_2/\sqrt{n'}))\}]\,dz_1 dz_2\right]$$

Now we can rewrite

$$\{z_1^{\,2} - 2\rho\sqrt{n/n'}z_1z_2 + z_2^{\,2}\} - \left\{2(n' - n\rho^2)/n')((t_1z_1/\sqrt{n}) + (t_2z_2/\sqrt{n'})\right\}$$
$$= \left\{(z_1 - \rho(\sqrt{n/n'})z_2) - (1 - n\rho^2/n')t_1/\sqrt{n}\right\}^2$$
$$+ (1 - n\rho^2/n')\left[\{z_2 - \rho t_1/\sqrt{n'} - t_2/\sqrt{n'}\}^2 - t_1^2/n - t_2^2/n' - 2\rho t_1t_2/n'\right]$$

Substituting

$$(z_1 - (\rho\sqrt{n/n'})z_2) - (1 - n\rho^2/n')(t_1/\sqrt{n}) = \{(n' - n\rho^2)/n'\}^{1/2}u \quad \text{and}$$

$$z_2 - \rho(t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = v$$

Then in this transformation, the Jacobian is given by

$$J = \frac{\partial(u,v)}{\partial(x_1,x_2)} = \sqrt{1 - (\frac{\rho^2 n}{n'})} \qquad dudv = |J|dz_1dz_2$$

When

$$z_2 = A \;,\; v = A - (\rho t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = A - \{(\rho t_1 + t_2)/\sqrt{n'}\} = A'$$

$$z_2 = B \;,\; v = B - (\rho t_1/\sqrt{n'}) - (t_2/\sqrt{n'}) = B - \{(\rho t_1 + t_2)/\sqrt{n'}\} = B'$$

Hence,

$$I_3^{'} = \{\sqrt{n'}/(2\pi\sqrt{n'-\rho^2 n})\}Exp(t_1\mu_y + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{v=A'}^{\infty} Exp(-(n'/2(n'-\rho^2 n))[((n'-\rho^2 n)/n')u^2 + \{(n'-n\rho^2)\}/n')[v^2 - \{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\}]\sqrt{(n'-\rho^2 n)/n'}\}dudv\right.$$

$$\left. + \int_{-\infty}^{\infty}\int_{-\infty}^{v=B'} Exp(-(n'/2(n'-\rho^2 n))[((n'-\rho^2 n)/n')u^2 + \{(n'-n\rho^2)/n')[v^2 - \{(t_1^2/n) + (t_2^2/n') + (2t_1t_2/n')\}]\sqrt{(n'-\rho^2 n)/n'}\}dudv\right]$$

$$I_3^{'} = \{1/(2\pi)\}Exp(t_1\mu_y + t_2\mu_x)\left[\int_{-\infty}^{\infty}\int_{v=A'}^{\infty} Exp(-(n'/2(n'-\rho^2 n))[((n'-\rho^2 n)/n')u^2 + (n'-n\rho^2/n')(v^2 - (t_1^2/n) - (t_2^2/n') - (2t_1t_2/n')]dudv\right.$$

$$\left. + \int_{-\infty}^{\infty}\int_{-\infty}^{v=B'} Exp(-(n'/2(n'-\rho^2 n))[((n'-n)/n')u^2 + (n'-n\rho^2/n')(v^2 - (t_1^2/n) - (t_2^2/n') - (2t_1t_2/n')]dudv\right]$$

$$= Exp\{t_1\mu_y + t_2\mu_x + (1/2)\{(t_1^2/n') + (t_2^2/n) + (2\rho t_1 t_2/n')\}$$
$$\left\{(1/2\pi)\int_{u=\infty}^{\infty}\int_{v=A'}^{\infty} Exp(-(1/2)(u^2+v^2)dudv + (1/2\pi)\int_{u-\infty}^{\infty}\int_{v=-\infty}^{B'} Exp(-(1/2)(u^2+v^2)dudv\right\}$$

$$= Exp\{t_1\mu_x + t_2\mu_y + (1/2)\{(t_1^2/n') + (t_2^2/n) + (2t_1 t_2\rho/n')\}$$
$$\left[(1/\sqrt{2\pi})\int_{u=-\infty}^{\infty}[Exp\{(-(1/2)(u^2)\}du]\int_{v=A'}^{\infty}[Exp\{(-(1/2)(v^2)\}dv\right.$$
$$\left. + (1/\sqrt{2\pi})\int_{u=-\infty}^{\infty}[Exp\{(-(1/2)(u^2)\}du]\int_{v=-\infty}^{B'}[Exp\{(-(1/2)(v^2)\}dv\right]$$

$$I_3' = Exp\{t_1\mu_y + t_2\mu_x + (1/2)\{(t_1^2/n') + (t_2^2/n) + (2\rho t_1 t_2/n')\}\{(1 - \Phi(A') + \Phi(B')\}$$

Now,

$$I_3 = \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{E(e^{t_1\bar{y}_{st}+t_2\bar{x}_{n'}})\right\} \quad if \quad \left|\bar{x}_{n'}\right| > Z_\alpha/\sqrt{n'}\right]_{t_1=t_2=0}$$

$$= \frac{\partial}{\partial t_1}\left[\frac{\partial}{\partial t_2}\left\{Exp(t_1\mu_y + t_2\mu_x + (1/2)\left\{(t_1^2/n)+(t_2^2/n')+(2\rho t_1 t_2/n')\{1-\Phi(A')+\Phi(B')\}\right\}\right\}\right]_{t_1=t_2=0}$$

Differentiating under an integral sign is given by the formula

$$\xi(y) = \int_{g(y)}^{h(y)} f(x,y)\,dx \quad , \quad \text{then}$$

$$\xi'(y) = \int_{g(y)}^{h(y)} f_y(x,y) + h'(y)f(h(y),y) - g'(y)f(g(y),y)$$

Differentiating I_3' under an integral sign and letting $t_1 = t_2 = 0$, we get

$$I_3 = (\partial / \partial t_1)\Big[\Big\{\mu_x e^{\mu_y t_1 + \mu_x t_2}\left(e^{1/2(t_1^2/n + t_2^2/n' + 2t_1 t_2 \rho/n')}\right)(1 - \Phi(A') + \Phi(B'))\Big\}$$
$$+ \Big\{e^{\mu_y t_1 + \mu_x t_2}\left(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2 \rho/n')}((\rho t_1 + t_2)/n')(1 - \Phi(A') + \Phi(B'))\Big\}$$
$$+ \{e^{\mu_y t_1 + \mu_x t_2}\left(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2 \rho/n')}\ (1/\sqrt{2\pi n'}\left(e^{(-1/2)(A - (\rho t_1 + t_2)/\sqrt{n'})^2} - e^{(-1/2)(B - (\rho t_1 + t_2)/\sqrt{n'})^2}\right)\}\Big]$$

$$= \Big[\Big\{\mu_x \mu_y e^{\mu_y t_1 + \mu_x t_2}\left(e^{1/2(t_1^2/n + t_2^2/n' + 2t_1 t_2 \rho/n')}\right)(1 - \Phi(A') + \Phi(B'))\Big\}$$
$$+ \Big\{\mu_x e^{\mu_y t_1 + \mu_x t_2}\left(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2 \rho/n')}\right)((t_1 + \rho t_2)/n')(1 - \Phi(A') + \Phi(B'))\Big\}$$
$$+ \Big\{\mu_x e^{\mu_y t_1 + \mu_x t_2}\left(e^{(1/2)(t_1^2/n + t_2^2/n' + 2t_1 t_2 \rho/n')}\right)(1/\sqrt{2\pi n'}\left(e^{(-1/2)(A - (\rho t_1 + t_2)/\sqrt{n'})^2} - e^{(-1/2)(B - (\rho t_1 + t_2)/\sqrt{n'})^2}\right)\Big\}\Big]$$

$$+\left\{\mu_y e^{\mu_y t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2\rho/n')}\right)((\rho t_1+t_2)/n')(1-\Phi(A')+\Phi(B'))\right\}$$

$$+\left\{e^{\mu_y t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2\rho/n')}\right)((\rho t_1+t_2)(t_1+\rho t_2)/n'^2\right.$$

$$\left.+\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2\rho/n')}\right)(\rho/n'))(1-\Phi(A')+\Phi(B'))\right\}$$

$$+\left\{e^{\mu_y t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2\rho/n')}\right)((\rho t_1+t_2)/n')\right.$$

$$\left.(1/\sqrt{2\pi n'}(e^{(-1/2)(A-(\rho t_1+t_2)\sqrt{n'})^2}-e^{(-1/2)(B-(\rho t_1+t_2)/\sqrt{n'})^2})\right\}$$

$$+\left\{\mu_y e^{\mu_y t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2\rho/n')}\right)\right.$$

$$\left.(1/\sqrt{2\pi n'}(e^{(-1/2)(A-(\rho t_1+t_2)\sqrt{n'})^2}-e^{(-1/2)(B-(\rho t_1+t_2)\sqrt{n'})^2})\right\}$$

$$+\left\{e^{\mu_y t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2\rho/n')}((t_1+\rho t_2)/n')\right.\right.$$

$$\left.\left.(1/\sqrt{2\pi n'})(e^{(-1/2)(A-(\rho t_1+t_2)\sqrt{n'})^2}-e^{(-1/2)(B-(\rho t_1+t_2)\sqrt{n'})^2})\right\}\right.$$

$$+\left\{e^{\mu_y t_1+\mu_x t_2}\left(e^{(1/2)(t_1^2/n+t_2^2/n'+2t_1t_2\rho/n')}\right)\right.$$

$$(1/\sqrt{2\pi n'})(e^{(-1/2)(A-(\rho t_1+t_2)/\sqrt{n'})^2}((-1/2)2(A-(\rho t_1+t_2)/\sqrt{n'})(-\rho/\sqrt{n'})\Big\}$$

$$\left.-e^{(-1/2)(B-(t_1+t_2)/\sqrt{n'})^2}((-1/2)2(B-(\rho t_1+t_2)/\sqrt{n'})(-\rho/\sqrt{n'})\Big\}\right]$$

When $t_1 = t_2 = 0$, then

$$I_3 = \mu_x\mu_y(1-\Phi(A)+\Phi(B)) + (\rho\mu_x/\sqrt{n'})(\varphi(A)-\varphi(B)) + (\rho/n')(1-\Phi(A)+\Phi(B))$$
$$+ \mu_y(\varphi(A)-\varphi(B))/\sqrt{n'} + (\rho/\sqrt{2\pi n'})(e^{-A^2/2}(A/\sqrt{n'}) - e^{-B^2/2}(B/\sqrt{n'}))$$

$$= (\mu_x\mu_y + \rho/n')(1-\Phi(A)+\Phi(B)) + (\mu_y + \rho\mu_x)(\varphi(A)-\varphi(B))$$
$$+ (\rho/n')(A\varphi(A) - B\varphi(B))$$

Hence,

$$E(\bar{y}_{st}, \bar{x}_{n'} / |\bar{x}_{n'}| > Z_\alpha/\sqrt{n'})$$

$$= (\mu_x\mu_y + \rho/n')\{(1-\Phi(A)+\Phi(B)\}$$
$$+ (1/\sqrt{n'})(\mu_y + \rho\mu_x)\{\varphi(A)-\varphi(B)\} + (\rho/n')\{A\varphi(A) - B\varphi(B)\}$$

where $\Phi(.)$ is the cumulative distribution of N(0,1) and $\varphi(.)$ is the density function.

APPENDIX 6

Evaluation of the Mean square error of t_4

The Combined regression estimator under double sampling is given by

$$t_4 = \bar{y}_{st} + b_{yx}(\bar{x}_{n'} - \bar{x}_{st})$$

where $\bar{y}_{st} = \sum_h W_h \bar{y}_h$ and $\bar{x}_{st} = \sum_h W_h \bar{x}_h$

The $MSE(t_4)$ can be derived as follows,

Let

$$\bar{y}_{st} = \mu_y + \varepsilon_1, \quad \bar{x}_{st} = \mu_x + \varepsilon_2, \quad \bar{x}_{n'} = \mu_x + \varepsilon_3,$$

$$b_{yx} = s_{yx}/s_x^2, \quad s_{yx} = S_{yx} + \varepsilon_4, \quad s_x^2 = S_x^2 + \varepsilon_5$$

where ε_i's are errors such that $E(\varepsilon_i) = 0$ for i = 1,2,3,4 and 5 and

S_x^2, S_y^2 and S_{yx} are population variances and covariance respectively.

Now,

$$MSE\ (t_4) = E\{t_4 - E(t_4)\}^2$$

$$= E\{\bar{y}_{st} + b_{yx}(\bar{x}_{n'} - \bar{x}_{st}) - \mu_y\}^2$$

$$= E\{(\mu_y + \varepsilon_1) + (s_{yx}/s_x^2)(\varepsilon_3 - \varepsilon_2) - \mu_y\}^2$$

$$= E[\varepsilon_1 + \{(S_{yx} + \varepsilon_4)/(S_x^2 + \varepsilon_5)\}(\varepsilon_3 - \varepsilon_2)]^2$$

$$= E[\varepsilon_1 + (S_{yx}/S_x^2)\{1 + (\varepsilon_4/S_{yx})\}\{1 + (\varepsilon_5/S_x^2)\}^{-1}(\varepsilon_3 - \varepsilon_2)]^2$$

Assuming

$\left|\varepsilon_5 / S_x^2\right| < 1$ we get,

$$\text{MSE}(t_4) = E[\varepsilon_1 + (S_{yx}/S_x^2)\{1+(\varepsilon_4/S_{yx})\}\{1-(\varepsilon_5/S_x^2)+.....\}\ (\varepsilon_3-\varepsilon_2)]^2$$

$$= E[\varepsilon_1 + (S_{yx}/S_x^2)\{1+(\varepsilon_4/S_{yx})\}\{\varepsilon_3-(\varepsilon_3\varepsilon_5/S_x^2)-\varepsilon_2+(\varepsilon_2\varepsilon_5/S_x^2)\}]^2$$

$$= E[\varepsilon_1 + (S_{yx}/S_x^2)\{\varepsilon_3-(\varepsilon_3\varepsilon_5/S_x^2)-\varepsilon_2+(\varepsilon_2\varepsilon_5/S_x^2)+(\varepsilon_4\varepsilon_3/S_{yx})-(\varepsilon_2\varepsilon_4/S_{yx})]^2$$

which upto to second order of approximation, gives

$$MSE\ (t_4) = E[\varepsilon_1 + (S_{yx}/S_x^2)(\varepsilon_3-\varepsilon_2)]^2$$

$$= E[\varepsilon_1^2 + (S_{yx}^2/S_x^4)(\varepsilon_3^2-2\varepsilon_3\varepsilon_2+\varepsilon_2^2)+2(S_{yx}/S_x^2)(\varepsilon_1\varepsilon_3-\varepsilon_1\varepsilon_2)]$$

$$= Var\ (\bar{y}_{st}) + (S_{yx}^2/S_x^4)Var\ (\bar{x}_{st}) + (S_{yx}^2/S_x^4)Var\ (\bar{x}_{n'})$$
$$- 2(S_{yx}^2/S_x^4)Cov\ (\bar{x}_{n'},\bar{x}_{st}) + 2(S_{yx}/S_x^2)\{Cov\ (\bar{x}_{n'},\bar{y}_{st}) - Cov\ (\bar{y}_{st},\bar{x}_{st})\}$$

Now, the variance of $\bar{x}_{st}$ and $\bar{y}_{st}$ is given by

$$Var(\bar{y}_{st}) = \sum(W_h^2 S_{y_h}^2/n_h)(1-f_h) \quad Var(\bar{x}_{st}) = \sum(W_h^2 S_{x_h}^2/n_h)(1-f_h)$$
$$= \sum W_h^2 S_{y_h}^2/n_h \qquad = \sum(W_h^2 S_{x_h}^2/n_h)$$

(Cochran, 1977)

(Under the consideration that the sampling fraction n_h/N_h are negligible.)

$S_{y_h}^2$ and $S_{x_h}^2$ are variance of Y and X in the h[th] stratum.

Also the variance of $\bar{x}_{n'}$ is given by

$$Var(\bar{x}_{n'}) = \{(1/n') - (1/N)\}S_x^2$$
$$= S_x^2/n'$$

(Cochran, 1977)

(for large population size N, 1/N is negligible)

The pair (X, Y) is considered to be a bivariate random variable with mean (μ_x, μ_y) and covariance matrix $\sum_{(1,1)}$ in which the variances are denoted by S_x^2 and S_y^2 and the correlation coefficient by ρ. The regression estimator depends on weather the covariance matrix is known or not. If $\sum_{(1,1)}$ is known, we may let $S_x^2 = S_y^2 = 1$ without the loss of generality. The strata population (X_h, Y_h) can also be considered as a bivariate random variable for every h, with mean (μ_{x_h}, μ_{y_h}). If the covariance matrix $\sum_{(X_h, Y_h)}$ of the pair (X_h, Y_h) is known, we can let $S_{x_h}^2 = S_{y_h}^2 = 1$(WLOG).

When the samples are selected with proportional allocation then the strata weights are given by

$$W_h = (N_h / N) = (n_h / n) \qquad \text{(Cochran, 1977)}$$

Thus $\sum_h W_h^2 / n_h = \sum_h W_h^2 / nW_h = (1/n)\sum_h W_h = (1/n)$ (as $\sum_h W_h = 1$)

Hence the covariance matrix of $(\bar{x}_{n'}, \bar{x}_{st}, \bar{y}_{st})$ reduces to

$$\sum = \begin{pmatrix} Var(\bar{x}_{n'}) & Cov(\bar{x}_{n'}, \bar{x}_{st}) & Cov(\bar{x}_{n'}, \bar{y}_{st}) \\ Cov(\bar{x}_{st}, \bar{x}_{n'}) & Var(\bar{x}_{st}) & Cov(\bar{x}_{st}, \bar{y}_{st}) \\ Cov(\bar{y}_{st}, \bar{x}_{n'}) & Cov(\bar{y}_{st}, \bar{x}_{st}) & Var(\bar{y}_{st}) \end{pmatrix}$$

i.e,
$$\sum = \begin{pmatrix} 1/n' & 1/n' & \rho/n' \\ 1/n' & 1/n & \rho/n \\ \rho/n' & \rho/n & 1/n \end{pmatrix}$$

Hence,

$$MSE(t_4) = (1/n) + (\rho^2/n) + (\rho^2/n') - 2(\rho^2/n') + 2\rho\{(\rho/n') - (\rho/n)\}$$

$$= (1/n)(1 - \rho^2) + (1/n')\rho^2$$

APPENDIX 7

Fortran 77 program to evaluate an integral (I_{11}) using Simpson's 1/3rd rule

```
 F(X)=0.398862*((PI+(X/SQRT(200.0)))*(PI+(X/SQRT(200.0))))*
    7EXP(-0.5*X*X)
     PI = -0.1
     DO 10 I=1,11
     B=15
     N=50
     PI=PI+0.1
     Z=2.58
     A=Z-PI*SQRT(200.0)
     PRINT 2,A
2    FORMAT(////"THE VALUE OF IOWER LIMIT A=",F10.4)
     H=(B-A)/N
     X0=A
     XN=B
     SUM=F(X0)+F(XN)
     DO 20 K=1,N-1
       XK=X0+K*H
       IF (MOD(K,2).EQ.0) THEN
         SUM=SUM+2*F(XK)
       ELSE
          SUM=SUM+4*F(XK)
       ENDIF
20   ENDDO
     SUM=(H/3)*SUM
     PRINT 6
6    FORMAT(//1X,"RESULT:",/1X)
     PRINT*,"THE VALUE OF THE INTEGRAL IS:",SUM
10   ENDDO
     STOP
     END
```

APPENDIX 8

Fortran 77 program to evaluate an integral (I_{12}) using Simpson's $1/3^{rd}$ rule

```
F(X)=0.398862*((PI+(X/SQRT(200.0)))*(PI+(X/SQRT(200.0))))*
   7EXP(-0.5*X*X)
    PI =-0.1
    DO 10 I=1,11
    A= -15
    N=50
    PI=PI+0.1
    Z=2.58
    B=-Z-PI*SQRT(120.0)
    PRINT 2,B
2    FORMAT(//"THE VALUE OF IOWER LIMIT B=",F10.4)
    H=(B-A)/N
    X0=A
    XN=B
    SUM=F(X0)+F(XN)
    DO 20 K=1,N-1
      XK=X0+K*H
      IF (MOD(K,2).EQ.0) THEN
        SUM=SUM+2*F(XK)
      ELSE
         SUM=SUM+4*F(XK)
      ENDIF
20   ENDDO
    SUM=(H/3)*SUM
    PRINT 6
6    FORMAT(//1X,"RESULT:",/1X)
    PRINT*,"THE VALUE OF THE INTEGRAL IS:",SUM
10   ENDDO
    STOP
    END
```

APPENDIX 9

Fortran 77 program to evaluate an integral (I_{21}) using Simpson's 1/3rd rule

```
C   PURPOSE
C       INTEGRATES F(U,V),U=A TO B,V=C TO D
C   DESCRIPTION OF PARAMETERS
C       A,C - THE LOWER LIMITS OF INTEGRATION
C       B,D - THE UPPER LIMITS OF INTEGRATION
C       N,M - NUMBERS OF SUB INTERVALS FOR U,V RESPECTIVELY.
C       SIM - ON RETURN IT CONTAINS THE DOUBLE INTEGRATION VALUE.
C   METHOD
C       SIMPSON RULE - DOUBLE INTEGRATION.
    IMPLICIT REAL*8 (A-Z)
    F(U,V)=(MI+(U/(SQRT(120.0))))*(MI+(V/(SQRT(120.0))))*0.38969*
   6EXP(-3*(U*U-2*0.9128*U*V+V*V))
    MI=-0.1
    DO 20 I=1,11
    B=10
    C=-15
    D=15
    N=50
    M=50
    MI=MI+0.1
    Z=1.96
    A=Z-MI*SQRT(120.0)
    PRINT 3,A
```

```
3    FORMAT(2X,"THE VALUE OF A=",F10.4/)
     H1=(B-A)/N
     H2=(D-C)/M
     B=A
     C1=C
     SIMP =0
     DO 2 J =1,N
     A=B
     B=A+H1
     D=C1
     DO 2 K=1,M
     C=D
     D=C+H2
     FUN=((D-C)*(B-A)/36)*(F(A,C)+F(A,D)+F(B,C)+F(B,D)+4
    6*(F(A,(C+D)/2)+F(B,(C+D)/2)+F((A+B)/2,C)+F((A+B)/2,D))
    7+16*(F((A+B)/2,(C+D)/2)))
2    SIMP=SIMP+FUN
     SIM=SIMP
C    SIM CONTAINS THE DOUBLE INTEGRATED VALUE
     PRINT 4,SIM
4    FORMAT(12X,"DOUBLE INTEGRATION VALUE=",F10.6///)
20   ENDDO
     STOP
     END
```

APPENDIX 10

Fortran 77 program to evaluate an integral (I_{22}) using Simpson's 1/3rd rule

```
C    PURPOSE
C        INTEGRATES F(U,V),U=A TO B,V=C TO D
C    DESCRIPTION OF PARAMETERS
C        A,C - THE LOWER LIMITS OF INTEGRATION
C        B,D - THE UPPER LIMITS OF INTEGRATION
C        N,M - NUMBERS OF SUB INTERVALS FOR U,V RESPECTIVELY.
C        SIM - ON RETURN IT CONTAINS THE DOUBLE INTEGRATION VALUE.
C    METHOD
C        SIMPSIN RULE - DOUBLE INTEGRATION.
      IMPLICIT REAL*8 (A-Z)
      F(U,V)=(MI+(U/(SQRT(120.0))))*(MI+(V/(SQRT(120.0))))*0.38969*
     6EXP(-3*(U*U-2*0.9128*U*V+V*V))
      MI=-0.1
      DO 20 I=1,11
      A=-15
      C=-15
      D=15
      N=50
      M=50
      MI=MI+0.1
      Z=1.96
      B=-Z-MI*SQRT(120.0)
```

```
      PRINT 3,B
3     FORMAT(2X,"THE VALUE OF B=",F10.4/)
      H1=(B-A)/N
      H2=(D-C)/M
      B=A
      C1=C
      SIMP =0
      DO 2 J =1,N
      A=B
      B=A+H1
      D=C1
      DO 2 K=1,M
      C=D
      D=C+H2
      FUN=((D-C)*(B-A)/36)*(F(A,C)+F(A,D)+F(B,C)+F(B,D)+4
     6*(F(A,(C+D)/2)+F(B,(C+D)/2)+F((A+B)/2,C)+F((A+B)/2,D))
     7+16*(F((A+B)/2,(C+D)/2)))
2     SIMP=SIMP+FUN
      SIM=SIMP
C     SIM CONTAINS THE DOUBLE INTEGRATED VALUE
      PRINT 4,SIM
4     FORMAT(12X,"DOUBLE INTEGRATION VALUE=",F10.6///)
20    ENDDO
      STOP
      END
```

BIBLIOGRAPHY

Ahmed, M.S. (1998). A note on regression-type estimators using multiple auxiliary information. *Australian and New Zealand Journal of Statistic*, **40**(3), 373 - 376.

Ali, A.M., and Saleh, A.K.Md.E. (1991). Preliminary test and empirical Bayes approach to shrinkage estimation of regression parameters. *Journal of Japan Statistical Society*, **21**(1), 401 - 416.

Bancroft, T. A., and Han, C. P. (1983). A Note on Pooling Variances, *Jour. Amer. Stat. Assoc.*, **78**(384), 981- 983.

Bock,M.E.,Yancey,T.A.,and Judge,G.G.(1973). The statistical consequences of preliminary test estimator in regression, *Jour.Amer.Stat.Assoc.*, **68**(341),109 - 116.

Census of India.(1991). Primary census abstract ,Registrar General of India.2A, Mansing Road, New Delhi.

Chand, L.(1975). Some ratio-type estimators based on two or more auxiliary variables.Unpublished Ph.D. Dissertation. Iowa State University,Iowa.

Chapra, S.C., and Canale, R.P. (1989). *Numerical Methods for Engineering.* Tata Mcgraw Hill. Second edition.

Chernyak, A. (2001). Optimal allocation in stratified and double random sampling with a non linear cost function, *Journal of Mathematical Sciences*, **103**(4), 525 - 528.

Cochran, W.G. (1977). *Sampling Techniques,* Wiley Eastern Ltd. Third edition.

Das, G. (1992). Preliminary test estimators in double sampling with two auxiliary variables, Unpublished Ph.D. thesis, North Eastern Hill University, Shillong, India.

Das, G. (2003). A generalized study of preliminary test estimator in double sampling, *Assam Statistical review*, **17** (2), 127 - 138.

Das, G., and Bez, K. (1995). Preliminary test estimators In double sampling with two auxiliary variables, *Communications in statistics (Theory and Methods)*, **24** (5), 1211 - 1226.

Davidov, O.,and Chang, Y. (2002). Stratified double sampling with continous outcomes: Design and analysis. *Statistic.*, **36**(2), 163 - 173.

Diana, G., and Tommasi, .C.(2004). Optimal use of two auxiliary variables in double sampling, *Statistical Methods and Applications*, **13**, 275 – 284.

Dorfman, A.H. (1994). A note on variance estimation for the regression estimator in double sampling. *Jour.Amer. Stat. Assoc.*, **89**,137 - 140.

Esimai, G.O., and Han, C.P. (1982). Regression estimation with double sample and partial information. *Jour. Ind. Stat. Assoc.*, **20**, 1 – 7.

Grimes, E., and Sukhatme, P.V. (1980). A regression type estimator base on preliminary test of significance. *Jour.Amer. Stat. Assoc.*, **75**, 957 - 962.

Gunst, R. F., and Mason, R. L. (1977). Biased estimation in regression: An evaluation using mean squared error. *Jour.Amer. Stat. Assoc.*, **72**, 616 - 628.

Han, C.P. (1973). Double sampling with partial information on auxiliary variables. *Jour.Amer. Stat. Assoc.*, **68**, 914 - 918.

Han, C.P. (1978). Non negative and preliminary test estimators of variance components. *Jour.Amer. Stat. Assoc.*, **73**, 855 - 858.

Han, C. P.(1990). A Monte carlo comparasion of regression estimators in stratified sampling. Proceedings papers of *American Statistical Assoc*, 209 - 212.

Han, C.P., and Bancroft, T.A. (1968). On pooling means when variance is unknown, *Jour.Amer. Stat. Assoc.*, **63**, 1333 - 1342.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of nonresponse in sample surveys. *Jour.Amer. Stat. Assoc.*, **41**, 517 - 529.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*. vol. 1 and 2, NewYork, Wiley.

Hidiroglou, M.A. and Sarndal, C.E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, **24**(1), 11 - 20.

Isaki,C.T. (1983). Variance estimation using auxiliary information. *Jour.Amer. Stat. Assoc.*,**78**, 117 - 123.

Jain,M.K., Iyengar,S.R.K., and Jain, R.K.,(2007). *Numerical Methods for Scientific and Engineering Computating*, New Age International, Fifth Edition.

Jhajj, H. S., Sharma, M. K., and Grover. L. K. (2006). A families of estimators of population mean using information on auxiliary attribute. *Pak. J. Statist.*, **22**(1), 43 – 50.

Johnson,J.P., Bancroft,T.A., and Han,C.P.(1977). A pooling methodology for regression in prediction. *Biometrics,* **33**(1), 57 – 67.

Khare, B.B., and Srivastava, S.R. (1980). On an efficient estimator of population mean using two auxiliary variables. *Proc. Nat. Acad. Sci. India,* **50***(A), IV,* 209 – 214.

Kibria, B,M,G. (1996). On preliminary test ridge regression estimators for linear restrictions in a regression model with non-normal disturbances. *Communications in Statistics (Theory and Methods)*, **25** (10), 2349 - 2369.

Kibria, B. M. G., and Saleh A.K. Md. E. (2004). Preliminary test ridge estimators with student's *t* errors and conflicting test-statistics. *Metrika*, **59**, 105 -124.

Kiregyera, B. (1980). A chain ratio type estimator in finite population double sampling using two auxiliary variables. *Metrika,* **27**, 217 - 223.

Kiregyera, B.(1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. *Metrika,* **31**, 215 - 226.

Mohanty, S. (1967). Combination of ratio and regression estimate. *Jour. Ind. Stat. Assoc.*, **5**, 16 - 19.

Mukerjee, R., and Chaudhuri, A. (1990). Asymtotic optimality of double sampling plans employing generalized regression estimators. *Journal of Statistical Planning and Inference*, **26**, 173 - 183.

Mukerjee, R., Rao, T.J., and Vijayan,V.(1987). Regression type estimators using multiple auxiliary information. *Australian Journal of Statistics,* **29**, 244 - 254.

Murthy, M.N. (1967). *Sampling theory and methods.* Statistical Publishing Society, Calcutta.

Neyman, J.(1938). Contribution to the theory of sampling human populations, *Jour.Amer. Stat. Assoc.*, **33**, 101 - 116.

Okafor, F. C., and Lee, H. (2000). Double sampling for ratio and regression estimation with sub sampling of non respondent. *Survey Methodology*, **26**(2), 183 - 188.

Pandey, B.N., and Singh, J. (1977). Estimation of variance of normal population using prior information. *Jour. Ind. Stat. Assoc.*, **15**, 141 - 150.

Pradhan, B.K. (2005). A chain regression estimator in two phase sampling using multi auxiliary information. *Bull. Malays.Math. Sci. Soc.* (2) **28**(1),81 - 86

Quenouille, M.H.(1956). Notes on Bias estimation. *Biometrika,* **43**, 353 - 360.

Raj, D.(1965). On Sampling over two occasions with probability proportional to size. *Ann. Math .Stat.,* **36**, 327 - 330.

Rajaraman,V.(1997). *Computer programming in Fortran 77*, PHI, New Delhi, Fourth Edition.

Reich,R.M., Bonham,C.D., and Kimberly,K. (1993). Technical notes; Double sampling revisited. *Journal of Range Management,* **46**(1), 88 - 90.

Reproductive and Child Health Report(2000): Rapid Household Survey. Phase I,1998., International Institute for Population Sciences (IIPS), Mumbai.

Roy, D.C. (2003). A regression type estimator in two phase sampling using two auxiliary variables, *Pak. J. Statist.*, **19** (3) , 281 - 290.

Sahoo, J., Sahoo, L.N., and Mohanty,S.(1993). A regression approach to estimation in two phase sampling using two auxiliary variables. *Current Science*, **65**(1), 73 - 75.

Sahoo, J., and Sahoo, L.N. (1999). An alternative class of estimators in double sampling procedures. *Calcutta Statistical Association Bulletin,* **49**, 79 - 83.

Sahoo, L.N., and Swain, A.K.P.C. (1989). On two modified ratio estimators in two phase sampling. *Metron,* **47**, 261 - 266.

Saleh, A.K.Md.E. and Han, C,P. (1990). Shrinkage estimation in regression analysis. *Estadistica.,* **42**, 40 - 63.

Saleh A. K. Md. E., and Kibria, B. M. G. (1993). Performance of some new preliminary test ridge regression estimators and their properties. *Communications in Statistics (Theory and Methods),* **22**(10), 2747 - 2764.

Samiuddin, M., and Hanif, M. (2007). Estimation of the population mean in single and two phase sampling with or without additional information. *Pak. J. Statist.*, **23**(2), 99 - 118.

Sarndal, C.E., and Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and non response. *Int. Statist. Rev.,* **55**(3), 279 - 294.

Senapati ,S,C., and Sahoo,L,N. (2006). An alternative class of estimators in double sampling. *Bull. Malays. Math. Sci. Soc.* (2) **29**(1) , 89 - 94.

Shabbir, J., and Gupta, S. (2007). On estimating the finite population mean with known population proportion of auxiliary variable.*Pakistan. J. Statist.*, **23**(1),1 -9.

Singh, A.K. and Singh, H.P. (1997). A Note on the efficiencies of three product type estimators under a linear model. *Jour.Ind.Soc.Agric.Stat.,***50** (2), 30 - 34.

Singh,H.P. and Kakran,M.S.(1993). A modified ratio estimator using known coefficient of kurtosis of an auxiliary character. Revised version submitted to *Jour.Ind.Soc.Agric.Stat* ., New Delhi. India.

Singh, H.P., Katyar, N.P., and Gangwar, D.K. (1996). A class of almost unbiased regression-type estimators in two phase sampling applying Quenouille's method. *Jour. Ind. Soc. Agric. Stat.,* **48**, 98 - 104.

Singh, P., and Srivastava, A. K. (1980). Sampling scheme providing unbiased regression estimator, *Biometrika*, **67**(1), 205 - 209.

Singh, S. (2006). Surveyed statisticians celebrate golden jubilee year 2003 of the linear regression estimator. *Metrika*, **63**, 1 - 18.

Sisodia, B.V.S. and Dwivedi,V.K. (1981). A modified ratio estimator using coefficient of variation of auxiliary variable. *Jour. Ind. Soc. Agric. Stat.,* **33**, 13 - 18.

Sisodia, B.V.S., and Srivastava, A.K.(1982). Modifying regression estimators with a preliminary test in double sampling, *Sankhya* B**44**(3), 295 – 303.

Sisodia, B.V.S.(1981). On the use of preliminary test of significance in repeated surveys. *Jour.Ind. Stat. Assoc.*, **1**, 47 - 68.

Sitter,R.R. (1997). Variance estimation for the regression estimator in two phase sampling. *Jour. Amer. Stat. Assoc.*, **92**, 780 - 787.

Srivastava, S.K. (1971). A generalized estimator for the mean of a finite population using multi auxiliary information. *Jour.Amer. Stat. Assoc.*, **66,** 404 - 407.

Srivastava, S.K. (1980). A class of estimators using auxiliary information in sample surveys. *Canadian .Jour. of Statistics,* **8**(2), 253 - 254.

Srivastava, S.K. (1981). A generalized two-phase sampling estimator. *Jour. Ind. Soc. Agric. Stat.,* **13** (1), 38 - 46.

Srivastava, S.K., and Jhajj, H,S.(1981). A class of estimators of the population mean in survey sampling using auxiliary information., *Biometrika*, **68**(1), 341 - 343.

Srivenkataramana, T. and Tracy, D.S. (1989). Two phase sampling for selection with probability proportional to size in sample surveys. *Biometika,* **76** (4), 818 - 821.

Stata Statistical Software (2003) Release 8.0. College Station, Texas, Stata Corporation.

Sukhatme,P,V., Sukhatme,B,V., Sukhatme,S., and Asok,C. (1984). *Sampling theory of surveys with applications.* Iowa state university press and IASRI. Third edition.

Sukhatme, P.V., and Tang, K.T. (1975). Allocation in stratified sample sub - subsequent to preliminary test of significance. *Jour.Amer. Stat. Assoc.*,**70**, 175 - 179.

Tamhane, A,C., (1978). Inference based on regression estimator in double sampling. *Biometrika*, **65**(2), 419 - 427.

Tripathi, T.P., and Ahmed, M.S.(1995). A class of estimators for a finite population mean based on multivariate information and general two phase sampling. *Calcutta Statist. Assoc. Bull.,* **45**, 203 - 218.

Upadhayaya, L.N., and Singh, H.P. (1999). Use of transformed auxiliary variable in estimating the finite population mean, *Biometrical Journal,* **41**(5), 627 - 636.

Williams,W.H. (1963). The precision of some unbiased regression estimators. *Biometrics*, **19**(2), 352 - 361.

Zhang, B. (1999). Bootstrapping with auxiliary information. *Canadian Jour. of Statistics.*, **27**(2), 237 - 249.

BIO – DATA

Name	Mr. Phrangstone Khongji
Date of Birth	22nd December 1973.
Nationality	Indian
Permanent Residence	Shillong, Meghalaya, India.
Present Affiliation	Assistant. Professor in Statistics School of Technology North Eastern Hill University Shillong, Meghalaya.

Academic Qualification

1. M.Sc, Mathematics,1998, North Eastern Hill University Shillong, Meghalaya.

2. Master in Population Stuidies, 2001, International institute for Population Sciences, Mumbai.

Teaching Experience

1. B. Sc. Mathematics Pass and Major courses , 2001 – 2006, Kiang Nangbah Government College, Jowai, Meghalaya.

2. B.Tech, Statistics, 2006 onwards, School of Technology, North Eastern Hill University, Shillong.

Workshops/Conferences

1. Visitor : Tata Institute of Fundamental Research, Mumbai,India, July – Aug,1998.

2. Workshop attended on *Relational Database Management System (RDBMS).* jointly organized by Indian Statistical Institute, Kolkata and St. Anthony's College, Shillong, March, 2005.

3. Workshop attended on *Neural Network and Genetic Algorithm* organised by the Department of Statistics, St. Anthony's College, Shillong. March, 2007.

4. Workshop attended on *Statistical Methods in Medical and Health Statistics* organised by the Department of Statistics, NEHU, February, 2009.