2500/-

NON-PARAMETRIC REGRESSION TO FINITE POPULATION ESTIMATION.

BY

TOBIAS MBITHI MWALILI

SIGNATURE _____. 17-7-97

SUPERVISOR.

DR. ROMANUS ODHIAMBO OTIENO

SIGNATURE _____. 17-7-97

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTERS OF SCIENCE (STATISTICS) AT KENYATTA UNIVERSITY.

JULY, 1997.

DEDICATED TO MY SON TEDDY MWALILI. MAY YOU GROW TO APPRECIATE THE VIRTUE OF HARD WORK

ACKNOWLEDGMENTS.

I would like to register my sincere appreciation to all those who assisted me in any way to do this piece of work.

I am grateful to my supervisor Dr. Romanus Odhiambo. It is through your patience, advice and reliability that this piece of work came to be. I am greately indebted to Dr. Wafula, you really motivated me to the field of statistics and research.

I would also like to thank Kenyatta University for granting me the opportunity to undertake my studies. To my course mates, Njeri, Mageto and Nderitu I appreciated your company and cooperation.

Most earnestly, I thank my father for his tireless efforts to ensure that I completed my studies. To my sisters and brother I say thank you for your material and moral support. Mother, your constant prayers and encouragement's are highly appreciated.

TABLE OF CONTENTS

CHAPTER FOUR

CHAPTER FIVE

CHAPTER SIX

ABSTRACT

Nonparametric regression is used here to estimate the finite population mean. The variance of the derived estimate is obtained and procedures for it's estimation suggested. The appropriateness of the variance estimators is established by derivation of their mean square error, which is shown to diminish with rise in sample size. Empirical study is performed using real and simulated data, and the outcomes support theoretical findings.

CHAPTER ONE

1.1 INTRODUCTION

The purpose of Design and analysis of sample surveys is to obtain information about finite Populations. In trying to obtain this information, there are two options that the sample surveyor can take. One is complete enumeration or census . In this method, each and every unit of population is observed .The above method is time consuming, expensive and unrealistic especially when dealing with large populations . The surveyor then opts for the second method of sampling. This involves observing part of the population called sample, then making inference based on the sample about the entire population. The theory of the sample survey aims at developing sampling strategies that result in the selection of a sample that is a 'good' representation of the whole population . It also provides methods for making inference about the characteristic of interest and finding criterion for comparing different strategies in order to obtain optimal results from a sample survey.

1.2 DEFINITION OF TERMS

1.2.1 FINITE POPULATION

This is a collection of N units where $N < \infty$ is the size of the population. In sample surveys N is the usually known. For instance, N could be the number of schools in a country , or the number of towns in that country.

1.2.2 SAMPLING UNITS

These are subdivisions or subsets of the whole population such that

$Ui \cap U_{i \neq j} = \varnothing, \Rightarrow \bigcup_{i=1}^{n} U_i = U$ where U is acollection of N units U_i's .

1.2.3 IDENTIFIABLE UNITS.

Units of a finite population are said to be identifiable if they can be uniquely labelled from 1 to N and the label of each unit is unique .*i.e.* There is a 1:1 correspondence between the units and the indices *1,2, ...,N* such that the population comprises of $U = \{u_1, u_2, u_3, ..., u_N\}$.

1.2.4 FRAME.

A clear and concise listing of all the sampling units by which population units can be identified unambiguously .Some times the formation of a frame is not feasible in some populations, for example fish populations.

1.2.5 CHARACTERISTICS OF INTEREST.

When carrying out survey sampling, there is a particular characteristic we are interested in. This could be the population total, mean wage earnings, or even the ratio of domestic income to expenditure in education. Thus, associated with each U_i is the characteristic of interest Y_i , $i=1,2,3,.....N$.

1.2.6 AUXILIARY INFORMATION.

Associated to each unity of $\underline{U} = (u_1, u_2, u_3, ...u_N)$ is a certain known characteristics vector $\underline{X} = (x_1, x_2, x_3, ...x_N)$ referred to as auxiliary information and is positively correlated to $\underline{Y} = (y_1, y_2, y_3, ...y_N)$. The auxiliary information is known before hand .The technique is to use the obtained sample, plus the auxiliary information to make inference about $\underline{Y} = (y_1, y_2, y_3, ...y_N)$ or afunction of Y_i 's .X_i's could be previous values of Y_i's when a complete census was done or X_i's could be any variable that is positively correlated with Y_i 's .

There could be also exist q characteristics positively correlated to Y, giving us an Nxq matrix of co-variates; $X_{n\times q}=((x_{ij}))_{n\times q}$.

1.2.7 SAMPLE

This is an ordered collection of units from $\underline{U} = (u_1, u_2, u_3, ...u_N)$ such that $S = (u_{1s}, u_{2s}, u_{3s}, ...u_{ns})$, Where $n \leq N$ is the sample size and $u_{(is)}$ denotes the i^{th} units in sample S.

1.3 SELECTION OF THE SAMPLE

Let S denote the set of all possible samples from a finite population U. Let also $P(s)$ denote the probability that a sample s is drawn. Then a probability sampling design assigns to each s$\in$ S a probability p(s)$\geq$ 0 such that $\sum_{S} p(s) = 1$.

1.4 SAMPLE SURVEY ESTIMATION PROBLEM

Essentially, there are two estimation problems that a sample surveyor seeks to solve. (I) Estimation of some well defined descriptive functions of $\underline{Y} = (y_1, y_2, y_3, \ldots\ldots y_N)^T$ the so called 'parameters' of the finite population. The common functions of interest are

(1) The finite population total, $T = \sum_{i=1}^{N} y_i$

(2) The finite population mean, $\bar{T} = \frac{\sum_{i=1}^{N} y_i}{N}$

(3) The finite population variance $V(y) = \frac{\sum (y_i - \bar{y})^2}{N}$

(II) To carry out an analytic inference about the internal structure of the data. For example, given the linear regression model $y = \alpha + \beta x_i$ the problem would be to estimate α and β .

1.5 APPROACHES TO SAMPLE SURVEY ESTIMATION PROBLEM

Basically, there are at least two approaches to sample survey estimation problem namely;

(I) The classical approach, otherwise referred to as the randomization approach.

(II)The prediction approach, otherwise referred to as the Super population approach.

There are fundamental differences in the way these two approaches view the population units. In the next section we shall outline these differences.

1.5.1 THE CLASSICAL APPROACH

In this approach, each population unit u_i $i=1,2,3,.....N$ is associated with a fixed but unknown real number which is the value of the variable under study . The sample chosen is a probability sample based on a certain sampling design d[s,p(s)].

Inference is thus based on the observed quantities $y_1, y_2, y_3, \ldots\ldots\ldots\ldots, y_n$

which were initially chosen to the sample through the design ***d[s,p(s)]***. The implication here is that inference will be tied to the chosen design d. That why it is commonly referred to as design based.

1.5.2 THE PREDICTION APPROACH.

Remark I

The set up in the classical approach assumes that the population units are labelled and that the statistician has access to the true and fixed value for the y_i of the i^{th} unit. This is both too demanding and unjustifiable. It is through this shortfall that the prediction approach came in to being.

In this approach, a superpopulation model is inherent in any given finite population. The model employed characterize the actual population values , both the observed and unobserved which are considered as a realization of random variables $Y_1, Y_2, Y_3 \ldots\ldots\ldots\ldots, Y_N$. The relationship among the variables is expressed as a model of the joint distribution of the random variables $Y_1, Y_2, Y_3 \ldots\ldots\ldots\ldots, Y_N$. We now give an example to illustrate this.

Example 1

Let $\hat{\bar{T}}(y)$ be the ratio estimator of the population mean. i.e.

$\hat{\bar{T}}(y) = \bar{y}_R = \frac{\bar{y}}{\bar{x}}\bar{X}$ where $\bar{x}, \bar{y}$ and $\bar{X}, \bar{Y}$ are the sample and population means respectively. A super population model traditionally associated with the ratio estimator is

$E(y_i)=\beta x_i$

$Var(y_i)=\sigma^2 x_i \qquad (1.1)$

$Cov(y_i, y_j)=0 \quad i \neq j$

Due to the above outlined differences in the approaches, the way in which an estimator is defined differ as well. For this reason we shall study Estimation in each approach separately.

1.6 CLASSICAL APPROACH: ESTIMATORS, UNBIASEDNESS, VARIANCE AND MEAN SQUARE ERROR.

1.6.1 ESTIMATORS

In this approach, an estimator $\hat{T}(y)/S$ is seen as a real valued function defined on $S \times R^N$ where $s \in S$ depends on Y only through y_i's for which unit *i* occurs in sample S. We thus calculate the proposed estimator $\hat{T}(y)/S$ based on the observed quantities $y_{1s}, y_{2s}, y_{3s}, \ldots\ldots\ldots\ldots, y_{ns}$. As an illustration consider the example below.

Example 2

The estimators of the population mean $\overline{Y}$ and total *Y* are given by

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} \text{ and}$$

$N\bar{y} = N\frac{\sum_{i=1}^{n} y_i}{n}$ respectively. The key point here is that the estimators are basically functions of the y_i's in S.

1.6.2 UNBIASEDNESS OF $\hat{T}(y)$

An estimator $\hat{T}(y)$ based on design P is said to be design unbiased for T(y) if,

$$E_p[T(y)/S] = \sum_{s \in S} T(y)P(s) = T(y)$$

where $E_p[T(y)/S]$ denotes the conditional expectation of $\hat{T}(y)$ given that sample S is chosen through design P.

Theorem 1

The estimator $\hat{T}(y) = \sum_{s \in S} l_{si} y_i$ for the population total T is unbiased for T.

Where, $\hat{T}(y) = \sum_{i \in s} l_{si} p(s) = 1, i \le l \le N$ and l_{si} depends on S and the i^{th} unit. (B.K Sinha 1991,Sec 2.4).

Proof

$$E_p[\hat{T}(y)/S] = \sum_{s \in S} T(y)P(s)$$

$$= \sum_{s \in S} P(s)\left[\sum_{i \in s} l_{is} y_i\right]$$

$$= \sum_{s \in S} y_i\left[\sum_{i \in s} l_{is} P(s)\right]$$

But $\sum_{i \in s} l_{is} P(s) = 1 \Rightarrow E_p[\hat{T}(y)/S] = \sum_{i=1}^{N} Y_i = T(y)$ hence $\hat{T}(y)$ is unbiased for T(y).

In situations where $\hat{T}(y)$ is biased we have

$E_p[\hat{T}(y)/S] = T(Y) + Bias - term$ which we can write as

$$E_p[\hat{T}(y)/S] = T(Y) + B_p[\hat{T}(Y)]$$

as such the Bias will be given by $B_p[\hat{T}(Y)] = E_p[\hat{T}(y)/S] - T(Y)$

1.6.3 THE VARIANCE AND MSE OF $\hat{T}(y)$

The variance of $\hat{T}(y)$ is given by

$$Var[\hat{T}(y)] = E_p\left[\hat{T}(y) - E_p[\hat{T}(y)/S]\right]^2 \qquad (1.2)$$

If the estimator is not unbiased then its mean square error is given by

$$MSE_p[T(y)] = E_p[\hat{T}(y) - T(y)]^2$$

$$= Var_p[\hat{T}(y)] + [B_p T(y)]^2 \qquad (1.3)$$

1.6.4 THE CRITERION FOR COMPARING COMPETING STRATEGIES

In the classical approach, the pair $[p, \hat{T}(y)]$ denotes a sampling strategy, where $\hat{T}(y)$ depends on P. The performance of any strategy $[p, \hat{T}(y)]$ is judged through the minimization of (1.2) and (1.3) above. The problem that arises immediately here is, if we opt to use the minimization criterion, then we shall have difficulties in choosing between an unbiased estimator with a small variance and a biased estimator with a small mean square error.

1.7 THE PREDICTION APPROACH : ESTIMATORS, UNBIASEDNESS, VARIANCE AND MEAN SQUARE ERROR.

1.7.1 ESTIMATORS

An estimator $\hat{T}(\underline{y})$ is said to be model unbiased for $T(Y)$ if

$$E_M[T(\underline{y}) / S,\underline{Y}] = E_M[T(\underline{Y})]$$

where $E_M[T(\underline{y}) / S,\underline{Y}]$ denotes the conditional expectation of $\hat{T}(\underline{y})$ given sample $(S,\underline{Y})$ with respect to a given model.

If $\hat{T}(\underline{y})$ is biased then the bias is given by

$$B_M[\hat{T}(\underline{Y})] = E_M[\hat{T}(\underline{Y}) - T(\underline{Y})]$$

1.7.2 THE VARIANCE OF AND MSE OF $\hat{T}(\underline{y})$

Under the prediction approach, variance and mean square error of $\hat{T}(\underline{y})$ is given by $Var_M[\hat{T}(\underline{Y}) / S,\underline{Y}] = E_M[\hat{T}(\underline{y}) - E_M\left[[\hat{T}(\underline{Y}) / S] / S,\underline{Y}\right]$ and

$$MSE_M[\hat{T}(\underline{Y}) / S,Y] = Var_M[\hat{T}(\underline{Y}) / S,\underline{Y}] + B_M[\hat{T}(\underline{Y})]^2$$

$\Rightarrow$ if the Bias is zero then

$$MSE_M[\hat{T}(\underline{Y}) / S,\underline{Y}] = Var_M[\hat{T}(\underline{Y}) / S,\underline{Y}].$$

1.8 TOWARDS A COMPROMISE

In the above section, we have reviewed the measures of uncertainty and how these differ depending on the approach employed . The question that arises is, which of the two approaches is superior in leading to the best strategies?.

Sinha and Hedayat (1991) feel that the classical approach does not lead to any definite optimal strategies. They advocate for use of superpopulation models as a means of reaching conclusive results when comparing competing strategies and eventually producing the most efficient strategies.

However it must be noted that these two approaches are not necessarily opposing . In fact, a blend of the two can be used to produce optimal strategies . Godambe and Thompson (1973) suggested the quantity

$$E_M E_P[\hat{T}(\underline{Y}) - T(\underline{Y})]^2 . \qquad (1.4)$$

Which could be used for minimization purposes in our search for optimal strategies . Expanding the (1.3) above we have

$$E_P E_M\left[\hat{T}(\underline{Y}) - T(\underline{Y})\right]^2 = E_P E_M\left[\hat{T}(\underline{Y}) - E_M\hat{T}(\underline{Y}) + E_M\hat{T}(\underline{Y}) - T(\underline{Y})\right]^2$$

Remark

E_p and E_m can be interchanged since P does not depend on y_i s .

Equation (1.4) now becomes

$$E_p\left[E_M\left(\hat{T}(\underline{Y}) - E_M\hat{T}(\underline{Y})\right)^2 + E_M\left(\hat{T}(\underline{Y}) - T(\underline{Y})\right)^2\right]$$

$$= E_P[Var_M\hat{T}(\underline{Y})] + E_P[B_M T(\underline{Y})]^2 \qquad (1.5)$$

(1.5)

Notice that (1.5) provides a measure of uncertainty based both on model and design based approaches . This makes a good expression for searching optimal strategies.

In the light of Remark I in section 1.5.2 we will henceforth adopt the prediction approach in this study .For now ,we review parametric methods of estimation in the next chapter.

CHAPTER TWO

2.0 PARAMETRIC ESTIMATION OF T :THE POPULATION TOTAL.

In the parametric estimation of T we shall consider the super population model;

$$E(y_i) = \beta x_i$$

$$Var(y_i) = \sigma^2 x_i \qquad (2.1)$$

$Cov(y_i, y_j) = 0$ for $y_i \neq y_j$ which is traditionally associated with the ratio estimator (Cochran 1953, sec 68).

Taking Expectation under model (2.1), we get,

$$E_M(T) = \sum E(y_i)$$

$$= \sum \beta x_i$$

Theorem 2

The ratio estimator $\hat{T}_R$ is unbiased for T under model (2.1).

Proof

The ratio Estimator is given by

$$\hat{T}_R = \left[\sum_{i=1}^{N} X_i \frac{\sum_s y_i}{\sum_s x_i} \right]$$

taking expectation under model (2.1) we have,

$$E_M\left[\hat{T}_R\right] = E_M\left[\sum_{i=1}^{N} X_i \frac{\sum_s y_i}{\sum_s x_i}\right]$$

which can be written as

$$= \left[\sum_{i=1}^{N} X_i \beta \frac{\sum_s x_i}{\sum_s x_i}\right] = \sum_{i=1}^{N} X_i \beta = E_M(T)$$

Hence $\hat{T}_R$ is unbiased for T under Model (2.1). However, if the model is misspecified, the estimator becomes biased. To illustrate this we consider an example below.

Example 3

Consider T under model

$$E(y) = \alpha + \beta x_i$$

$$Var(y_i) = \sigma^2 x_i \qquad (2.2)$$

$$Cov(y_i, y_j) = 0 \text{ for } i \neq j$$

Now we know that , $\bar{Y} = \frac{\sum_{i=1}^{i=N} Y_i}{N}$

So that,

$$E_M(\bar{Y}) = \frac{\sum_{i=1}^{N} y_i}{N}$$

$$= \frac{1}{N}\sum(\alpha + \beta x_i)$$

$$= \alpha + \beta\bar{x}$$

Next, $E(\hat{\bar{Y}}) = E\frac{\bar{y}}{\bar{x}}\bar{X}$

$$= \frac{\bar{X}}{\bar{x}} E \frac{\sum_s y_i}{n}$$

$$= \frac{\bar{X}}{n\bar{x}} \sum_s E(y_i)$$

$$= \frac{\bar{X}}{n\bar{x}} \sum_s (\alpha + \beta x_i)$$

$$= \frac{\bar{X}}{n\bar{x}} (n\alpha + \beta \sum_s x_i)$$

$$= \frac{\bar{X}\alpha}{n\bar{x}} + \beta\bar{x} \qquad (2.3)$$

Note: (2.3) above implies that $\hat{\bar{T}}_R$ is biased under model (2.1) and the bias is given by

$$B_M = E_M\left[\hat{T}(y) - T(y)\right]$$

$$= \bar{X}\alpha + \beta\bar{X} - \alpha - \beta\bar{X}$$

$$= \alpha\frac{(\bar{X} - \bar{x})}{\bar{x}}$$

this bias tends to zero when the sample is balanced on x *i.e.* when $\bar{X} = \bar{x}$ (Royall and Herdson 1973a). We stated earlier that, the accuracy of an estimate is measured through the mean square error and its variance.

For the ratio estimator of the Population mean; there are several methods that have been used to estimate its variance, using the prediction approach. In the next section, we intend to review these methods.

2.1 THE RATIO ESTIMATOR OF THE TOTAL AND ESTIMATORS OF ITS ERROR VARIANCE.

Now the true total of the survey measuremaent is given by;

$$T = \sum_s y_i + \sum_r y_i \quad \text{Where } s \cap r = \varnothing \text{ and } s \cup r = N.$$

Using the prediction approach the estimate for total is

$$\hat{T}_R = \sum_s y_i + \sum_r \hat{\beta} x_i \ , \qquad (2.3)$$

where $\hat{\beta} = \dfrac{\sum_s y_i}{\sum_s x_i}$. So that (2.3) can be written as

$$\hat{T}_R = \sum_s y_i + \frac{\sum_s y_i}{\sum_s x_i} \sum_r x_i$$

$$= \sum_s y_i \left[\frac{\sum_s x_i + \sum_r x_i}{\sum_s x_i} \right] = \sum_s y_i \left[\frac{\sum_N x_i}{\sum_s x_i} \right]$$

The difference between $\hat{T}_R$ and T is the prediction error and will in this case be given by,

$$(\hat{T}-T)=\sum_s y_i\left[\frac{\sum_{i=1}^{N}x_i-\sum_s x_i}{\sum_s x_i}\right]-\sum_r y_i$$

$$=\sum_s y_i\left[\frac{\sum_r x_i}{\sum_s x_i}\right]-\sum_r y_i. \qquad (2.4)$$

Next we obtain the error variance of (2.4) i.e.

$$Var_M(\hat{T}-T)=\left[\frac{\sum_r x_i}{\sum_s x_i}\right]^2 Var\left(\sum_s y_i\right)+Var\left(\sum_r y_i\right)$$

$$=\left[\frac{\sum_r x_i}{\sum_s x_i}\right]^2 \sigma^2\left(\sum_s x_i\right)+\sigma^2\left(\sum_r x_i\right)$$

$$=\sigma^2\left\{\left[\frac{(N-n)\bar{x}_r}{n\bar{x}_s}\right]^2 n\bar{x}_s+(N-n)\bar{x}_r\right\}$$

$$=(N-n)\bar{x}_r\sigma^2\left[\frac{(N-n)\bar{x}_r+n\bar{x}_s}{n\bar{x}_s}\right]$$

$$=\frac{N^2\left(1-n/N\right)}{n\bar{x}_s}\bar{X}\bar{x}_r\sigma^2,$$

Which implies that $Var_M(\hat{\bar{T}}-\bar{T})=\frac{1}{N^2}Var(\hat{T}-T)=\left(\frac{1-f}{n\bar{x}_s}\right)\bar{X}\bar{x}_r\sigma^2.$ (2.5)

Where $f=n/N$

Since $\hat{T}_R$ is unbiased for T, its mean square error will be equal to its variance *i.e.*.

$$M.S.E\left(\hat{\bar{T}}\right) = Var\left(\hat{\bar{T}}\right).$$

2.1.1 CONSISTENCY OF $\hat{T}_R$.

Considering expression (2.5) *i.e* .

$Var_M(\hat{T}-T) = \left(\frac{1-f}{n\bar{x}_s}\right)\bar{X}\bar{x}_r\sigma^2$, it is easy to see that $Var(\hat{\bar{T}}) \xrightarrow{P} 0$ as long as $n \to \infty, N \to \infty, f = {}^{n}/_{N} \to 0$ and $\bar{X}, \bar{x}_s, \bar{x}_r \le b < \infty$ *i.e.* if we are dealing with a stable population. Hence we conclude that $\hat{T}_R$ is a consistent estimator of *T*.

2.2 ESTIMATORS OF OF THE ERROR VARIANCE: A REVIEW.

Here we review several estimator of variance due to Royall and Cumberland [1978,1981] .

Substituting σ^2 in expression (2.5) with its estimate from weighted least squares gives a popular variance expression,

$$V_L = \left(\frac{1-f}{n\bar{x}_s}\right)\bar{X}\bar{x}_r \frac{1}{n-1}\sum_s\left(\frac{y_i - \hat{\beta}x_i}{\sqrt{x_i}}\right)^2.$$ this statistic is unbiased under model (2.1) but can be badly biased if the model fails; that is if $Var(y_i)$ is not proportional to x_i .(Royal and Eberhardt 1975).

Another estimator suggested by Royall and Cumberland (1978) through direct substitution of σ^2 with the squared residuals is

$$V_D = \left(\frac{1-f}{n^2}\right)\frac{\overline{X}\overline{x}_r}{\overline{x}_s^2}\sum_{i=1}^{n}\left(\frac{e_i^2}{1-k_i}\right)$$

where $e_i=(y_i-\hat{\beta}x_i)^2$ and $k_i = \dfrac{x_i}{\sum_s x_i}$.

This estimator is unbiased under model (2.1) , approximately unbiased for more general variance models (Royal and Cumberland 1978a).

A statistic given in most text books for use with simple random sampling is $V_c = \dfrac{N}{n}(N-n)\sum\dfrac{(y_i-\hat{\beta}x_i)^2}{n-1}$. This can have a serious bias under model (2.1) (Royall and Cumberland 1981). Royall and Emberhart (1 975) adjusted V_C to remove its bias under model (2.1) in a balanced sample obtaining,

$$V_H = \frac{V_c\left(\dfrac{\overline{x}_r\overline{X}}{\overline{x}_s^2}\right)}{\left(1-\dfrac{V_s^2}{n}\right)}$$

where $V_s^2 = \dfrac{1}{n-1}\sum_s\dfrac{(x_i-\overline{x}_s)^2}{\overline{x}_s^2}$.

In addition to the above estimators another estimator that has been derived using the chews procedure i.e. method of moments is

$$V_{CH} = \frac{N-n}{N}\frac{\overline{X}\overline{x}_r}{n^2\overline{x}_s + D_s}\sum_s\frac{e_i^s}{\left(1-g_i/n\right)}$$ where $g_i = 2x_i/\overline{x}_s$ and

$$D_s = \left(\frac{1}{n\bar{x}_s^2}\right)\sum_{i=1}^{n}\frac{x_i^2}{1-g_i/n}$$

Clearly for V_{CH} to be non-negative $2x_i \le n\bar{x}$ for all i in the sample, this condition will not always be satisfied .This is the greatest impediment to the use of this estimator . For this reason we shall not consider it in the subsequent sections.

Another that is worth mentioning is the jackknife variance estimator due to (Turkey 1958, Jones1962) given by the expression

$V_J = N(N-n)(n-1)\bar{x}^2\sum_s \frac{D_{(j)}^2}{n}$ where, for every j in s, $D_{(j)}$ is the difference between the ratio $(n\bar{y}_s - y_j)/(n\bar{x}_s - x_j)$ and the average of these n ratios.

Royall and Eberhard (1975) showed that under mild conditions, V_H, V_D, V_J are asymptotically equivalent i.e. $V_H = V_j\{1 + 0(1)\}$.Wu and Deng(1984) showed that V_j is stochastically larger than V_D While V_D is larger than V_H.

REMARKS

We have shown that estimators based on Parametric procedures are unbiased only under the specified model. They however become biased once the model conditions are violated. This points at robustness problem characteristic to parametric approach to Estimation. This is the major motivation to this study. Henceforth in this study we adopt a Nonparametric approach as a means of obtaining robust estimators.

CHAPTER THREE

3.0 NON PARAMETRIC ESTIMATION OF THE FINITE POPULATION TOTALS.

3.1 INTRODUCTION

In this chapter we shall consider the nonparametric procedures in estimation of population mean. As a motivation to this , recall that in chapter 2 we considered the parametric approach under the model ;

$$E(y_i) = \beta x_i$$

$$Var(y_i) = \sigma^2 x_i$$

$$Cov\ (y_i, y_j) = 0 \quad i \neq j$$

Here we showed that the estimators obtained using parametric procedure are unbiased only if the model assumptions are met. However, the estimators become biased unbiased when the model is misspecified or the the variance condition is violated. This points at robustness problem. Many practising statisticians are not comfortable with this approach due to uncertainties in the choice of the model.

For this reason, a new estimator based on Non-parametric regression is suggested. Here, we weaken the assumptions conserning Concerning the relationship between y_i and x_i. In particular we shall consider the model,

$$E(Y_i / X_i = x_i) = M(x_i)$$

$$Var(Y_i / X_i = x_i) = \sigma^2(x_i) \qquad (3.1)$$

$$Cov\,(y_i, y_j) = 0 \quad i \neq j$$

Further we will assume that the functions $m(x_i)$, $\sigma(x_i)$ are Lipschits continous (i.e Smooth). Under this, several Non-Parametric procedures can be used to estimate the population total $T = \sum_{i=1}^{N} y_i$. The widely used smoothing procedures are ;

1) Smoothing splines [Wahba (1975)]

2) K- Nearest neighbour [K-N-N]

3) Kernel smoothers, i.e

(I) Priestly chao [Priestly and Chao (1972), Gasser and Muller (1979)].

(II) Nadaraya- Watson [Nadaraya (1964),Watson(1964)].

Non of these smoothing functions is uniformly best. However, Kernel smoothers have been found to have optimal minimax properties [Gasser and Engel (1990)]. As such, in this study we shall focus on Kernel functions of the Nadaraya - Watson and Priestly Chao type.

These smoothers are given as follows;

1) Priestly Chao (PC) weight represented by

$$W_h(x_i, x_j) = \left(\frac{x_j - x_{j-1}}{h}\right) k\left(\frac{x_i - x_j}{h}\right)$$

2) The Nadaraya Watson (NW) weight represented by

$$W_h(x_i, x_i) = \frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{j \in s} k\left(\frac{x_i - x_j}{h}\right)}$$

$j \in s$, $i \in U$

3.2 NON-PARAMETRIC ESTIMATOR OF THE FINITE POPULATION TOTAL T:

The Non parametric estimator for $M(x_i)$ is given by

$\hat{M}_{np}(x_i) = \sum_{j \in s} w_h(x_i, x_j)$ where $w_h(x_i, x_j)$ is the chosen smoothing wieght. Specifically for the PC smoother ,

$$\hat{M}_{pc}(x_i, x_j) = \sum_{j \in s}\left(\frac{x_j - x_{j-1}}{h}\right) k\left(\frac{x_i - x_j}{h}\right) y_j$$

Hence the estimator for the population based on tha *PC* smoother $\hat{T}_{pc}$ will be

$$\hat{T}_{pc} = \sum_{j \in s}(y_i) + \sum_{i \in r}[\hat{E}(y_i / x_i)] = \sum_{j \in s}(y_i) + \sum_{i \in r} \hat{M}_{pc}(x_i)$$

Which can be written as,

$$\hat{T}_{pc} = \sum_{j \in s}(y_i) + \sum_{i \in r}\sum_{j \in s}\left(\frac{x_j - x_{j-1}}{h}\right) k\left(\frac{x_i - x_j}{h}\right) y_j$$ similarly, for the *NW* smoother

$$\hat{M}_{nw}(x_i) = \frac{\sum_{j \in s} k\left(\frac{x_i - x_j}{h}\right)}{\sum_{j \in s} k\left(\frac{x_i - x_j}{h}\right)} y_j$$

Hence the estimator for T, based on NW smoother is $\hat{T}_{nw}$, where

$$\hat{T}_{nw} = \sum_{j\in s} y_j + \sum_{i\notin r}\sum_{j\in s} \frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{j\in s} k\left(\frac{x_i - x_j}{h}\right)} y_j$$

Let both $\hat{T}_{nw}$ and $\hat{T}_{pc}$ be represented by $\hat{T}_{NP}$ as

$$\hat{T}_{NP} = \sum_{j\in} y_j + \sum_{i\notin r}\left\{\sum_{j\in s} w_h(x_i, x_j) y_j\right\}$$

where *NP* stands for non parametric. Having derived these estimators, its imperative that we study the asymptotic properties of the estimators. We now turn to the asymptotic study of our estimators.

3.3 THE ASYMPTOTIC PROPERTIES OF $\hat{T}_{pc}$ AND $\hat{T}_{nw}$.

We shall study the asymptotic properties under the following conditions.

(i) $M(x_i) \geq c_o > 0$

(ii) $M(x_i)$ is twice continously differentiable

(iii) $\left|M(x_i) - M(x_j)\right| \leq \left|x_i - x_j\right|^{\beta}$ for some $\beta \in [0,1]$ i.e a lipschits contnuous

(iv) $K(u) = 0$ for all $|u| \geq 1$

(v) $K(u)$ is an even function

(vi) $|K(u)-K(v)| \leq M|u-v|^{\alpha}$, for some $\alpha \in [0,1]$ and M is a constant .

(vii) $\int_{-\infty}^{\infty} k(u)du = 1$, $\int_{-\infty}^{\infty} uk(u)du = 0$

and $\int_{-\infty}^{\infty} u^i k(u)du \neq 0$

assume

Further we shall ~~assuming~~ that x_i's are equispaced in the compact interval $[o,1]$.

then for $\hat{T}_{pc}$ where

$$E[\hat{M}(x_i)] = \sum_{j\in s}\left(\frac{x_j - x_{j-1}}{h}\right)k\left(\frac{x_i - x_i}{h}\right)M(x_j) \text{ can be written as}$$

$$= \frac{1}{nh}\sum_{j\in s}k\left(\frac{x_i - x_j}{h}\right)M(x_j) \approx$$

$$\frac{1}{h}\sum_{j\in s}\left[\int_{x_{j-1}}^{x_j}k\left(\frac{x_i - x_j}{h}\right)\right]dsM(x_j) \approx$$

$$\frac{1}{h}\sum_{j\in s}\left\{\int_{x_{j-1}}^{x_j}\left[k\left(\frac{x_i - x_j}{h}\right) - k\left(\frac{x_i - t}{h}\right)\right]m(t)dt + \int_{x_{j-1}}^{x_j}k\left(\frac{x_i - t}{h}\right)m(t) + \int_{x_{j-1}}^{x_j}k\left(\frac{x_i - x_j}{h}\right)\Big[m(x) - m(t)dt\Big]\right\}$$

$$\leq \frac{1}{h}\sum_{j\in s}\left[\int_{x_{j-1}}^{x_j}\underbrace{\left|\frac{t - x_j}{h}\right|^{\beta}}_{a}m(t)dt + \int_{x_{j-1}}^{x_j}k\left(\frac{x_i - t}{h}\right)m(t) + \int_{x_{j-1}}^{x_j}k\left(\frac{x_i - x_j}{h}\right)\underbrace{m|x_i - t|^{\alpha}}_{b}\right]$$

Now, consider part ..... *(a)* and let us write it as

$$\sum_{j\in s}\int_{x_{j-1}}^{x_j}\left[\frac{t}{h} - \frac{x}{h}\right]dt = \sum_{j\in s}\left[\left|\frac{t^2}{2h} - \frac{x}{h}\right|_{x_{j-1}}^{x_j}\right]$$

$$= \sum_{j\in s}\left[\frac{x_j^2}{2h} - \frac{x_j^2}{h} - \frac{x_{j-1}^2}{2h} + \frac{x_j - x_{j-1}}{h}\right]$$

$$= \sum_{j\in s}\left[\frac{1}{2h}\left(x_j^2 - x_{j-1}^2\right) - \frac{1}{h}x_j\left(x_j - x_{j-1}\right)\right]$$

$$= \sum_{j\in s}\left[\frac{1}{2h}(x_j + x_{j-1})(x_j - x_{j-1}) - \frac{1}{h}x_j(x_j - x_{j-1})\right]$$

$$= \sum_{j\in s}\left[\frac{x_j + x_{j-1}}{2nh} - \frac{x_j}{nh}\right]$$

$$= \frac{n\bar{x}_s + n\bar{x}_s - x_1}{2hn} - \frac{2n\bar{x}_s}{2hn}$$

$$= \frac{x_1}{2nh} \quad which \to 0 \text{ as } hn \to \infty \quad which\ implies\ that\ .....(a) \to 0 \text{ as}$$

$nh \to \infty$ Similarly for .....*(b)* $|x_i - t|^{\alpha} \to 0$ as $nh \to \infty$. So our equation now reduces to

$$E\hat{M}(x_i) \approx \frac{1}{h}\sum \int_{x_{j-1}}^{x_j} k\left(\frac{x_i - t}{h}\right) m(t)dt \text{ but } m(t) \text{ can be expanded by the}$$

Taylor's expansion as,

$$m(t) = m(x_i) + hum'(x_i) + \frac{h^2u^2}{2}m''(x_i). \text{ So,}$$

$$E(m(xi)) \approx \frac{1}{h}\sum m(xi) \int_{xj-1}^{xj} k\left(\frac{x_i - t}{h}\right)\left[m(x_i) + hum'(x_i) + \frac{h^2u^2m''(x_i)}{2}\right]$$

$$\approx \frac{1}{h}\sum_{j\in s} m(xi) \int_{\frac{x_i - x_{j-1}}{h}}^{\frac{x_i - x_j}{h}} kuhdu + \frac{h^2m'(xi)}{h}\underbrace{\sum_{j\in s}\int_{\frac{x_i - x_{j-1}}{h}}^{\frac{x_i - x_j}{h}} uk(u)du}_{c} + h^2m'(xi)\sum_{j\in s}\int_{\frac{x_i - x_{j-1}}{h}}^{\frac{x_i - x_j}{h}} u^2k(u)du$$

But c = 0 from condition (vii)so we now remain with

$$E(\hat{m}x_i)) = m(xi)\underbrace{\left[\sum_{j\in s}\int_{\frac{x_i - x_{j-1}}{h}}^{\frac{x_i - x_i}{h}} k(u)du\right]}_{d} + h^2 + \frac{h^2m''(xi)}{2}\sum_{j\in s}\int_{\frac{x_i - x}{h}}^{\frac{x_i - x_j}{h}} u^2kudu.$$

Now consider ... *(d)*, wecan expanded it as

$$m(xi)\left[\int_{\frac{x_i-x_0}{h}}^{\frac{x_i-x_1}{h}} kudu + \int_{\frac{x_x-x1}{h}}^{\frac{x_i-x_2}{h}} kudu + \ldots\ldots\ldots\ldots\ldots\ldots + \int_{\frac{x_i-x_{n-2}}{h}}^{\frac{x_i-x_{n-1}}{h}} kudu + \int_{\frac{x_i-x_{n-1}}{h}}^{\frac{x_i-x_n}{h}} kudu\right]$$

if we now let $u_j = \left(\frac{x_i - x_j}{h}\right)$ and $\int kudu = f(u)$ then *(d)* can be written as

$$d = m(x_i)\left[f(u_1) - f(u_2) + f(u_2)\ldots\ldots\ldots\ldots - f(u_{n-1}) + f(u_{n-1}) - f(u_n)\right]$$

$$= m(x)\left[f(u_n) - f(u_0)\right]$$

$$= m(x_i)\int_0^1 k(u)du$$

$= m(x_i)$ from condition *(vii)* . We now have

$$E\left(\hat{m}(x_i)\right) = m(x_i) + \frac{h^2 m''(x_i)}{2}\int_0^1 u^2 k(u)du$$

$$\sum_r E\left(\hat{m}(x_i)\right) = \sum_r m(x_i) + \sum_r \frac{h^2 m''(x_i)}{2}\int_0^1 u^2 k(u)du.$$ Hence

$$E\left(\hat{T}_{pc}\right) = \sum_s m(x_i) + \sum_r m(x) + \sum_r \frac{h^2 m''(x_i)}{2}\int_0^1 u^2 k(u)du$$

Therefore

$$E\left(\hat{T}_{pc} - (T)\right) = \sum_s m(x_i) + \sum_r m(x_i) + \sum_r \frac{h^2 m''(x_i)}{2}\int_0^1 u^2 k(u)du - \sum_s m(x_i) - \sum_r m(x_i)$$

$$E\left(\hat{T}_{pc} - T\right) = \sum_r \frac{h^2 m''(x_i)}{2}\int_0^1 u^2 k(u)du$$

hence $\hat{T}_{pc}$ is biased for T and the bias is given by

$$\sum_r \frac{h^2 m''(x_i)}{2} \int_0^1 u^2 k(u)du = \frac{h^2 c}{2}, \ldots c = \sum_r m''(x_i) \int_0^1 u^2 k(u)du .$$

This bias goes to 0 as $h \to 0$ as such $\hat{T}_{pc}$ is asymtotically unbiased for T.

In a similar manner, it can be shown that ,

$$E(\hat{T}_{nw}) = \sum_s m(x_j) + \sum_r m(x_i) + \frac{h}{2} \sum_r m''(x_i) \int_0^1 u^2 kudu$$

Notice that $E(\hat{T}_{nw}) = E(\hat{T}_{pc})$ under the above asymptotic conditions.

REMARKS

(I) This result is true only if the above assumptions hold.However,at the boundary of the interval [0,1] ,there could be a problem .The sparse sample (boundary) problem.

(II) Clearly the estimators based on Non-parametric procedures are biased. This bias can be eliminated if in addition to the above conditions we let $h \Rightarrow 0$. In essence this implies that the new procedures proposed here are biased-robust to misspecification of $E(Y_i / X_i = x_i)$.From remark (II) it is imperative that we consider MSE as a criterion for assesing the accuracy of $\hat{T}_{np}$.This is what we will study in the next chapter.

3.4 THE CONDITIONAL MEAN OF $(\hat{T}np - T)$

Now,The prediction error is given by,

$$(\hat{T}np - T) = \sum_s y_i + \sum_r \hat{m}(x_i) - \sum_s y_i - \sum_r y_i$$

$$= \sum_r \hat{m}(x_i) - \sum_r y_i$$

Therefore, $E(Tnp - T) = \sum_r E[\hat{m}(x_i)] - \sum_r E[y_i] = \sum_r \left[E[\hat{m}(x_i) - E[y_i]\right]$

$$= \sum_{r \in \bar{s}} \left[\sum_{j \in s} w_h(x_i x_i) m(x_j) - m(x_i) \right]$$

In the special case when $w_h(x_i,x_j)$ is the *nw* smoother, then the prediction error is given by,

$$E(\hat{T}np - T) = \sum_{r \in s} \left[\sum_{j \in s} \left(\frac{\frac{1}{hn} k\left(\frac{x_i - x_j}{h}\right) m(x_j)}{\frac{1}{hn} \sum_{j \in s} k\left(\frac{x_i - x_j}{h}\right)} \right) - m(x_i) \right]$$

$$= \sum_{r \in s} \left\{ d_s(x_j) \right\}^{-1} \sum_{j \in s} \frac{1}{hn} k\left(\frac{x_i - x_j}{h}\right) m(x_j) - m(x_i) \text{ Where}$$

$$d_s\left\{x_j\right\} = \frac{1}{hn} \sum_{j \in s} k\left(\frac{x_i - x_j}{h}\right).$$

We now turn our attention to the error variance and its estimation.

CHAPTER FOUR

4.0 THE ERROR VARIANCE AND ITS ESTIMATORS

The prediction error is given by

$$(\hat{T}_{np} - T) = \sum_r \hat{m}(x_i) - \sum_r (y_i).$$

Therefore its variance is

$$Var(\hat{T}_{np} - T) = Var\sum_r \hat{m}(x_i) - \sum_r Var(y_i) \qquad (4.1)$$

$$= Var\left[\sum_r \hat{m}(x_i)\right] + \sum_r \sigma^2(x_i)$$

If we now consider $Var\left[\sum_r \hat{m}(x_i)\right]$, this can be written as

$$Var\left[\sum_r \hat{m}(x_i)\right] = \sum_{i\in r}\sum_{k\in r} Cov([\hat{m}(x_i)],[\hat{m}(x_k)])$$

where $Cov(\hat{m}(x_i),\hat{m}(x_k)) = \sum_{j\in s}\sum_{l\in s} w_h(x_i,x_j)w_h(x_k,x_l) \Rightarrow$

$$Cov(\hat{m}(x_i),\hat{m}(x_k)) = \sum_{i\in r}\sum_{k\in r}\sum_{j\in s} w_h(x_i,x_j)w_h(x_k,x_j)\sigma^2(x_j).$$

For the special case of *NW* smoother,

$$Var(\hat{T}_{np} - T) = \sum_{i\in r}\sum_{k\in r}\sum_{j\in s}\left[\frac{\frac{1}{nh}k\left(\frac{x_i - x_j}{h}\right)\ \frac{1}{nh}k\left(\frac{x_k - x_j}{h}\right)}{\frac{1}{nh}\sum_{j\in s}k\left(\frac{x_i - x_j}{h}\right)\ \frac{1}{nh}\sum_{j\in s}k\left(\frac{x_k - x_j}{h}\right)}\sigma^2(x_j)\right] + \sum_{i\in r}\sigma^2(x_i)$$

.....(4.2)

$$= \sum_{i\in r}\sum_{k\in r}\sum_{j\in s}\frac{1}{nh}k\left(\frac{x_i - x_j}{h}\right)\frac{1}{nh}k\left(\frac{x_k - x_j}{h}\right)\sigma^2(x_j) + \sum_{i\in r}\sigma^2(x_i).$$

Since $\frac{1}{nh}\sum_{j\in s}k\left(\frac{x_i - x_j}{h}\right) = 1$ [Odhiambo (1996)].

But $\sigma^2(x_j)$ can be expanded by the Taylor's expansion as

$$\sigma^2(x_i) = \sigma^2(x_i) + hu\sigma'^2(x_i) + \frac{(hu)^2}{2}\sigma''^2(x_i)$$

$$\cong \sigma^2(x_i)$$

so that,

$$Var(\hat{T}_{np} - T) \cong \sum_{i\in r}\sum_{k\in r}\sum_{j\in s}\frac{1}{(nh)^2}k\left(\frac{x_i - x_j}{h}\right)k\left(\frac{x_k - x_j}{h}\right)\sigma^2(x_i) + \sum_{i\in r}\sigma^2(x_i).$$

Gasser and Muller (1984) gives an integral approximation of $Cov[\hat{m}(x_i), \hat{m}(x_j)]$ as $Cov[\hat{m}(t_i), \hat{m}(t_j)] = \frac{\sigma^2(x_i)}{hn}\int_o^1 k\left(\frac{x_i - s}{h}\right)k\left(\frac{x_k - s}{h}\right)ds$. Using this, we can approximate

$\sum_{j\in s}\frac{1}{(nh)^2}k\left(\frac{x_i - x_j}{h}\right)\frac{1}{nh}k\left(\frac{x_k - x_j}{h}\right)\sigma^2(x_j)$ by,

$$\frac{\sigma^2(x_i)}{hn}\int_{\frac{x_i}{h}}^{\frac{x_{i-1}}{h}} k(u)k\left(\frac{x_k - x_i}{h} + u\right)du \text{ where } u = \left(\frac{x_i - x_j}{h}\right).$$ Hence the error variance can be expressed as

$$Var(\hat{T}_{np} - T) \approx \frac{1}{hn}\left[\sum_{i\in S}\sum_{k\in r}\sigma^2(x_i)\int_{\frac{x_i}{h}}^{\frac{x_{i-1}}{h}} k(u)k\left(\frac{x_k - x_i}{h} + u\right)du\right] + \sum_{i\in r}\sigma^2(x_i)$$

Clearly

$$Var\left[\frac{(\hat{T}_{np} - T)}{N}\right] = \frac{1}{hn}\left[\frac{1}{N^2}\sum_{i\in S}\sum_{k\in r}\sigma^2(x_i)\int_{\frac{x_i}{h}}^{\frac{x_{i-1}}{h}} k(u)k\left(\frac{x_k - x_i}{h} + u\right)du\right] + O(N^{-1}).$$

This implies $Var\left[\frac{(\hat{T}_{np} - T)}{N}\right] \to 0$ *as* $nh \to \infty$, a fact implicitly implied by this result is that

$\hat{T}_{np} \to T$. Hence $\hat{T}_{np}$ is a consistent estimator of T.

4.1 THE VARIANCE OF $\hat{T}_{np}$

This is given by

$$\hat{T}_{np} = \sum_{i\in s} y_i + \sum_{i\in r}\sum_{j\in s} w_h(x_i, x_j)y_j$$

$$\Rightarrow \quad Var(\hat{T}_{np}) = \sum_{i\in s}\sigma^2(x_j) + Var\sum_{i\in r}\sum_{j\in s} w_h(x_i, x_j)y_j$$

$$= \sum_{i\in s}\sigma^2(x_j) + \sum_{i\in r}\sum_{k\in s} w_h^2(x_i, x_j)Var(y_j)$$

$$= \sum_{i\in s}\sigma^2(x_j) + Var[\hat{m}(t_i)]$$

$$Var\left[\frac{(\hat{T}_{NP})}{N}\right] = O(N^{-1}f) + \frac{1}{N^2}\sum_{i\in r}\sum_{k\in s}\frac{\sigma^2(x_i)}{nh}\int_0^1 k^2(u).$$

Therefore,

$$Var(\hat{T}_{np}) \cong \sum_{i\in\sigma}\sum_{k\in\sigma}\frac{\sigma^2(x_i)}{nh}\int_0^1 k^2(u)$$

Having obtained the error variance, we now turn our attention to the estimation of this error variance.

4.2 THE ESTIMATION OF THE ERROR VARIANCE

In section 4.0 We showed that the error variance is given by,

$$Var(\hat{T}_{NP} - T) = Var\sum_r \hat{m}(x_i) + \sum_{j\in\sigma}\sigma^2(x_j)$$

$$= \sum_r \left[w_h(x_i, x_j)\right]^2 \sigma^2(x_i) + \sum_{j\in\sigma}\sigma^2(x_j).$$

Let us rewrite this as

$$Var(\hat{T}_{NP} - T) = \sum_{i\in s} w_i^2\sigma^2(x_i) + \sum_{j\in\sigma}\sigma^2(x_j).$$

We observe that this is a function of $\sigma^2(x_i)$'s

for i=1,2,3.....N., which are unknown. The purpose of this section is to propose procedures of estimating $\sigma^2(x_i)$'s.

Dofman (1994) applied the method of moments to estimate $\sigma^2(x_i)$, in particular, he defines $\hat{\sigma}^2(x_i) = E(y_i^2) - \left[E(y_i)\right]^2$. He then proceeds to estimate $E(y_j^2)$ and $E(y_j)$

by $\sum_{j\in s} w_h(x_i x_j)y_j^2$ and $\sum_{j\in s} w_l(x_i x_j)y_j$ respectively. Where l and h are the chosen bandwidth in each case and $w_h(x_i\, x_j)$ is the chosen smoothing weight.

Consequently, he obtains

$$\hat{\sigma}^2(x_i) = \sum_{j\in s} w_l(x_i x_j) y_j^2 - \left[\sum_{j\in s} w_l(x_i x_j) y_i\right]^2 . \qquad (4.3)$$

This leads to

$Var(\hat{T}_{np} - T) = \sum_{i\in s} w_h^2 \hat{\sigma}^2(x_i) + \sum_{j\in r} \hat{\sigma}^2(x_j)$. Considering Expression ..(4.3)

there is a possibility that $\sum_{j\in s} w_h(x_i x_j) y_j^2 < \left[\sum_{j\in s} w_l(x_i x_j) y_i\right]^2$. In such a case, $\hat{\sigma}^2(x_i)$ will take negative value. As if to reiforce this short coming, Dorfman(1994) in one run when $h=0.25$ obtained a negative variance. As a solution to this problem he suggested that h=1. Clearly this is not realistic as h=l may not be the optimal bandwidth in either case. To solve this problem, we suggest two new procedures of estimating $\sigma^2(x_i)$. Our procedure will be based on

(I) Kernel procedures

(II) Model based bootstrap procedure.

4.3 KERNEL PROCEDURE

Here we shall use the *NW* and the *PC* smoothers considered in section 3.1. We shall base our method on the fact that the squared residuals $(y_i - \hat{m}(x_i))^2$ is approximately unbiased estimator of $\sigma^2(x_i)$ Horn, Duncan(1975),Royal and Cumberland(1978). That is $\sigma^2(x_i) \cong (y_i - \hat{m}(x_i))^2$.

If we now let $\hat{e}_j = \left(y_j - \hat{m}(x_j)\right)^2$ to be our naive estimator for $\sigma^2(x_i)$, we shall then seek to obtain an improved estimate by smoothing $\hat{e}_j = \left(y_j - \hat{m}(x_j)\right)^2$ for j∈s and x_i, x_j are sample points close to x_j, y_j. Where closeness of $x_i y_j$ to $x_i x_j$ is measured in terms of the distance $|x_i - x_j|$.

Now let $w_h(x_i, x_j)$ be a linear smoothing function, then we can write,

$$\hat{\sigma}^2_{NP}(x_i) = \sum_{j \in s} w_h(x_i x_j)\hat{e}_j^2 . \qquad (4.4)$$

So that the error variance given in equation (4.1) becomes can be estimated by

$$Var(\hat{T}_{NP} - T) = \sum_{i \in s} w_i^2 \hat{\sigma}^2_{NP}(x_i) + \sum_{j \in r} \hat{\sigma}^2_{NP}(x_i) .$$

For the special case of *NW* smoother,

$$w_h(x_i, x_j) = \frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{j \in s} k\left(\frac{x_i - x_j}{h}\right)}$$

(4.4) becomes,

$$\sigma^2_{nw}(x_i) = \sum_{j \in s}\left[\frac{k\left(\frac{x_i - x_j}{h}\right)}{\sum_{j \in s} k\left(\frac{x_i - x_j}{h}\right)}\right]\hat{e}_j^2 . \qquad (4.5)$$

Similarly for the PC smoother (4.4) becomes

$$\sigma^2_{pc}(x_i) = \sum_{j \in s} k\left(\frac{x_j - x_{j-1}}{h}\right) k\left(\frac{x_i - x_j}{h}\right)\hat{e}_j^2 . \qquad (4.6)$$

(4.5) and (4.6) above gives rise to two proposed kernel based estimators of the error variance, for PC and NW smoothers repetitively. Which we shall denote by:-

$$V_{pc} = \sum_{i \in s} w_i^2 \sigma_{pc}^2(x_i) + \sum_{j \in r} \sigma_{pc}^2(x_i) \quad (4.7)$$

$$V_{nw} = \sum_{i \in s} w_i^2 \sigma_{nw}^2(x_i) + \sum_{j \in r} \sigma_{nw}^2(x_i).$$

Having derived our estimators, it is most natural that we study their asymptotic properties and this is what we turn to in the next section.

4.4 THE ASYMPTOTIC PROPERTIES OF V_{pc} AND V_{nw}

We shall study the asymptotic properties of these estimators under the following conditions.

(I) $\sigma^2(x_i)$ is twice continuously differentiable.

(II) $\sigma^2(x_i) \geq c_o > 0$ for all $x_i \in [a,b]$.

(III) $\left|\sigma^2(x_i) - \sigma^2(x_j)\right| \leq \left|x_i - x_j\right|^{\beta}$ for some $\beta \in [0,1]$.

i.e. $\sigma^2(x_i)$ is alipschits continuous.

(IV) $k(u)=0$ for all ≥ 1.

(V) $k(u)$ is an even function.

(VI) $|k(u)-k(v)| \leq m|u-v|^{\alpha}, u,v \in (0,1)$ for some $\alpha \in (0,1)$ and m is a constant.

(VII) $\int_{-\infty}^{\infty} k(u)du = 1, \int_{-\infty}^{\infty} uk(u)du = 0, \int_{-\infty}^{\infty} u^i k(u)du \neq 0,$ for $I=2$.

(VIII) $E_{\xi}(y_j^2) = \mu^4(x_j) < \infty$.

4.4.1 THE ASYMPTOTIC PROPERTIES OF V_{PC}

Since the PC and NW weights are theoretically equivalent, it will be sufficient for us to study the asymptotic properties of the PC.

Taking the limiting factor as n→ ∞ of equation (4.7) we have

$$Lim_{n\to\infty}E_{\xi}[V_{pc}] = Lim_{n\to\infty}\left\{\sum_{i\in s} w_i^2 E_{\xi}\sigma_{pc}^2(x_i) + \sum_{i\in r} E_{\xi}\sigma_{pc}^2(x_i)\right\} \qquad (4.8)$$

but from (4.6) we know that $\sigma_{pc}^2(x_i) = \sum_{j\in s} k\left(\frac{x_j - x_{j-1}}{h}\right)k\left(\frac{x_i - x_j}{h}\right)\hat{e}_j^2$

therefore,

$$E_{\xi}\sigma_{pc}^2(x_i) = \sum_{j\in s} k\left(\frac{x_j - x_{j-1}}{h}\right)k\left(\frac{x_i - x_j}{h}\right)E_{\xi}\hat{e}_j^2.$$

Further, it is known that

$E_{\xi}(e_j^2) = \sigma^2(x_j) + O(n^{-1})$ subject to some mild conditions (

(See Royal and Cumberland 1978). In the light of this we can write

$$E_{\xi}\sigma^2(x_i) \cong \frac{1}{h}\sum_{j\in s}\left\{\int_{x_{j-1}}^{x_j} k\left(\frac{x_i - x_j}{h}\right)\sigma^2(x_i)dt\right\}$$

$$\cong \frac{1}{h}\sum_{j\in s}\left\{\int_{x_{j-1}}^{x_j}\left[k\left(\frac{x_i - x_j}{h}\right) - k\left(\frac{x_i - t}{h}\right)\right]\sigma^2(t)dt + \int_{x_{j-1}}^{x_j}\left[k\left(\frac{x_i - x_j}{h}\right)\right]\sigma^2(t)dt + \int_{x_{j-1}}^{x_j}\left[k\left(\frac{x_i - x_j}{h}\right)\left(\sigma^2(x_j) - \sigma^2(t)\right)dt\right]\right\}$$

.

This implies that

$$\left|E_\xi\sigma^2(x_i)-\frac{1}{h}\sum_{j\in s}\int_{x_{j-1}}^{x_j}\left[k\left(\frac{x_i-t}{h}\right)\right]\sigma^2(t)dt\right| \le \left|\frac{M}{h}\sum_{j\in s}\int_{x_{j-1}}^{x_j}\left[k\left(\frac{x_i-t}{h}\right)^\alpha\right]\sigma^2(t)dt\right|$$

$$+\left|\frac{L}{h}\sum_{j\in s}\int_{x_{j-1}}^{x_j}(x_i-t)^\beta\left[k\left(\frac{x_i-t}{h}\right)^\alpha\right]dt\right|$$

$$\le M\sup\left\{\frac{1}{h}\sum_{j\in s}\sigma^2(x_j)\left[k\left(\frac{x_i-x_{j-1}}{h}\right)\right]_{\frac{1}{\alpha+1}}^{\alpha+1}+L\sum_{j\in s}\frac{k(u)(x_i-x_{j-1})^{\beta+1}}{\beta+1}\right\};u=\frac{x_i-x_j}{h} \qquad (4.9)$$

$$\le o(nh)^{-(\alpha+1)}+\frac{Lh}{n^\beta(\beta+1)}\int uk(u)du = o(nh)^{-(\alpha+1)}.$$

Clearly, the R.H.S of (4.9) goes to zero if nh$\to\infty$ as n $\to\infty$. Hence,

$$E_\xi\left\{\sigma^2{}_{pc}(x_i)\right\}\cong\frac{1}{h}\sum_{j\in s}\int_{x_{j-1}}^{x_j}K\left[\frac{x_i-t}{h}\right]\sigma^2(t)dt.$$

If we expand $\sigma^2(t)$ about x_i, by Taylor's expansion we shall obtain,

$$E_\xi\sigma^2_{pc}(x_i)\cong\sigma^2(x_i)+\frac{h^2}{2}\sigma''^2(x_i)\int_{\frac{x_i-x_{j-1}}{h}}^{\frac{x_i-x_j}{h}}u^2k(u)du \quad \cong\sigma^2(x_i)+\frac{h^2}{2}\sigma''^2(x_i)d_k$$

where $d_k=\int_{\frac{x_i-x_{j-1}}{h}}^{\frac{x_i-x_j}{h}}u^2k(u)du$. Substituting for $E_\xi\left\{\sigma^2{}_{pc}(x_i)\right\}$ in (4.7) will lead to

$$E_\xi[V_{pc}]=\left\{\sum_{i\in s}\left[w_i^2\sigma^2(x_i)+\frac{h^2}{2}\sigma''^2(x_i)d_k\right]+\sum_{i\in r}\left[w_i^2\sigma^2(x_i)+\frac{h^2}{2}\sigma''^2(x_i)d_k\right]\right\}$$

The Bias of V_{pc} given by $B_\xi(V_{pc}) = E_\xi(V_{pc}) - Var_\xi(T_{pc} - T)$ will therefore be

$B_\xi[V_{pc}] = d_k \frac{h^2}{2} \sum_{i \in s} w_i \sigma''^2(x_i) + d_k \frac{h^2}{2} \sum_{j \in r} \sigma''^2(x_i)$, hence the relative Bias is

$$B_\xi\left\{\frac{V_{pc}}{Var(\hat{T}_{pc} - T)}\right\} = d_k \frac{h^2}{2}\left[\frac{\sum_{i \in s} w_i \sigma''^2(x_i) + \sum_{j \in r} \sigma''^2(x_i)}{\sum_{i \in s} w_i \sigma(x_i) + \sum_{j \in r} \sigma(x_i)}\right]$$

$$= d_k \frac{h^2}{2}\left[\frac{\sum_{i \in s} w_i \sigma''^2(x_i)}{\sum_{i \in s} w_i \sigma(x_i) + \sum_{j \in r} \sigma(x_i)}\right] + d_k \frac{h^2}{2}\left[\frac{\sum_{j \in r} \sigma''^2(x_i)}{\sum_{i \in s} w_i \sigma(x_i) + \sum_{j \in r} \sigma(x_i)}\right]$$

We note that the terms in brackets are bounded. Hence if $h \to 0$ as $n \to \infty$ The relative bias goes to zero. This implies that

$$E_\xi(V_p) \cong Var_\xi(\hat{T}_{pc} - T).$$

We therefore conclude that V_p is asymptotically unbiased for the error variance.

Another estimation method that in gaining popularity in recent research is the bootstrap method. We shall study this method and its application in estimation of error variance in the next section.

4.5 MODEL BASED BOOTSTRAP METHOD OF ESTIMATING THE ERROR VARIANCE.

4.5.0 INTRODUCTION

This is a resampling technique where repeated samples are taken from the initial samples and an estimator constructed from the repeated samples.

This procedure can be used to provide standard error estimates and confidence interval for the parameters of interest.

Efron (1989) gave an extensive study of this method in the identically independently distributed case to obtain standard error estimates for nonparametric confidence intervals for any parametor of interest.

Odhiambo (1995) asserts that Bootstrap potentially offers a method of overcoming robustness problem associated with model based surveys.

4.5.1 THE PROPOSED PROCEDURE

We shall assume that the population follows the super population model (3.1). Under this model , and following nonparameteic procedures in section (3.4), we derived the prediction error as

$$\hat{T}_{np} - T = \sum_{i \in r} \sum_{j \in s} w_h(x_i, x_j) y_j - \sum_{i \in r} y_i \qquad (4.10)$$

From conditions of model (3.1) we can rewrite (4.10) as

$$(\hat{T}_{NP} - T) = \sum_{j \in s}\sum_{i \in r} w_h(x_i x_j) y_j - \sum_{i \in r} y_i$$

Let us now express the prediction error as

$$\sum_{j \in s} a_j y_j - \sum_{i \in r} y_i$$

where $a_j = \sum_{i \in r} w_h(x_i x_j) y_j$

and $w_h(x_i x_j)$ is the chosen smoothing weight. we shall use this expression to obtain the error variance.

4.5.2 THE ERROR VARIANCE

This will be given by $Var(\hat{T}_{NP} - T) = Var\sum_{j \in s} a_j y_j + Var\sum_{i \in r} y_i$

$$= Var\sum_{j \in s} a_j e_j + Var\sum_{i \in r} e_i$$

Something wrong here

With us now ,are two components to estimate, that is

$Var\sum_{j \in s} a_j e_j$ and $Var\sum_{i \in r} e_i$

We first estimate the first component computing the quantities

$\{a_1e_1, a_2e_2, a_3e_3, \ldots\ldots\ldots\ldots a_ne_n\}$ and treat them as the parent sample. We shall then draw B samples of size n using simple random sampling with replacement.

So that we have

$$\hat{a}_{11}\hat{e}_{11}, \hat{a}_{12}\hat{e}_{12}, \hat{a}_{13}\hat{e}_{13}, \ldots\ldots\ldots\ldots\ldots\ldots \hat{a}_{1n}\hat{e}_{1n}$$

$$\hat{a}_{21}\hat{e}_{21}, \hat{a}_{22}\hat{e}_{22}, \hat{a}_{23}\hat{e}_{23}, \ldots\ldots\ldots\ldots\ldots\ldots \hat{a}_{2n}\hat{e}_{2n}$$

$$\hat{a}_{31}\hat{e}_{31}, \hat{a}_{32}\hat{e}_{32}, \hat{a}_{33}\hat{e}_{33}, \ldots\ldots\ldots\ldots\ldots\ldots \hat{a}_{3n}\hat{e}_{3n}$$

. . . . . .

. . . . . .

. . . . . .

$$\hat{a}_{B1}\hat{e}_{B1}, \hat{a}_{B2}\hat{e}_{B2}, \hat{a}_{B3}\hat{e}_{B3}, \ldots\ldots\ldots\ldots\ldots\ldots \hat{a}_{Bn}\hat{e}_{Bn}$$

From each of the B samples we compute

$$\left\{R_1^* R_2^*, R_3^*, R_4^* \ldots\ldots\ldots\ldots R_{B-1}^*, R_B^*\right\}$$

where $R_b^* = \sum_{j=1}^{n} \hat{a}_{bj}\hat{e}_{bj}$.

Then the variance estimator $Var\sum_{j\in s} a_j e_j$ will be given by the Monte Carlo formula

as $\frac{1}{B-1}\sum_{b=1}^{B}(R_b^* - \overline{R}^*)^2$ Where $\overline{R}^* = \frac{1}{B}\sum_{b=1}^{B} R_b^*$.

To estimate the second component $Var\sum_{i\in r} e_i$, we recall that

$Var\sum_{i\in r} y_i \cong Var\sum_{i\in r} \hat{m}_{np}(x_i) = \sum_{i\in r} \sigma^2{}_{np}(x_i)$ which we got as the nonparametric estimator of $Var\sum_{i\in r} y_i$ in section 4.2.1

Combining the two terms together we obtain

$$V_{B-NP} = \frac{1}{B-1}\sum_{b=1}^{B}(R_b^* - \overline{R}^*)^2 + \sigma^2{}_{NP}(x_i) \qquad (4.11)$$

Thus we shall have two estimators of the variance depending on the smoothing weight adopted. i.e

$$V_{B-pc} = \frac{1}{B-1}\sum_{b=1}^{B}(R_b^{*pc} - \overline{R}^*)^2 + \sigma^2{}_{pc}(x_i) \text{ and}$$

$$V_{B-nw} = \frac{1}{B-1}\sum_{b=1}^{B}(R_b^{*nw} - \overline{R}^*)^2 + \sigma^2{}_{nw}(x_i)$$

CLOSING REMARK

So far we have studied six different estimators of the error variance, namely V_D, V_L, V_{pc}, V_{nw} $V_{B\text{-}pc}$ and $V_{B\text{-}nw}$. This study would be incomplete without an empirical study on the efficiency of our estimators. To find out which estimator is more efficient and in what circumstances ,we now embark on empirical study in the next chapter.

CHAPTER FIVE

5.0 EMPIRICAL STUDY

5.1 DESCRIPTION OF THE POPULATIONS

Properties of the six variance estimators V_L , V_D , V_{NW}, V_{PC} , $V_{B\text{-}NW}$ and $V_{B\text{-}PC}$ are studied in five populations: four artificial populations and one natural population. The artificial populations were constructed in the following manner.

(I) First we generated 300 points of x_i [$x_i \sim U(0,1)$] mutually independent across i.

(II) We generated $e_i \sim N(0,1)$ $i = 1, 2 \ldots\ldots\ldots 300.$

(III) The population data points y_i were then generated by use of the following modes.

Structure	Population	Model	Variance
Linear	*AP1*	$y_i = 100\, x_i + \sqrt{x_i} e_i$	$x_i \sigma^2$
	AP2	$y_i = 100\, x_i + x_i e_i$	$x_i^2 \sigma^2$
Quadratic	*AP3*	$y_i = 100 x_i^2 + \sqrt{x_i} e_i$	$x_i \sigma^2$
	AP4	$y_i = 100 x_i^2 + x_i e_i$	$x_i^2 \sigma^2$

for $i = 1, 2, 3. \ldots\ldots\ldots\ldots, 300$ *and* $\sigma^2 = 1$

The natural population (NP) of size 195 is the production by mass of dairy products by different farms in Kenya for the year 1992 and 1996. *[Source; Central Bureau of Statistics]. The 1992 values served as the auxiliary information (x_i) and 1996 values served as the variable under study y_i for $i = 1,2,3 \ldots\ldots, 195$.

We note that AP1 is perfectly linear, it agrees with the assumptions of the ratio model (2.1).We violated the variance conditions, and retained the linear relationship of Y_i to X_i. This gave the rise to population *AP2*. To study the effect of the linear restriction we introduced the quadratic relationship for Y_i to X_i but retained the variance condition. This gave rise to population *AP3*. For *AP4* we violated both the linearity condition and the variance condition.

[Scatter diagram of NP is given figure VI].

5.2 DESIGN OF THE STUDY

Using simple random sampling without replacement we drew 1000 samples of size $n=40$ for each of the five populations. For the Bootstrap variance estimators, $B=100$ Bootstrap resamples were generated for each of the 1000 parent samples.

The optimal Kernel.

$$k(u) = \frac{3}{4}[1-u^2]$$

$$= 0 \text{ if } |U^2| > 1$$

was used in the study of the non-parametric estimators.

5.3 THE SEARCH FOR AN OPTIMAL BANDWIDTH.

An optimal bandwidth h_{nw} and h_{pc} for Nadaraya Watson and Priestly Chao smoothers respectively was searched within the interval:

$\frac{\sigma}{4n^{\frac{1}{5}}} \le h \le \frac{3\sigma}{2n^{\frac{1}{5}}}$ where σ is the standard deviation of x_i's as given in Silverman(1986).

We settled for h_{nw} and h_{pc} that minimised MSE_{nw} and MSE_{pc} respectively.

For each of the five population and for each sample we computed,

$\overline{T} = \sum_{i=1}^{N} \frac{y_i}{N}$ and the prediction errors

$$E_R = (\hat{\overline{T}}_R - \overline{T})$$
$$E_{nw} = (\hat{\overline{T}}_{nw} - \overline{T})$$
$$E_{pc} = (\hat{\overline{T}}_{pc} - \overline{T})$$

The variance estimators V_D, V_L, V_{nw}, V_{pc}, $V_{B\text{-}nw}$ *and* $V_{B\text{-}pc}$ were computed for each sample. Three mean square error values,

$MSE_R = \frac{\sum E_R^2}{1000}$ $MSE_{nw} = \frac{\sum E_{nw}^2}{1000}$ and $MSE_{pc} = \frac{\sum E_{pc}^2}{1000}$ were computed for each population. The following un conditional values were computed for each population.

(I) Bias given by

$$B[V(.)] = \frac{\sum V(.)}{1000} - MSE(.)$$

(II) Relative Bias ,

$$REBI[(.)] = \frac{\sum V(.)}{1000\,MSE(.)} - 1$$

(III) Root Mean Square error,

$$RMSE[V(.)] = \left[\frac{\left(\sum V(.) - MSE(.)\right)^2}{1000}\right]^{\frac{1}{2}}$$

To see how each of the variance estimators is effective in tracking the MSE, we sorted the samples in an increasing order of and then grouped the samples into groups of 50 samples in that order. Further we computed

$MSE_{(.)} = \frac{\sum E_{(.)}^2}{50}$ for each of the three prediction errors and the averages $\frac{\sum V(.)}{50}$ for each of the six variance estimators in each of the resulting 20 groups.

The graphs of increasing $\bar{x}_s$ against $MSE_{(.)}$ and $\frac{\sum V(.)}{50}$ were plotted for each population to gauge how the different variance estimators are effective in tracking the

$MSE_{(.)}$. The graph of respective populations are given in figures 1, 2, 3, 4 and 5.

5.4 RESULTS OF THE EMPIRICAL STUDY

Results of this study are summarized in Table I through VII and figure 1 through 6

The mse

Table I summarizes the results of MSE arising from the three different procedures of estimation. We notice that for linear populations the ratio estimator is more superior than the nonparametric estimators in estimating population Total. This case still holds even after disturbing the Variance condition in AP2. The situation becomes reversed when we consider non linear models. Its quite noticeable that MSE_{PC} remained larger in all cases depicting T_{PC} as a poor estimator.

The Bias

In all the populations the V_{pc} was the poorest resulting in large biases as compared to other variance estimators. V_D and V_L under estimated MSE_R and MSE_{nw} but the bias was small for the linear populations. As expected V_D and V_L grossly over estimated both MSE_R and MSE_{nw} in the nonlinear populations. For the NP V_D and V_L resulted in smaller bias with $V_{B\text{-}nw}$ being a close rival.

The RMSE

V_{NW} was the best overall in minimizing the $RMSE_R$ performing poorly for the NP. For V_D and V_L the same picture of getting poorer as more linear restrictions were violated was repeated , but when it came to the natural population they were hard to beat. For $RMSE_{NW}$ and $V_{B\text{-}NW}$ was the best in all populations with V_D and V_L performing poorly

The REBI

For $REBI_{NW}$ V_{NW} and $V_{B\text{-}nw}$ produced small relative bias in quadratic populations but were beaten by V_D and V_L when it came to linear populations. $REBI_R$ for V_D and V_L were small in all populations.

Performance of $V_{(.)}$'s in tracking MSE_R and MSE_{nw}

Figure 1 reveals that MSE_{pc} and its variance estimators V_{pc} and $V_{B\text{-}pc}$ were so much removed from the from the others.

This outcome shows estimators based on PC smoothers are actually poor estimators. This result was independently established by Odhiambo (1996).

For AP1 V_D and V_L were better in tracing the MSE_R , we notice that at balanced case i.e. when $\bar{x}_s = \bar{X}$ MSE_R is almost equal to MSE_{nw} . V_{nw} remains consisted and does seem to vary with $\bar{x}_s$ .The same picture is reflected in figure2 for AP2.

Considering AP3 in Figure 3 we notice that MSE_R is very high in the unbalanced cases while MSE_{nw} continuously comes down as $\bar{x}_s$ grows large. In the balanced case MSE_{nw} and MSE_R are almost equal while V_D and V_L over estimated both Mean square errors. The most consistent estimator for MSE_{nw} is $V_{B\text{-}nw}$. This scenario is repeated in AP4.

For NAP competition is very stiff in fact, MSE_R and MSE_{NW} behaved almost the same way. V_D and V_L were overestimates while V_{NW} and $V_{B\text{-}nw}$ being underestimates. However , V_{NW} was much lower than the other variance estimators.

5.4.1 DISCUSSION

The above results reinforced our theory in many ways. The robustness problem customarily associated with ratio estimators came out so well, as they performed poorly immediately the model conditions are violated. The Nonparametric estimators proved to be more robust, in fact there mean square error did not change a great deal as we changed the model specifications.

The most interesting result came from the natural population. The ratio estimators became close competitors of the nonparametric estimators. This did not

come as a surprise as a quick glance at the scatter diagram in figure 6 will indicate a linear concentration. Nonetheless, the mean square error from nonparametric estimators still remained low.

Table I

Mean square error values for the five populations

	AP1	AP2	AP3	AP4	NP
MSE(R)	1.1544E-02	7.4964E-03	3.050406	3.066129	9.2551E-04
MSE(pc)	4.445365	3.373698	3.518624	4.134322	2.1335E-03
MSE(nw)	0.9284943	0.8716868	0.7401577	0.8639173	8.3717E-04

Table II

Unconditional Values for the Bias of Variance Estimates X 1000 for MSE(R)

	AP1	AP2	AP3	AP4	NP
V_D	-0.0389	0.1409	251.0	242	0.1004
V_L	0.0588	-1.5860	41.75	13.927	0.2877
V_{PC}	2.322	2453	-1506	-1710	15.039
V_{NW}	-0.01152	-7.4787	-3050	-3066	-0.9246
$V_{B\text{-}PC}$	1464	1550	-2039	-2221	14.140
$V_{B\text{-}NW}$	1.784	24.534	-3030	-3051	-0.6717

Table III

Unconditioned Values of the bias of Variance estimates X 1000 for MSE (nw)

	AP1	AP3	AP2	AP4	NP
V_D	-917	2444	2561	-864	0.1003
V_L	-916	2216	2352	-865	0.2877
V_{PC}	1405	491	803	1 589	15.0394
V_{NW}	-928	-863	-740	-871	-0.9246
$V_{B\text{-}PC}$	548	-1908	270	686	14.1404
$V_{B\text{-}NW}$	-899	-848	-719	-839	-0617

Table IV

Unconditional Values for the RMSE(R) X1000 for MSE(R)

	AP1	AP2	AP3	AP4	NP
V_D	364	232	104305	104308	0.3241
V_L	383	187	97686	92302	0.3833
V_{PC}	73811	7782	48730	42770	504
V_{NW}	0.2404	0.3208	95.642	962	2199
$V_{B\text{-}PC}$	46683	4926	31862	26618	476
$V_{B\text{-}NW}$	928	1012	544.8	379	7.995

Table V

Unconditional Values of the RMSE X1000 for MSE(nw)

	AP1	AP2	AP3	AP4	NP
V_D	334	205	104378	1045779	32
V_L	354	159	97759	97372	38
V_{PC}	73782	72796	48803	42840	504
V_{NW}	28	27	22.58	26	0.00059
$V_{B\text{-}PC}$	46650	49240	3193	26688	476
$V_{B\text{-}NW}$	899	985	617	449	7.99

Talbe VI

Unconditional Values for REBI X 1000 for MSE(nw)

	AP1	AP2	AP3	AP4	NP
V_D	-987	-991	3460.5	0.788	225.4
V_L	-986	-993	3177	4.542	449.1
V_{PC}	-1513	1823	1086	-.557	18070
V_{NW}	-999	-999	-999.965	-999	-999.0
$V_{B\text{-}PC}$	+588	-787	365	-724	16996
$V_{B\text{-}NW}$	-968	-963	-972	-925	-696.9

Table VII

Unconditional Values for REBI X1000 for MSE (R)

	AP1	AP2	AP3	AP4	NP
V_D	-3.45	18.804	82.308	2828	108.5
V_L	50.1	-211	13.688	2565	310.8
V_{PC}	201174	227931	-493	569	16249
V_{NW}	-998	-997	-999	-999	-999.1
$V_{B\text{-}PC}$	126868	206833	-6686	-2029	15278
$V_{B\text{-}NW}$	1545	3272	-993	-982	-725.8

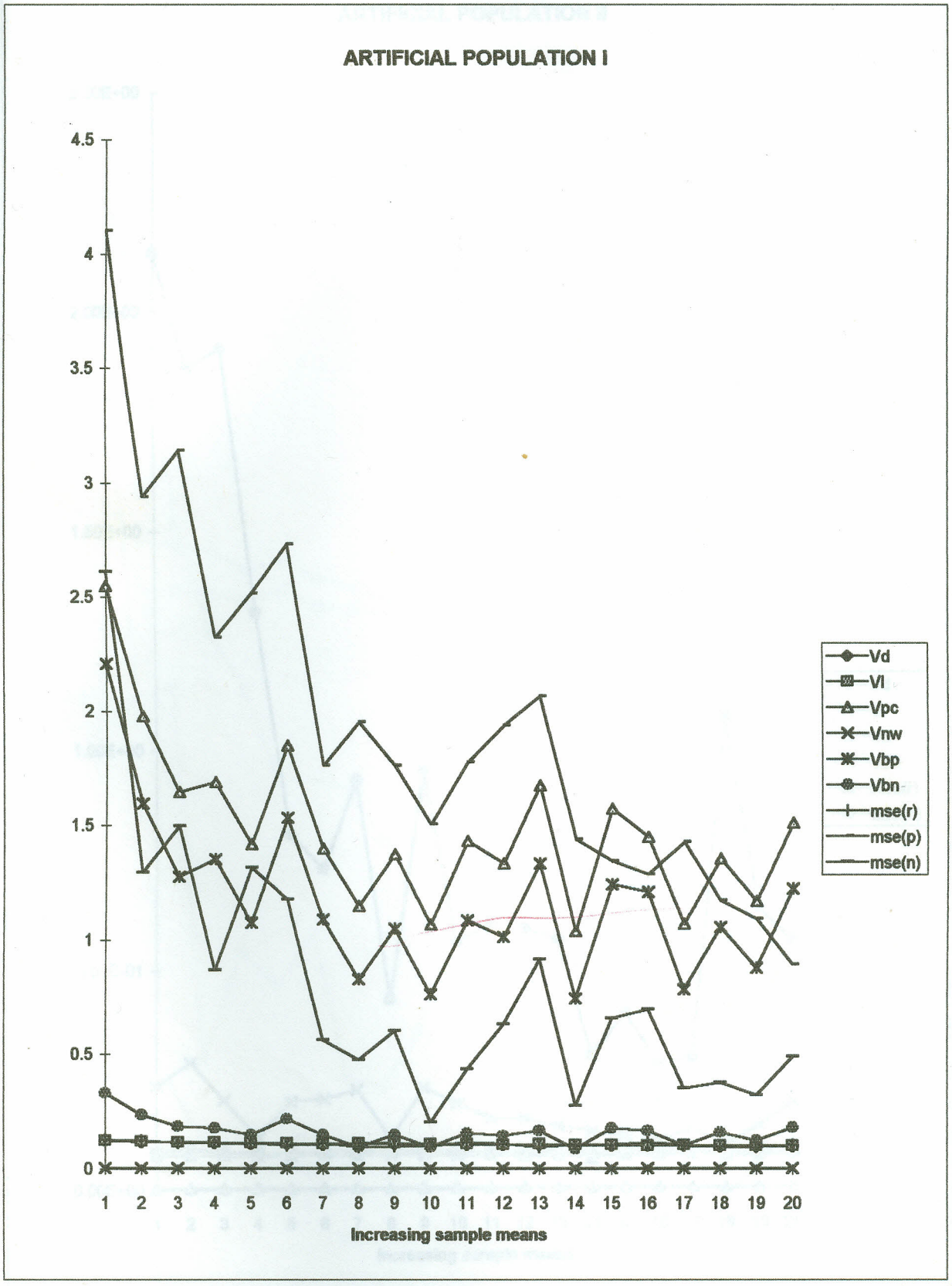
ARTIFICIAL POPULATION I
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Increasing sample means
Vd
Vl
Vpc
Vnw
Vbp
Vbn
mse(r)
mse(p)
mse(n)

Figure I

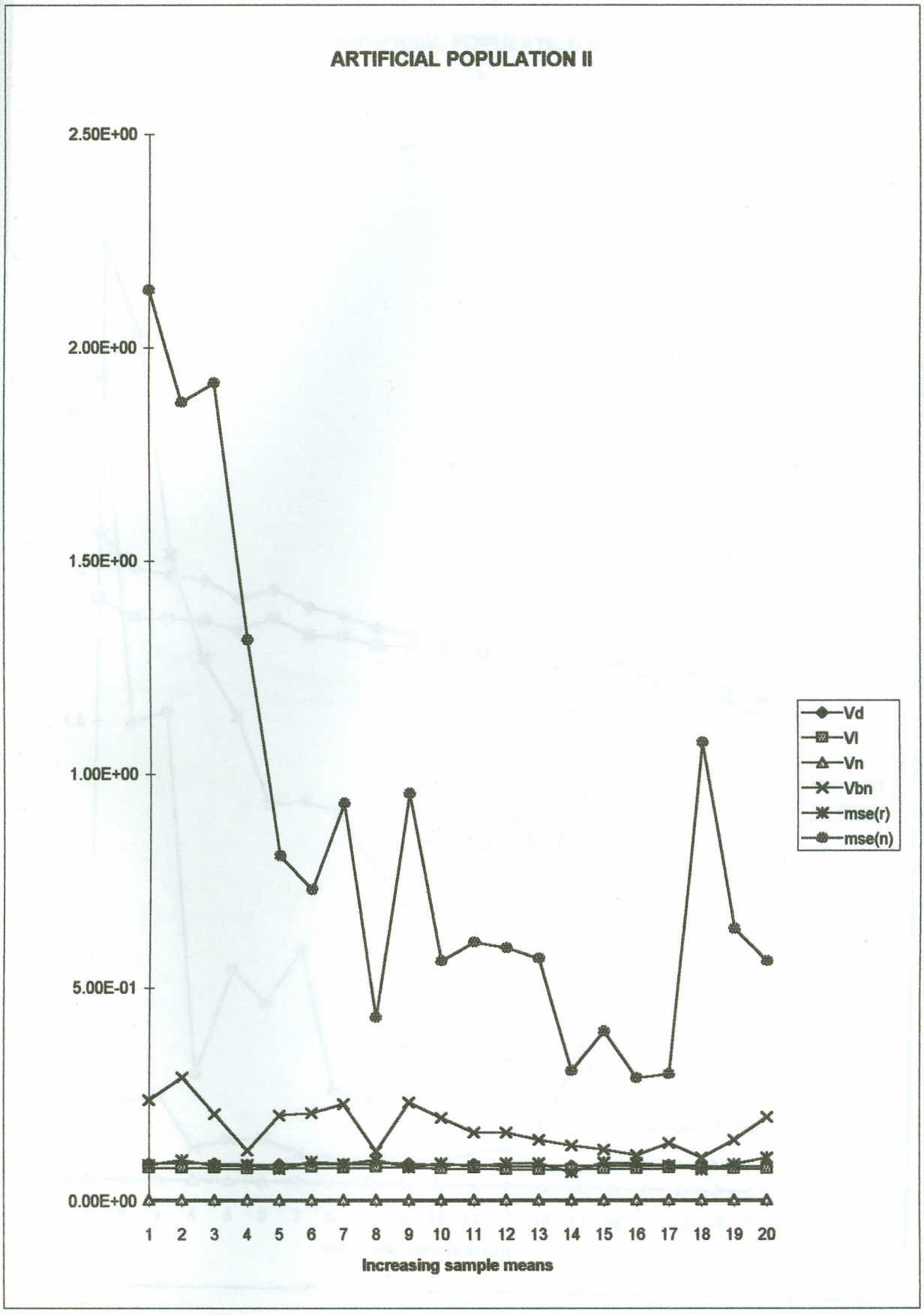
ARTIFICIAL POPULATION II
2.50E+00
2.00E+00
1.50E+00
1.00E+00
5.00E-01
0.00E+00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Increasing sample means
Vd
Vl
Vn
Vbn
mse(r)
mse(n)

Figure 2

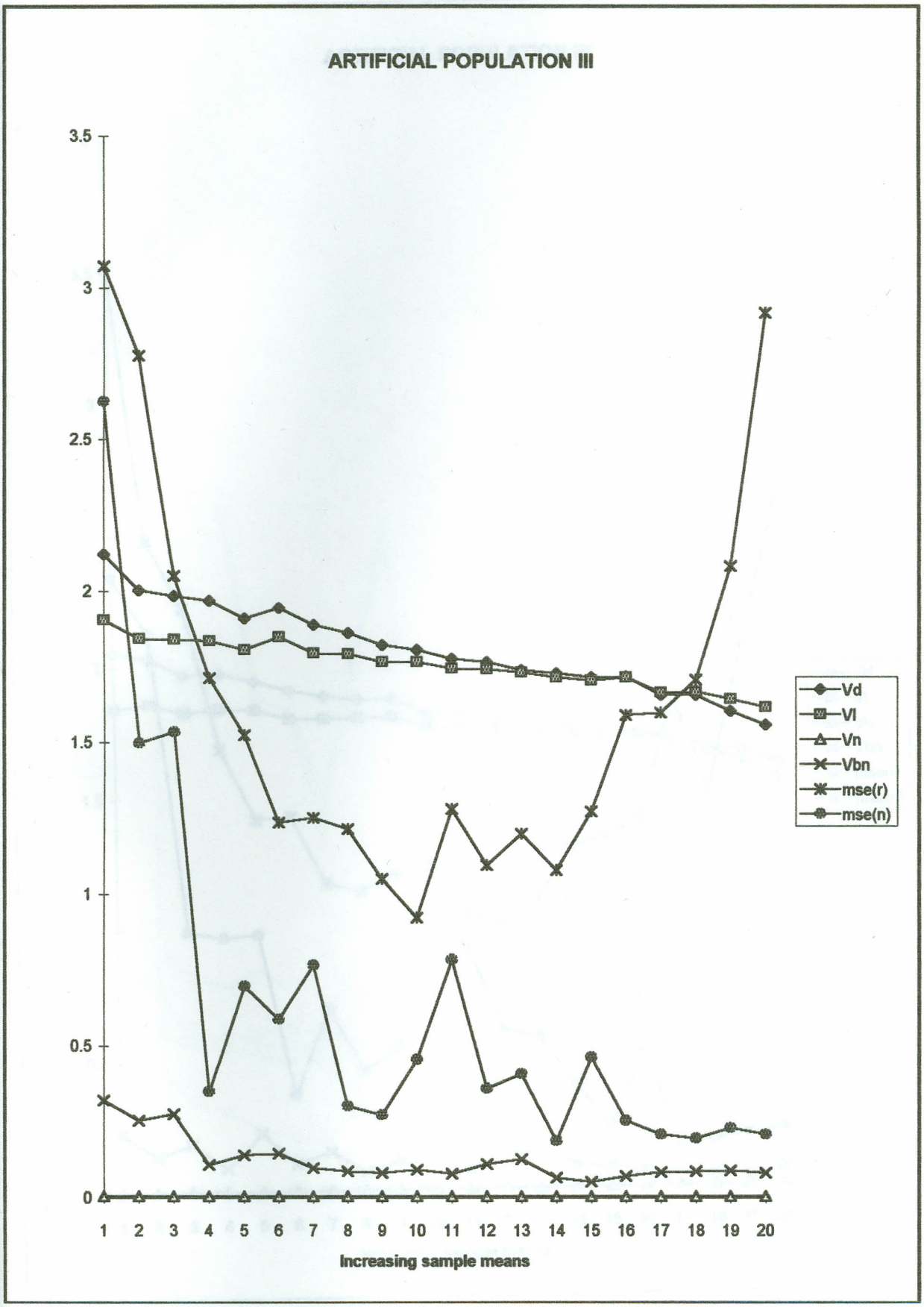
ARTIFICIAL POPULATION III
3.5
3
2.5
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Increasing sample means
Vd
Vl
Vn
Vbn
mse(r)
mse(n)

Figure 3

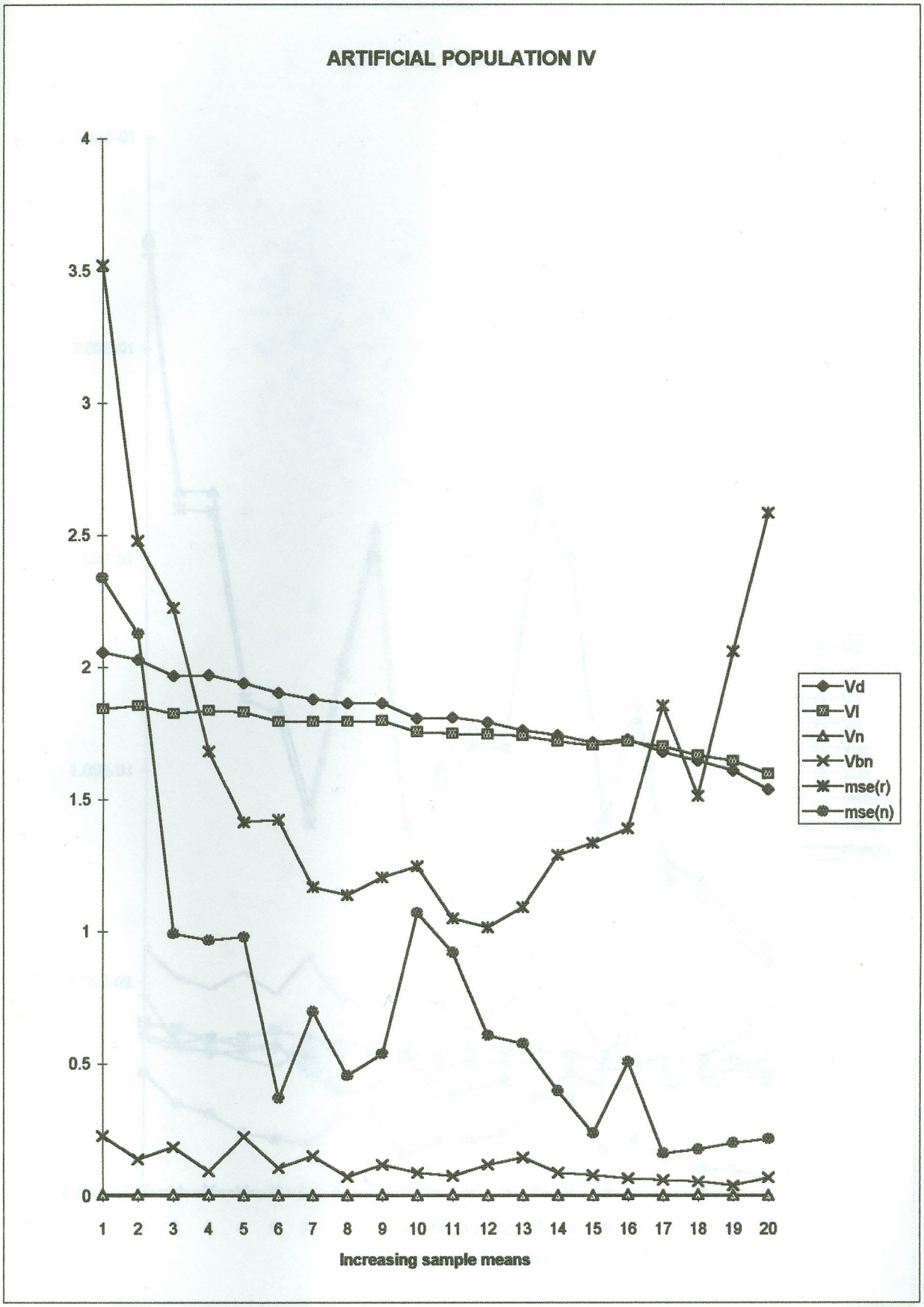
ARTIFICIAL POPULATION IV
4
3.5
3
2.5
2
1.5
1
0.5
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Increasing sample means
Vd
Vl
Vn
Vbn
mse(r)
mse(n)

Figure 4

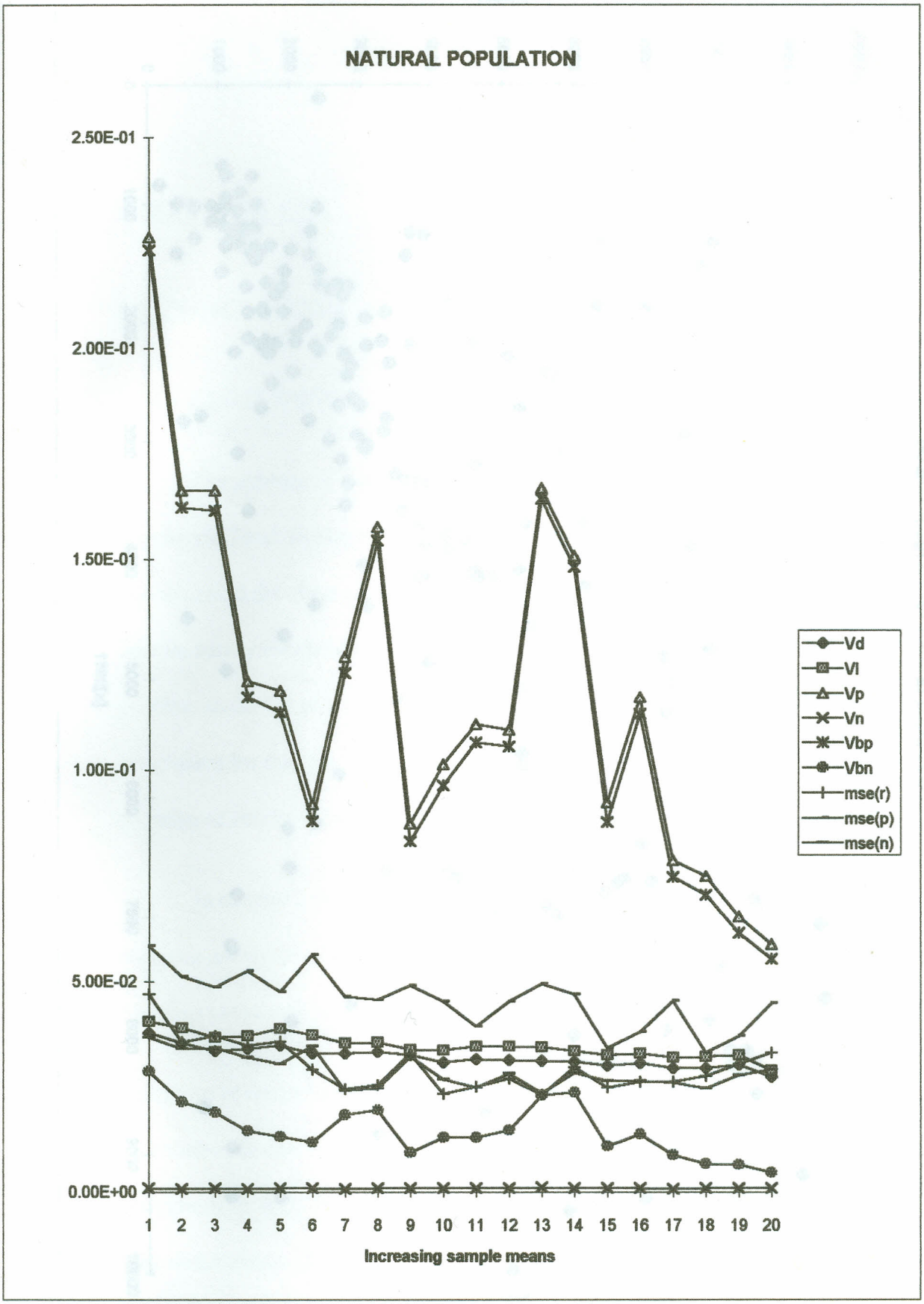
NATURAL POPULATION
2.50E-01
2.00E-01
1.50E-01
1.00E-01
5.00E-02
0.00E+00
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Increasing sample means
Vd
Vl
Vp
Vn
Vbp
Vbn
mse(r)
mse(p)
mse(n)

Figure 5

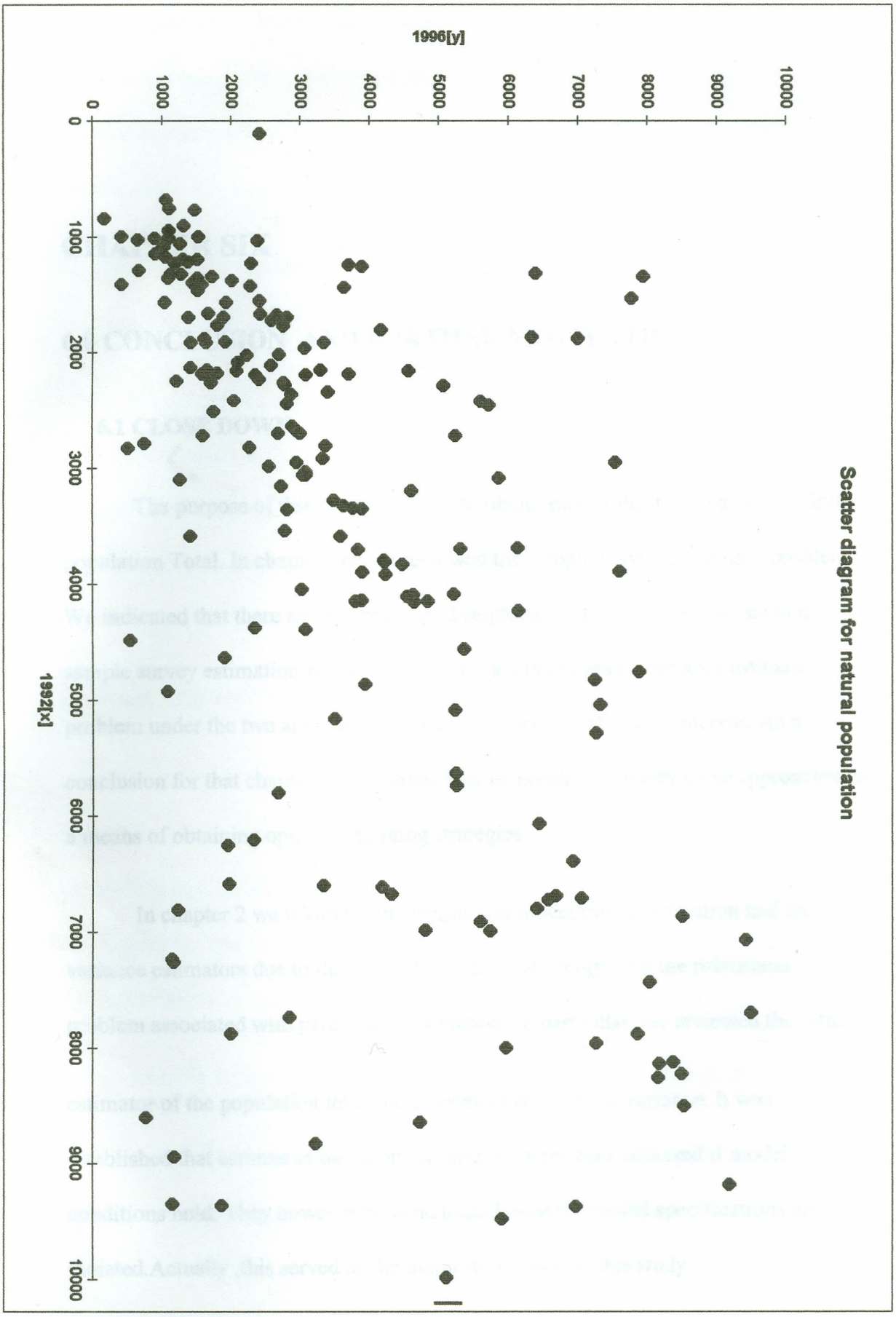
Scatter diagram for natural population
1996[y]
1992[x]
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

Figure 6

CHAPTER SIX

6.0 CONCLUSION AND FURTHER MOTIVATION.

6.1 CLOSE DOWN

The purpose of this study has been to obtain more robust estimators for finite population Total. In chapter one we reviewed the sample survey estimation problem. We indicated that there are two main, philosophically different approaches to the sample survey estimation problem. Hence as, a consequence, variance estimation problem under the two approaches being motivated by different concerns. As a conclusion for that chapter we remarked that its possible to marry these approaches as a means of obtaining optimal sampling strategies.

In chapter 2 we reviewed the parametric procedure of estimation and the variance estimators due to this procedure. Here we brought out the robustness problem associated with parametric estimation. In particular, we reviewed the ratio estimator of the population total and estimators of its error variance. It was established that estimators based on this approach are only unbiased if model conditions hold. They however become biased once the model specifications are violated.Actually ,this served as the major motivation to this study.

In chapter 3 and 4 we endeavoured to search for more robust estimators. We considered nonparametric estimation of population Totals as a means of obtaining variance estimators that have good robustness properties under more general model specifications. We suggested new estimation procedures by use of kernel smoothers. We showed that estimators based on this procedure are robust to model conditions violation. Our claim could only be established through an empirical study.

We dedicated chapter 5 to Empirical study of our estimators. Findings of this study matched well with our theory. Nonparametric estimators proved to be more robust to model conditions violation. This seems as a major step in solving the robustness problem that has haunted sample survey researchers for along time. However , caution must me expressed as we would indicate in the section below.

6.2 FURTHER MOTIVATION.

In our study for the asymptotic properties of our estimators we imposed some restrictions which may not be realistic. One obvious and most disturbing assumption we made is that the X_i's are equispaced , clearly this is not always the case ~~the case~~. No wonder for the natural population the nonparametric estimators were highly rivalled by the ratio estimators.

Another restriction we imposed on the model conditions is that the $Cov(y_i, y_j) = 0$ for $i \neq j$ .It would be interesting to study the behaviour of these estimators without this restriction. This is still an open area for research.

REFERENCES

1) **Dorfman, A.H (1994)** . Nonparametric regression for estimating Totals in finite populations submitted to Biometrika

2) **Efron, B. (1987)** . Better Bootstrap confidence intervals (with discussions). Journal of American Statistical; Association, 82, 171-200

3) **Gasser, T and Englel.). (1990)** . The choice of weights in Kernel Regression Estimation. Biometrika, 97, 377-381.

4) **Gasser, T. And Muller, H.G (1979)** . Kernel Estimation of Regression Functions. Smoothing Techniques for curve Estimation, eels. T. Gasser and M. Rosenblatt, New York; springer Verlag, 23-68.

5) **Horn, S.D, Horn, R.A and Duncan, D.B., (1975).** Estimating Heteroscedistic variances in linear Models. Journal of the American statistical association 78, 776-807.

6) **Kovar, J.G, Rao, J.N.K., and Wu,CF.J., (1980)** . Bootstrap and other methods to measure Errors in survey Estimates. The Canadian journal of statistics, 16, 25-45

7) **Morgan T.J, Elements of simulation (1984)**. Chapman and Hall, New York, London.

8) **Muller, H.G and StadtMuller, U., (1987)** . Estimation of Hetroscedasticity in regression Analysis. Annals of statistics 15, 610-635.

9) **Nadaraya E.A (1964)** . On Estimating regression. Theory of Probability Application 9, 141-142.

10) Odhiambo, R.O (1991). A study of the Robustness Properties of the variance estimators of the ratio estimator, unpublished M.Sc Dissertation, Kenyatta University.

11) **Odhiambo, R.O (1996)** .Robust Variance Estimation for finite population sampling. Unpublished Ph.D. thesis Kenyatta University.

12) **Priestly, M.B and Chao M.T; (1972)** . Non-Parametric function fitting , journal of the Royal Statistical Society, B, 34, 385-392.

!3) **Yang, M.C.K and Robinson, D.H (1996)** . Learning and understanding statistics by computer. Computer series, vol. 4

14) **Royal, R.M (1971)** . Linear Regression Models infinite population sampling theory. In foundations of statistical Inference (V.P, Godambe and D.A, sprotts, eds.) Holt, Rinhart and Winston, Toronto.

15) **Royall ,R.M and Cumberland. W.G (1978)** . Variance Estimation infinite population sampling. Journal of the American statistical Association, 73, 351-358

16) **Royall, R.M and Cumberland. W.G., (1985)** .Conditional Convergence Properties of Finite Population confidence intervals. Journal of the American Statistical Association, 80, 355-359.

17) **Royall, R.M and Cumberland. W.G., (1981)** .An Empirical study of the Ratio Estimator and Estimators of its variance.

18) **Royall, R.M and Elberhardt, K.R. (1975)** . Variance Estimates for the ratio Estimator. Sankhya Ser. C, 37, 43-52.

19) **Royall, R.M and Herson, J., (1973)** . Robust Estimation infinite population sampling, I, Journal of the American Statistical Association, 68, 880-889.

20) **Silver-Man, B., (1988)** . Some aspects of spline smoothing approach to Non-parametric curve fitting. Journal of the Royal statistical Society. A, 139, 183-195.

21) **Silverman, B. (1986)**.Density Estimation Chapman and Hall, London.

22) **Smith J.M.F and Njenga. E.G, (1983)** Robust Model-Based methods of Analytical Surveys. Statistics Canada, 18,2,187-208.

23) **Wafula, C. (1988)** . Some contributions of Variance Estimation in Sample Surveys. Unpublished Ph.D. thesis, university of Kent at Canterbury.

24) **Wahba, G. (1975)** Smoothing Noisy Data in Spline Functions. Numerical Mathematics 24, 386-393.