

A Design Unbiased Variance Estimator of the Systematic Sample Means

Festus A. Were, George Orwa, Romanus Odhiambo

Jomo Kenyatta University of Agriculture and Technology, School of Mathematical Sciences, Nairobi, Kenya

Email address:

Weref87@gmail.com (F. A. Were), orwa@fscjkuat.ac.ke (G. Orwa), romanusotieno@yahoo.com (R. Odhiambo)

To cite this article:

Festus A. Were, George Orwa, Romanus Odhiambo. A Design Unbiased Variance Estimator of the Systematic Sample Means. *American Journal of Theoretical and Applied Statistics*. Vol. 4, No. 3, 2015, pp. 201-210. doi: 10.11648/j.ajtas.20150403.27

Abstract: Systematic sampling is normally used in surveys of finite populations because of its appealing simplicity and efficiency. When properly applied, it can reflect stratification in the population and thus can be more precise than SRS. In systematic sampling technique, the sampling units are evenly spread over the whole population. This sampling scheme is very sensitive to correlation between units in the entire population. A positive autocorrelation reduces the precision while a negative autocorrelation will improve the precision compared to simple random sampling. The limitation of this sampling method is that, it is not possible to estimate the design variance that is unbiased. This study proposes an estimator for the design variance based on a non-parametric model for the population using local polynomial regression as the estimation technique. The non-parametric model is more flexible that it can hold for many practical situations. A simulation study is performed to enable the comparison of the efficiency of the proposed estimator to the existing ones. The performance measures used include: Relative Bias (RB) and Mean Square Error (MSE). From the simulation results, it can be seen that local polynomial estimator based on nonparametric model is consistent and design unbiased for the variance of systematic sample mean. The simulation study gave smaller values for the relative biases and mean squared errors for proposed estimator.

Keywords: Systematic Sampling, Local Polynomial Regression, Non-Parametric Model, Design Variance

1. Introduction

1.1. Background of the Study

Systematic sampling is a probability sampling technique where a sample is obtained by selecting every k^{th} element of the population where k is an integer greater than 1. The first number of the sample must be selected randomly from within the first k elements. The selection is done from an ordered list. It is a popular method of selection especially when units are many and are serially arranged from 1 to N . Suppose that N the total number of units is a multiple of the required sample size n and an integer k , such that $N = nk$, a random number is selected between 1 and k .

A sample which comprises of the first unit is selected randomly and every k^{th} unit, until the required sample size is obtained. The interval k divides the population into groups. In this method we are selecting one cluster of units with probability $\frac{1}{k}$. Since the first number is drawn at random from 1 to k , each unit in the supposedly equal clusters gets the same probability of selection $\frac{1}{k}$.

Systematic sampling is widely used in surveys of finite populations such as forest where other sampling scheme cannot be easily applied. This is due to its appealing simplicity and efficiency. When properly applied, the method picks up any obvious or hidden stratification in the population and thus can be more precise than simple random sampling. Also, systematic sampling is easy to implement, thus reducing costs.

Since a systematic sample can be regarded as a random selection of one cluster, it is not possible to give an unbiased or even consistent design based estimator of the variance and this is the challenge faced by researchers who apply it in practice. There are two approaches that are proposed to solving this problem. One is to postulate a superpopulation model characterizing the population structure and to obtain model-unbiased estimators of variance Cochran(1977). The superpopulation model is used to describe the relationship between the auxiliary variable and the study variable. This approach may not yield satisfactory results because the model assumption is usually hard to verify in practice and the

unbiased estimators of variance can be sensitive to model assumption. The second approach is to take additional observations (supplementary sample) typically of smaller size than the first sample via simple random sampling Zinger(1980) or systematic sampling.

Nonparametric regression is motivated by the fact that it provides a flexible way of studying the relationships between variables and also results in good estimators thus increasing their efficiency compared to estimators obtained using designed based approaches.

In this framework, this study is concerned with the estimation of the variance of systematic sample mean using a nonparametric approach (local polynomial regression technique) with the aid of a superpopulation model. It also offers the methodology to study the convergence properties of the proposed estimator.

1.2. Literature Review

Zinger(1980) pursued an approach defined as partially systematic sampling in which he obtained an unbiased estimator for the variance of systematic sample mean, however, his proposed estimator faced the challenge of not being able to prove for non negative variance except for the case of $\beta = \frac{1}{2}$. Wu(1984) suggested difference estimator to tackle the problem faced by Zinger (1980) which was non negative for all $\beta \geq \frac{K-1}{2K}$. Rana (1989) following the work done by Zinger(1980) proposed a different estimator for variance of systematic sample mean that was unbiased and non negative for all Values of β .

Wolter (2007) gives more comprehensive review on eight biased variance estimators and guidelines for choosing among them is given. The above variance estimation procedures are conditional on the design. In other words, they are design-based in the way that the finite population is treated as fixed.

There also exist some model-based variance estimators where the populations are considered random realizations from a super population model. Montanari. G and Bartolucci. F(1998) came up with a model based variance estimator using OLS which was approximately unbiased for the variance of systematic sample mean under linear super-population model. However this estimator lacked some accuracy and efficiency due to a higher contribution of the bias if the systematic component of the super-population is not linear. Montanari. G and Bartolucci. F(2006) later proposed a new class of unbiased estimators of the variance of systematic sample mean that included some simple nonparametric estimators under the assumption that the population follows a super population model that satisfied some mild assumptions. They showed that, the estimator based on local polynomial regression as the estimation technique under the assumption that the population follows a linear trend and the errors are homoscedastic and uncorrelated. The simulation results showed that the LPR estimator performed better in terms of relative bias and mean square error as they all had small values.

X. Li and J. Opsomer (2010) and Ayora. O(2014) also using the work that was proposed by Later Montanari. G and

Bartolucci. F(2006) considered a broadly applicable model for the data, in which both the mean and the variance are left unspecified subject only to smoothness assumptions. They then came up with a model-based nonparametric variance estimator, in which both the mean and the variance functions of the data are estimated nonparametric ally using local polynomial regression as the smoothing technique. In their simulation experiment performed, it was evident that this estimator perform better giving small relative bias and mean square errors as compared to the other classical estimators discussed in Wolter (2007).

This study considers a more applicable model in which the mean function is unspecified but the variance function is homoscedastic. The researcher proposes a model-based nonparametric estimator using local polynomial regression as the smoothing technique for variance of systematic sample means. It will later shows that the estimator proposed is model consistent for the design variance of the survey estimator, subject to the population smoothness assumptions.

1.3. Statement of the Problem

Variance estimation for systematic sample mean still remains an issue that has not been addressed as only estimation procedures which are not so robust have been proposed. In view of this, exact computation of a robust estimator for variance in the systematic sample mean or total mean still remains an open area of research.

2. Methodology

2.1. Introduction

In this study, the researcher will first review Systematic sampling and the existing estimators of variance of systematic sample mean. Assumption used in developing the proposed estimator will be reviewed, then propose an estimator based on local polynomial regression using a nonparametric super population model. It will also provide the proof for the consistency of the proposed estimation and lastly compare the performance of the proposed estimator through a simulation study.

In the current study, let $Y_j, j = 1, 2, \dots, N$ be finite population measurements of size N representing some survey characteristics and $X_j, j = 1, 2, \dots, N$ be a vector of auxiliary variables which is considered fixed. Let k be the sampling interval and $P(.) = \frac{1}{K}$ be the probability of each element in the sample being selected from the population, then the systematic sample will consist of the observations $i, i + k, \dots, i + (n - 1)k$ where n is the sample size and the i^{th} systematic sample will be $y_i, \dots, y_{(n-1)(k+i)}$. Let \bar{Y} be the population mean, and \bar{y}_i be the i^{th} systematic sample mean, then, the study is interested in estimating the variance of \bar{y}_i which is defined by equation (3). To estimate this variance, the study uses the local polynomial regression function $\hat{\mu}_i = e_1^T (X_{ij}^T W_{ij} X_{ij})^{-1} X_{ij}^T W_{ij} Y_i$ discussed in Wand and Jones (1995) estimated from the i^{th} where $W_{ij} = \text{diag} \left(K \left(\frac{x_i - x_j}{h} \right) \right)$ for $i = 1, 2, \dots, n$ with h being the

smoothing parameter and $K(\cdot)$ a kernel function.

2.2. Review of Systematic Sampling

Suppose that the population size is N units and the study variable $Y_i, i = 1, 2, \dots, N$. Then the population mean is given as

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i \quad (1)$$

To draw a systematic SYS we first sort the population using some criterion. For example we can sort by one of the auxiliary variables in X_i . If the study variable Y and auxiliary variable X are related through a certain function, sorting by X may provide a good spread of Y 's so that a systematic sample can pick up hidden structures in the population. If we sort the population by some criterion that is not related to Y at all, for instance sort by a variable Z which is independent of Y , then we will have a random permutation of the population. In this case systematic sampling is equivalent to SRSWOR. After sorting the population we randomly choose an element from the first k ones say the i^{th} one, then, this systematic sample consists of the observations $i, i+k, \dots, i+(n-1)$. Thus Systematic sampling amounts to the selection of a single complex sampling unit that constitutes the whole sample. A systematic sample is a SRS of one cluster unit from a population of k cluster units. Table 1 illustrates this procedure. Each column corresponds to a possible sample systematic sample. The interval k divides the population into n rows of k elements each. One element from each row is selected and each element has the same location on each row.

Table 1. Composition of k systematic Samples.

Sample number						
1	2	3	i	k
y_1	y_2	y_3	y_i	y_k
...
...
$y_{(n-1)(k+1)}$	$y_{(n-1)(k+2)}$	$y_{(n-1)(k+3)}$	$y_{(n-1)(k+i)}$	y_{nk}
\bar{y}_1	\bar{y}_2	\bar{y}_3	\bar{y}_i	\bar{y}_k

The population mean is estimated by the i^{th} sample mean given as

$$Y_s = \frac{1}{n} \sum_{i=1}^n Y_i \quad (2)$$

The design based variance for this mean was first derived by Madow and Madow (1944) and is give by

$$Var_p(\bar{Y}) = \frac{1}{k} \sum_{i=1}^n (\bar{Y}_s - \bar{Y})^2 \quad (3)$$

But there is no unbiased design based estimate of $Var_p(\bar{Y})$ for the general variable Y . Among the eight estimators evaluated by Wolter (2007) as the estimates of $Var_p(\bar{Y})$ we look at the three main ones which are used in practice. One of the approaches is to treat the systematic sample as if it had been obtained by SRS. This estimator is defined by

$$\hat{V}_{SRS} = \frac{1-f}{n} \frac{1}{n-1} \sum_{i \in s} (Y_i - \bar{Y}_s)^2 \quad (4)$$

where

$$f = \frac{n}{N}$$

The other two estimators are based on pairwise differences and are recommended in Wolter (2007) as the best general purpose estimators of $Var_p(\bar{Y})$, these estimators are defined as

$$\hat{V}_{OL} = \frac{1-f}{n} \frac{1}{2(n-1)} \sum_{i=2}^n (Y_i - Y_{i-1})^2 \quad (5)$$

Which uses all successive pairwise differences and hence uses OL. The other estimator is defined by

$$\hat{V}_{NO} = \frac{1-f}{n} \frac{1}{N} \sum_{i=1}^n (Y_{2i} - Y_{2i-1})^2 \quad (6)$$

This takes on successive NO. The three estimators are designed biased for $Var_p(\bar{Y})$ in general.

The first estimator \hat{V}_{SRS} is viewed suitable when the ordering of the population is thought to have no effect on \bar{Y}_s or is considered as a conservative estimator when the ordering is related to the variable Y . However, as discussed by X. Li and Opsomer (2010), the unbiasedness of \hat{V}_{SRS} for uninformative ordering only holds if one averages over samples and over orderings of the population, so not design strictly design unbiasedness.

The bias of \hat{V}_{SRS} for a fixed ordering of the population can be larger and either positive or negative. The last two estimators tended to have smaller bias in the simulation experiments discussed in Wolter (2007). To obtain an unbiased estimate of $Var_p(\bar{Y})$, the following three designs have to be considered.

1) Multiple systematic sampling using a randomly determined starting position for each systematic sampling stage.

2) Systematic stratified - Two or more systematic samples (each with a different random start position) are taken within each stratum

3) Two stage sampling where the sub samples are collected according to systematic sampling design

4) Complementary systematic and random sampling where a systematic sample is supplemented by a random sample of size m from the remaining population units.

2.3. Review of Local Polynomial Regression

Nonparametric regression has become a rapidly developing and growing field of statistics. Nonparametric approaches to regression are flexible and data-analytic ways to estimate the regression function without the specification of a parametric model, that is, to let data find a suitable function that well explains the data. The Local modeling techniques with kernel weights provide a basic and easily understood nonparametric approach to regression. Local polynomial regression is a generalization of kernel regression since the regression function at a point x in kernel regression is estimated by a locally weighted average, which can be shown to correspond to fitting degree zero polynomials, that is, Nadaraya Watson estimator. Wand and Jones [1995] give a clear explanation of kernel smoothing including local polynomial regression.

Local polynomial regression has several advantages of other nonparametric approaches. This particular method is readily adapted to highly clustered, random, fixed designs and close to uniform designs, and on both interiors and boundaries. Local polynomial regression estimators don't have boundary bias, that is, they adapt automatically to the boundary effect, and thus there is no need for any modifications for correcting the large bias problem at the boundary.

Local polynomial estimators have high mini-max efficiency among the class of linear smoothers, including those ones produced by kernel smoothers and spline technique, both in the interior and at the boundary points. Fan [1992], Fan [1992] discusses in detail the local linear fit in comparison with the local constant fit and shows that the local linear regression smoothers have the desirable mean squared error (MSE), the design adaptation property, no boundary effects, and high asymptotic mini-max efficiency properties. Fan [1992] in their work were able to show that, local linear regression estimator adapts automatically to estimation at the boundary and they give expressions for the conditional MSE and mean integrated squared error (MISE) of the estimator. Wand and Jones [1995] extend the results of Fan [1992] on asymptotic bias and variance to the case of local polynomial estimators. Fan and Gijbels [1996] in their work, they emphasizes on methodologies with a particular focus on applications of local polynomial modeling techniques to various statistical problems including survival analysis, least square regression, nonlinear time series, robust regression, generalized linear models. Breidt and Opsomer [2000] apply local polynomial regression to model-assisted survey sampling.

One of the important issues in nonparametric regression is the choice of the smoothing parameter (bandwidth). In most scenarios, bandwidth is often selected subjectively by eye, but there are other situations where it is necessary to have the bandwidth automatically selected from the data. In data-driven smoothing parameter selection, all methods try to estimate the optimal bandwidth value that minimizes the mean squared error (MSE) at a point x or the MSE over all values of x . Most bandwidth selection methods attempt to find a value for the MISE (Mean integrated squared error)-minimizing bandwidth, and thus those are called global bandwidth selection methods. Cross-validation (CV) technique is a well-known method of optimizing the bandwidth, using the leave one-out prediction technique. However, the smoothing parameter computed by the CV method is very variable and normally tends to under-smooth in practice that is, the chosen bandwidths tend to be very small. In the case of linear smoothers, calculation of the CV method is easy since the expression of the leave-one-out predictor is a linear function of the complete data predictor. Another approach to bandwidth selection is to estimate MISE directly based on the data. This method estimates the variance and the bias of the estimator, thus it minimizes the estimated MISE with respect to the bandwidth. This "plug-in" method is used mostly in kernel regression and local polynomial regression. Plug-in technique gives more stable performance. The theory, the choice of a global variable bandwidth based on the plug-in procedure for the local linear smoothers was discussed by Fan

[1992].

Wand and Jones [1995] developed a simple direct plug-in bandwidth selector for local linear regression that is seen to work well in practice for a wide variety of functions and is shown to have appealing theoretical and practical properties. Fan and Gijbels [1995] propose a data-driven variable bandwidth selection procedure based on a residual squares criterion and show that local polynomial fitting using the variable bandwidth has spatial adaptation properties.

2.4. Trade-Off Between Bias and Variance

The choice of the bandwidth, h is of crucial importance tool for local polynomial regression. Smaller bandwidth results in less smoothing while larger bandwidth oversmooths the curve. There is a trade-off between variance and bias. Large values of bandwidth will reduce the variance since more points will be included in the estimate. However, as the bandwidth increases, the average distance between the local points and x_0 will increase. This can result in a larger bias in the estimator. A natural way to choose a bandwidth and balance this trade-off is by minimizing the mean square error (MSE) Fan and Gijbels [1996]. Therefore one should choose an optimal bandwidth to minimize MSE so as to balance the trade-off between the bias and variance.

In addition to selecting the optimal bandwidth, it is also important to select the appropriate order of polynomial to fit as when choosing a bandwidth, there is also trade-off between bias and variance. Higher order polynomials allow for precise fitting meaning the bias will be small but the order increases, so does the variance, but this increase is not constant. The asymptotic variance of $\hat{\mu}(x)$ only increases whenever the order goes from odd to even. There is no loss when going from $p = 0$ to $p = 1$ but going from $p = 1$ to $p = 2$ will increase asymptotic variance. This suggests only considering odd-ordered polynomials since the gain in bias appear to be free with no associated cost in variance Fan and Gijbels [1996], Wand and Jones [1995].

Fan and Gijbels [1996] suggests an adaptive method of choosing the correct order of polynomial based on local factor, allowing p to vary for different points in the support of data. The resulting estimator has the property of being robust to bandwidth. This means that if the chosen bandwidth is large is too large, a higher order polynomial is chosen to better model the boundaries of the data. If the chosen bandwidth is too small, a lower order polynomial is chosen to help make the estimate numerically stable and reduce the variance. Therefore one should select an appropriate bandwidth and order of the polynomial to balance the trade-off between the bias and variance in order to give an appropriate amount of smoothing.

2.5. Assumptions Used in Developing the Estimator in the Current Study

To prove the convergence property of the proposed estimator, the study adopts a theoretical framework in which both the population size N , the sample size n and the sampling interval tend to infinity. A sample is the selected as described in section 3.1

We make the following additional assumption on the study variable, the design and the smoothing parameter.

A1: The errors ε_j are independent with a mean of zero and variance δ_{NP}^2 and compact support, uniformly for all N

A2: For each N , we consider the x_j 's as fixed with respect to the superpopulation model. The x 's are independent and identically distributed.

$F(x) = \int_{-\infty}^x f(t)dt$, where $f(\cdot)$ is the density function with compact support $[a_x, b_x]$ and $f(x) > 0$

For all $X \in [a_x, b_x]$

A3: The sample size n and the sampling interval k are positive integers with $nk = N$. It is assumed that $n, N \rightarrow \infty$ and allow $k = O(1)$ or $k \rightarrow \infty$

A4: As $n \rightarrow \infty$, it is assumed $h^* \rightarrow 0$ and $nh^* \rightarrow \infty$ where h^* is h_m

A5: The kernel function $K(\cdot)$ is a compactly supported, bounded, symmetric kernel with $\int u^{p+1} k(u)du = u_{p+1}(k)$ assume that $u_{p+1}(k) \neq 0$

A6: The $(p+1)^{th}$ derivative of the mean function $\mu(\cdot)$ exists and is bounded on $[a_x, b_x]$

2.6. The Proposed Estimator

This study employs a model based approach in which a consistent variance estimator of systematic sample means is proposed under a nonparametric model using local polynomial regression as the method of estimation. In the estimator the bias correction term considered by Montanari, G and Bartolucci, F(1998),(2006) is not considered here and also the variance function of the model is assumed to be Homoscedastic. In the estimation let $X = (X_1, X_2, X_3, \dots, X_N)$ be a vector of univariate auxiliary variable, then, the non parametric superpopulation model is given by

$$Y = \mu(x_j) + \varepsilon_i \quad (7)$$

where

$$E_p(Y_j|X_j) = \mu(x_j)$$

$$Var_{NP}(\varepsilon) = \delta_{NP}^2$$

and

$$diag(w_1, w_2, w_3, \dots, w_N) = \delta_{NP}^2 \Sigma$$

Now the design variance in equation (3) can be written as

$$Var_p(\bar{Y}) = \frac{1}{kn^2} Y^T N Y$$

where

$N = M^T H M$ with $H = I_n^T \otimes 1_k$ here \otimes is the Kronecker product and 1_k is a column vector of 1's of length k .

Let $\mu(\cdot)$ be a continuous and bounded function and define $\mu = (\mu(x_1), \dots, \mu(x_N))$, it is assumed that w_j 's are bounded and positive where $j = 1, 2, 3, \dots, N$

Under model (7), the expected value of $Var_p(\bar{Y}_s)$ is

$$E_{NP}(Var_p(\bar{Y}_s)) = \frac{1}{kn^2} \mu^T N \mu + \frac{1}{kn^2} tr(N \Sigma) \delta_{NP}^2 \quad (8)$$

To estimate $E_{NP}(Var_p(\bar{Y}_s))$, the following local polynomial regression estimator for variance of systematic sample means is proposed

$$\hat{V}_{NP}(\bar{Y}_s) = \frac{1}{kn^2} (\hat{\mu}_i^T N \hat{\mu}_i) + \frac{1}{kn^2} tr(N \Sigma) \delta_{NP}^2 \quad (9)$$

Here

$$\hat{\delta}_{NP}^2 = \frac{(Y_i - \hat{\mu}_i)^T \Sigma_i^{-1} (Y_i - \hat{\mu}_i)}{n}$$

and

$$\hat{\mu}_i = \hat{\mu}(x_1), \dots, \hat{\mu}(x_n)$$

Where $\hat{\mu}(x_j)$ is the local polynomial regression estimator obtained from the i^{th} sample

$$\hat{\mu}_i = e_1^T (X_{ij}^T W_{ij} X_{ij})^{-1} X_{ij}^T W_{ij} Y_i$$

Where e is the $(p+1) \times 1$ vector of the identity matrix having 1 in the first entry and other entries 0. p denotes the degree of local polynomial regression.

$$X_{ij} = \begin{bmatrix} 1 & (x_1 - x_j) & \dots & (x_1 - x_j)^p \\ \vdots & & \ddots & \vdots \\ 1 & (x_n - x_j) & \dots & (x_n - x_j)^p \end{bmatrix}$$

$$W_{ij} = diag(K(\frac{x_i - x_j}{h})) \quad i = 1, 2, \dots, n$$

Where h is the smoothing parameter and $K(\cdot)$ the kernel function.

In developing the current estimator, reference is made to Wand and Jones(1995) version of the local polynomial regression estimator.

Under assumption A1-A6, the design variance is model consistent for the anticipated variance in the sense that

$$Var_p(\bar{Y}_s) = E_{NP}(Var_p(\bar{Y}_s)) = O\left(\frac{1}{\sqrt{N}}\right) \quad (10)$$

And the local polynomial variance estimator is model consistent for the anticipated variance for the design variance in the sense that

$$\hat{V}_{NP}(\bar{Y}_s) - E(Var_p(\bar{Y}_s)) = O_p(h_m^{p+1}) + O_p\left(\frac{1}{\sqrt{nh}}\right) \quad (11)$$

$$\hat{V}_{NP}(\bar{Y}_s) - Var_p(\bar{Y}_s) = O_p(h_m^{p+1}) + O_p\left(\frac{1}{nh_m}\right) \quad (12)$$

And the best bandwidth should satisfy the condition

$$h_m^{p+1} = O\left(\frac{1}{nh_m}\right)$$

which leads to $h_m = C_n^{-\frac{1}{2(p+3)}}$ the usual optimal rate for local polynomial regression. The bandwidth selection procedures such as plug-in or cross validation methods can be used in this case. This study provides the proof for the equation (11). The proof for equations 10 and 12 see X. Li(2006)

2.7. Proof of Equation (11)

From equation 11,

$$\hat{V}_{NP}(\bar{Y}_s) - \text{Var}_p(\bar{Y}_s) = \frac{1}{kn^2} (\hat{\mu}_i^T N \hat{\mu}_i - \mu^T N \mu) + \frac{1}{kn^2} \text{tr}(N \Sigma) (\delta_{NP}^2 - \delta_{NP}^2) = \# + \# \quad (13)$$

The first term on the right hand of equation 13 can be written as

$$\begin{aligned} \frac{1}{kn^2} (\hat{\mu}_i^T N \hat{\mu}_i - \mu^T N \mu) &= \frac{1}{kn^2} (\hat{\mu}_i - \mu)^T N (\hat{\mu}_i - \mu) + \frac{1}{kn^2} \mu^T N (\hat{\mu}_i - \mu) + \frac{1}{kn^2} (\hat{\mu}_i - \mu)^T \mu N \\ &= \frac{1}{kn^2} (\hat{\mu}_i - \mu)^T N (\hat{\mu}_i - \mu) + \frac{2}{kn^2} \mu^T N (\hat{\mu}_i - \mu) = A + 2B \end{aligned}$$

Note that $\mu^T N (\hat{\mu}_i - \mu) = \mu^T N \mu$ because they are both scalars.

By definition of matrix A , A can be written as

$$\begin{aligned} A &= \frac{1}{k} \sum_{i=1}^k \left(\frac{1}{n} \sum_{i \in s_i} (\hat{\mu}(x_i) - \mu(x_i)) - \frac{1}{N} \sum_{j \in U} (\hat{\mu}(x_j) - \mu(x_j)) \right)^2 \\ &= \frac{1}{k} \sum_{i=1}^k \left[\left(\frac{1}{n^2} \sum_{i \in s_i} (\hat{\mu}(x_i) - \mu(x_i)) \right)^2 + \frac{1}{N^2} \left(\frac{1}{n^2} \sum_{j \in U} (\hat{\mu}(x_j) - \mu(x_j)) \right)^2 - \frac{2}{nN} \sum_{i \in s_i} (\hat{\mu}(x_i) - \mu(x_i)) \sum_{j \in U} (\hat{\mu}(x_j) - \mu(x_j)) \right] \\ &= \frac{1}{k} \sum_{i=1}^k (a + b + c) \end{aligned}$$

where

$$\begin{aligned} a &= \left(\frac{1}{n^2} \sum_{i \in s_i} (\hat{\mu}(x_i) - \mu(x_i)) \right)^2 \\ b &= \frac{1}{N^2} \left(\frac{1}{n^2} \sum_{j \in U} (\hat{\mu}(x_j) - \mu(x_j)) \right)^2 \\ c &= -\frac{2}{nN} \sum_{i \in s_i} (\hat{\mu}(x_i) - \mu(x_i)) \sum_{j \in U} (\hat{\mu}(x_j) - \mu(x_j)) \end{aligned}$$

note that

$$\begin{aligned} \hat{\mu}(x_i) - \mu(x_i) &= S_{ij} Y_j - \mu(x_i) = S_{ij} (\mu_i - \varepsilon_j) - \mu(x_i) \\ &= b_i(x_j) + S_{ij} \varepsilon_j \end{aligned}$$

Here S_{ij} is the smoother matrix and $S_{ij} = e_1^T (X_{ij}^T W_{ij} X_{ij})^{-1} X_{ij}^T W_{ij}$ where X_{ij} and W_{ij} are defined. In this case for simplicity we will use K_{ij} to denote $K\left(\frac{x_i - x_j}{h}\right)$. Now expanding the parentheses in a the following expression is obtained

$$\begin{aligned} E(a) &= \frac{1}{n^2} \sum_{j \in s_i} b_i^2(x_j) + \frac{1}{n^2} E\left(\sum_{j \in s_i} S_{ij} \varepsilon_i \varepsilon_i^T S_{ij}^T\right) + \\ &\quad \frac{1}{n^2} E\left(\sum_{j \in s_i} \sum_{l \in s_i, l \neq j} b_i(x_j) b_i(x_l)\right) + \\ &\quad \frac{1}{n^2} E\left(\sum_{j \in s_i} \sum_{l \in s_i, l \neq j} S_{ij} \varepsilon_i \varepsilon_i^T S_{lj}^T\right) \quad (14) \end{aligned}$$

The right hand side of equation 14 has four terms; each part will be calculated one by one.

(i) first let us investigate $\frac{1}{n^2} \sum_{j \in s_i} b_i^2(x_j)$. Using the technique similar to the one used by Wand and Jones(1995).

Let

$$\mu_i = (\mu(x_1), \dots, \mu(x_N))$$

Then by Taylor theorem

$$\mu_i = X_{ij} \begin{pmatrix} \mu(x_j) \\ D_{\mu(x_j)} \end{pmatrix} + R_{\mu(x_j)}$$

where

$$D_{\mu(x_j)} = \left(\mu'(x_j), \frac{1}{2} \mu''(x_j), \dots, \frac{1}{p} \mu^p(x_j) \right)$$

And $R_{\mu(x_j)}$ is a vector of Taylor series remainder terms, therefore,

$$\begin{aligned} b_i(x_j) &= S_{ij} \mu_i - \mu(x_j) = S_{ij} R_{\mu(x_j)} \\ &= e_1^T \begin{pmatrix} Z_{11j} & \cdots & Z_{1(p+1)j} \\ \vdots & \ddots & \vdots \\ Z_{(p+1)1j} & \cdots & Z_{(p+1)(p+1)j} \end{pmatrix} \begin{pmatrix} t_{1j} \\ \vdots \\ t_{(p+1)j} \end{pmatrix} \equiv e_1^T Z_j^{*-1} T \end{aligned}$$

Under assumption A2 and A3 by lemma 2(ii) Bredit and Opsomer (2000) for a certain point x_j there are atleast $p+1$ points in the interval $[x_j - h, x_j + h]$. So Z_j^* is invertible.

Lemma 1: Assume that the kernel function K_{ij} is bounded above, then

$$\frac{1}{nh} K_{ij}(x_i - x_j)^r = O(h^r)$$

Where $r = 0, 1, 2, \dots$

The proof of lemma 1 is provided by X.Li(2006). Thus, suppose A4 holds, by lemma 1, we have

$$z_{stj} = O(h^{s+t-2})$$

$$t_{kj} = O(h^{k+p})$$

and

$$b_i(x_j) = e_1^T \begin{pmatrix} O(1) & \cdots & O(h^p) \\ \vdots & \ddots & \vdots \\ O(h^p) & \cdots & O(h^{2p}) \end{pmatrix}^{-1} \begin{pmatrix} O(h^{p+1}) \\ \vdots \\ O(h^{p+1}) \end{pmatrix}$$

Note that the order of a matrix is the same as its inverse, therefore,

$$b_i(x_j) = O(h^{p+1}) + O(h^{3p+1}) = O(h^{p+1}) \quad (15)$$

thus

$$\frac{1}{n^2} \sum_{j \in S_i} b_i^2(x_j) = \frac{1}{n^2} \sum_{j \in S_i} O(h^{2p+1}) = O\left(\frac{h^{2p+2}}{n}\right) \quad (16)$$

(ii) Now we compute $\frac{1}{n^2} E(\sum_{j \in S_i} S_{ij} \varepsilon_i \varepsilon_i^T S_{ij}^T)$ in equation 14

$$\begin{aligned} \frac{1}{n^2} E \left(\sum_{j \in S_i} S_{ij} \varepsilon_i \varepsilon_i^T S_{ij}^T \right) &= \frac{1}{n^2} E \left(\sum_{j \in S_i} S_{ij} \sum_{l \in S_i} S_{il}^T \right) \\ e_1^T (X_{ij}^T W_{ij} X_{ij})^{-1} X_{ij}^T W_{ij} \sum_{l \in S_i} X_{il} (X_{il}^T W_{il} X_{il})^{-1} e_1 \\ &= e_1^T Z_j^{*-1} C_j^* Z_j^{*-1} e_1 \end{aligned}$$

Where \sum_i is the variance covariance matrix of the model 7 and

$$\begin{aligned} \sum_i &= \text{diag}(w_1, w_2, \dots, w_n) \\ C_j^* &= \begin{pmatrix} c_{11j} & \cdots & c_{1(p+1)j} \\ \vdots & \ddots & \vdots \\ c_{(p+1)1j} & \cdots & c_{(p+1)(p+1)j} \end{pmatrix} \\ c_{slj} &= \frac{1}{n^2 h^2} \sum_{i=1}^n K_{ij}^2 w_i (x_i - x_j)^{s+t-2} = O\left(\frac{h^{s+t-3}}{n}\right) \end{aligned}$$

By lemma 1 X.Li (2006) shows that

$$e_1^T Z_j^{*-1} C_j^* Z_j^{*-1} e_1 = O\left(\frac{1}{nh}\right)$$

and thus

$$\frac{1}{n^2} E(\sum_{j \in S_i} S_{ij} \varepsilon_i \varepsilon_i^T S_{ij}^T) = O\left(\frac{1}{n^2 h}\right) \quad (17)$$

(iii) Thirdly we now compute $\frac{1}{n^2} E(\sum_{j \in S_i, l \in S_i, l \neq j} b_i(x_j) b_i(x_l))$ in 14 and using the results in 15 we get

$$\frac{1}{n^2} E(\sum_{j \in S_i, l \in S_i, l \neq j} b_i(x_j) b_i(x_l)) = O(h^{2p+2}) \quad (18)$$

(iv) the last term on the right hand side of 14 is $\frac{1}{n^2} E(\sum_{j \in S_i} \sum_{l \in S_i, l \neq j} S_{ij} \varepsilon_i \varepsilon_i^T S_{lj}^T)$.

X.Li (2006) shows that

$$\frac{1}{n^2} E(\sum_{j \in S_i} \sum_{l \in S_i, l \neq j} S_{ij} \varepsilon_i \varepsilon_i^T S_{lj}^T) = O\left(\frac{1}{n}\right) \quad (19)$$

Assumption A3 implies that $nh \rightarrow \infty$ and by 16, 17 and 19

$$E(a) = O(h^{2p+2}) + O\left(\frac{1}{n}\right) \quad (20)$$

Similarly, $E(b)$ and $E(c)$ is calculated under A3 $nh \rightarrow \infty$, therefore

$$E(b) = O(h^{2p+2}) + O\left(\frac{1}{n}\right) \quad (21)$$

$$E(c) = O(h^{2p+2}) + O\left(\frac{1}{n}\right) \quad (22)$$

Also note that $A > 0$ and $|A| = A$ thus by 20, 21 and 22

$$\begin{aligned} E|A| &= E(A) = \frac{1}{k} \sum_{i=1}^k (E(a) + E(b) + E(c)) \\ &= O(h^{2p+2}) + O\left(\frac{1}{n}\right) \end{aligned}$$

this implies that

$$A = O(h^{2p+2}) + O\left(\frac{1}{n}\right)$$

Next using a similar approach to that of A

$$B = O(h^{p+1}) + O\left(\frac{1}{\sqrt{n}}\right)$$

Thus

$$\# = A + 2B = O(h^{2p+2}) + O\left(\frac{1}{\sqrt{n}}\right) \quad (23)$$

Now let us calculate $\# \#$ in 13

$$\begin{aligned} \delta_{NP}^2 - \delta_{NP}^2 &= \frac{1}{n} \sum_{j=1}^n \frac{(Y_j - \mu(x_j))^2}{w_j} - \delta_{NP}^2 \\ &+ \frac{1}{n} \sum_{j=1}^n \frac{(\hat{\mu}(x_j) - \mu(x_j))^2}{w_j} \\ &+ \frac{2}{n} \sum_{j=1}^n \frac{(Y_j - \hat{\mu}(x_j))(\hat{\mu}(x_j) - \mu(x_j))}{w_j} \end{aligned}$$

X. Li (2006) shows that

$$\frac{2}{n} \sum_{j=1}^n \frac{(Y_j - \mu(x_j))^2}{w_j} - \delta_{NP}^2 = O_p\left(\frac{1}{\sqrt{n}}\right) \quad (24)$$

$$\frac{1}{n} \sum_{j=1}^n \frac{(\hat{\mu}(x_j) - \mu(x_j))^2}{w_j} = O_p(h^{2p+2}) + O_p\left(\frac{1}{nh}\right) \quad (25)$$

and

$$\frac{2}{n} \sum_{j=1}^n \frac{(Y_j - \hat{\mu}(x_j))(\hat{\mu}(x_j) - \mu(x_j))}{w_j} = O_p(h^{p+1}) + O_p\left(\frac{1}{\sqrt{nh}}\right) \quad (26)$$

Since

$$\frac{1}{kn^2} \text{tr}(N\Sigma) = O(1)$$

And by 24, 25 and 26, we have

$$\frac{1}{kn^2} \text{tr}(N\Sigma) \hat{\delta}_{NP}^2 - \delta_{NP}^2 = O_p(h^{p+1}) + O_p\left(\frac{1}{\sqrt{nh}}\right) \quad (27)$$

Therefore by 23 and 26

$$\hat{V}_{NP}(\bar{Y}_s) - E\left(\text{Var}_p(\bar{Y}_s)\right) = O_p(h_m^{p+1}) + O_p\left(\frac{1}{\sqrt{nh}}\right)$$

hence the result.

2.8. Simulation Study

To further investigate the statistical properties of the above variance estimators, a simulation study was performed. For simplicity, the researcher considered the case where there was only one auxiliary variable x . It is also assumed that the errors are independently and normally distributed with homogeneous variances. Two super population models are examined. One is the linear model

$$y_j = 2x_j + \varepsilon_j \quad (28)$$

Where $j = 1, 2, \dots, N$

And $\varepsilon_j \sim N(0, \delta_1^2)$

The quadratic model

$$y_j = 1 + 2(x_j - 0.5)^2 + \varepsilon_j \quad (29)$$

$$\varepsilon_j \sim N(0, \delta_2^2)$$

The bigger the R^2 , the bigger the predictive power of the model. The two levels of R^2 , that are achieved are

$R^2 = 0.75$ "precise" model and $R^2 = 0.25$ the diffuse model.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{j \in U} \varepsilon_j^2}{\sum_{j \in U} (Y_j - \bar{Y}_U)^2}$$

To draw a systematic sample, the population first needs to be sorted. Three ways are considered: (1) Sort by auxiliary variable x ; (2) Sort by z_1 , where $z_{1j} = x_j + \theta_{1j}$ and $\theta_{1j} \sim N(0, \delta_{\theta_{1j}}^2)$. Choose $\delta_{\theta_{1j}}^2$ to make $R_{z_1}^2 = 0.75$ (3) Sort by z_2 , where $z_{2j} = x_j + \theta_{2j}$ and $\theta_{2j} \sim N(0, \delta_{\theta_{2j}}^2)$. Choose $\delta_{\theta_{2j}}^2$ to make $R_{z_2}^2 = 0.25$. Populations of size $N = 2000$ is generated. To achieve this, 2000 values of model variable x from the uniform distribution on $[0, 1]$ and 2000 values of error " from $N(0, 1)$ were generated. Then 2000 values of response variable y computed by model 28 and 29. Two systematic samples of size $n = 500$ and $n = 100$, with corresponding sampling intervals $k = 4$ and $k = 20$ are considered respectively. To draw a systematic sample, the data first sorted, either by x or z , from the smallest to the largest, then randomly choose an observation from the first k observations, say the i^{th} one. Then, the selected sample consists of the observations with the following subscripts: $i, i + k, \dots, i + (n - 1)k$.

For each simulation, the corresponding $\text{Var}_p(\bar{Y}_s)$, $E(\text{Var}_p(\bar{Y}_s))$, \hat{V}_{OL} , \hat{V}_{NO} and \hat{V}_{SRS} is calculated. For $\hat{V}_{NP}(\bar{Y}_s)$ it is calculated using two bandwidth values: $h = 0.50$ and $h = 0.25$, each simulation setting is repeated $B = 10\,000$ times. The researcher then compare the performance of the nonparametric variance estimator $\hat{V}_{NP}(\bar{Y}_s)$ with the overlapping differences \hat{V}_{OL} , the non-overlapping differences estimator, \hat{V}_{NO} , which are recommended by Wolter [2007] and the simple random sampling estimator \hat{V}_{SRS} . The relative bias (RB) and the mean squared error (MSE) are calculated. Let \hat{V} represent $\text{Var}_p(\bar{Y}_s)$, $\hat{V}_{NP}(\bar{Y}_s)$, \hat{V}_{OL} , \hat{V}_{NO} and \hat{V}_{SRS}

$$RB = \frac{E_{MP}(\hat{V}) - E_M(\text{Var}_p(\bar{Y}_s))}{E_M(\text{Var}_p(\bar{Y}_s))}$$

$$MSE = E_{MP}(\hat{V} - E_M(\text{Var}_p(\bar{Y}_s)))^2$$

where E_M denotes the expectation under the superpopulation model M , and E_{MP} denotes the expectation under both the model and design.

3. Simulation Results and Discussion

3.1. Introduction

This section presents the results obtained through the simulation discussed in section 2.7.

3.2. Results and Discussion

Table 2 gives the relative biases of $\hat{V}_{NP}(\bar{Y}_s)$, \hat{V}_{OL} , \hat{V}_{NO} and \hat{V}_{SRS} for the sample of size $n=500$ for different sorting variables with homoscedastic errors. The relative biases for non parametric estimators are computed at different bandwidth. The results from table show that given a proper bandwidth is chosen, non parametric estimator performs well overall than other three estimators resulting to smaller biases with most biases being less than zero. Especially when the super-population model is linear, $\hat{V}_{NP}(\bar{Y}_s)$ tends to favor bigger bandwidth. This is because local linear regression was used in the calculation of $\hat{V}_{NP}(\bar{Y}_s)$. The bigger bandwidth results in more points in the neighborhood of x_j and because the local polynomial regression is local linear which is correct one for this population with linear trend, so having more points will increase the precision of each local linear regression.

When the super-population model is quadratic, it tends to favor small bandwidth. This is because, as discussed above, for parametric estimation, linear regression will not estimate quadratic trend well. In other words, the wider the neighborhood, the more likely a quadratic trend will be seen there. Therefore local linear regression on that neighborhood could be bad. When the bandwidth is small, then the trend within each local interval will be approximated well by a linear trend.

The estimator based on simple random sampling \hat{V}_{SRS} performed poorly, resulting in large biases in both cases as it

overestimated the true variance.

It can also be seen that \hat{V}_{OL} , and \hat{V}_{NO} have smaller biases under linear and quadratic models when the population is sorted by x before drawing a systematic sample. This is because \hat{V}_{OL} and \hat{V}_{NO} capture the population trend very well and thus very efficient. When the sorting variable is not related to the population that is sorting the population by z_1 and z_2 , overlapping and non overlapping difference estimators cannot capture the population trend well hence resulting to large biases and MSE.

Table 3 gives the ratios of MSE for $\hat{V}_{NP}(\bar{Y}_s)$ that is evaluated at two bandwidth values ($h=0.25$ and $h=0.5$) obtained by dividing the MSE of other estimators by the MSE of $\hat{V}_{NP}(\bar{Y}_s)$ evaluated at a bandwidth ($h=0.1$). The MSE measures the variability of an estimator and smaller MSE values are normally desired. Therefore, it can be seen from this study that, $\hat{V}_{NP}(\bar{Y}_s)$ performs better than the variance estimators \hat{V}_{OL} , \hat{V}_{NO} and \hat{V}_{SRS} as it has smaller MSE values in almost all the cases of linear and quadratic models.

Table 2. Relative Bias(%) for $\hat{V}_{NP}(\bar{Y}_s)$ with bandwidth($h=0.1, 0.25, 0.5$), \hat{V}_{OL} , \hat{V}_{NO} and \hat{V}_{SRS} with $n=500$.

Mean function	LINEAR			QUADRATIC		
Sorting variable	1	0.75	0.25	1	0.75	0.25
$h=0.1$	-0.937	-0.918	-0.783	-0.994	-0.974	-0.920
$h=0.25$	-0.956	-0.526	-0.693	-0.940	-0.940	-0.905
$h=0.5$	-0.946	-0.841	-0.870	-0.952	-0.929	-0.856
\hat{V}_{OL}	-0.298	0.160	0.473	0.111	0.343	-2.025
\hat{V}_{NO}	-0.304	0.162	0.465	0.120	0.341	-1.020
\hat{V}_{SRS}	0.300	0.169	4.58	12.036	0.476	-1.026

Table 3. MSE(%) for $\hat{V}_{NP}(\bar{Y}_s)$ with bandwidth(0.25, 0.5), \hat{V}_{OL} , \hat{V}_{NO} and \hat{V}_{SRS} with $n=500$ divided by MSE of $\hat{V}_{NP}(\bar{Y}_s)$ with bandwidth $h=0.1$ and Homoscedastic errors.

Mean function	LINEAR			QUADRATIC		
Sorting variable	1	0.75	0.25	1	0.75	0.25
$h=0.25$	0.3	0.58	1.00	1.00	0.99	10.00
$h=0.5$	0.56	0.23	1.00	1.10	1.01	10.99
\hat{V}_{OL}	4.34	3.98	17.15	7.44	15.45	174.17
\hat{V}_{NO}	4.39	4.12	17.25	7.47	16.47	174.10
\hat{V}_{SRS}	43.86	38.69	27.02	17.32	17.33	172.40

4. Conclusions and Recommendation

The aim of this study was to develop design unbiased estimator of variance of the systematic means using local polynomial regression as the estimation technique. This study reveals that, the estimator based on non parametric model (7) using local polynomial regression as the estimation technique $\hat{V}_{NP}(\bar{Y}_s)$ is a consistent estimator for the $Var_p(\bar{Y}_s)$. In comparison to other estimator discussed in Wolter (2007), the local polynomial estimator $\hat{V}_{NP}(\bar{Y}_s)$ performed better in all the three cases. Therefore, this estimator has proved to be consistent and unbiased in estimating the design variance of systematic sample mean.

Hence, in practice, this study recommends the use of non parametric estimator $\hat{V}_{NP}(\bar{Y}_s)$ for estimating the variance of systematic sample mean over the estimators proposed by Wolter (2007).

Nomenclature

MSE- Mean squared Error
 RB - Relative Bias
 NO -Non Overlapping difference
 OL-Ordinary Least Square
 SRS-Simple random sampling
 SRSWOR-Simple Random Sampling without Replacement
 NP-Non parametric model

Acknowledgement

I thank the almighty God for granting me the knowledge to carry out this study. My sincere gratitude goes to my colleagues for the company and readiness to read this material and their corrections. I can't overlook the love and support from my family and friends.

Special thanks to my supervisor Dr. G. Orwa for his presence since the start of this work until the end. I highly appreciate the moral and academic support he gave me. I do also thank my supervisor prof. R. Odhiambo for the scholarly and professional assistance and his presence to correct my work. I also express my gratitude to the staff of statistics department for their friendly guidance throughout my study. May God bless you all.

References

- [1] Ayora. O. Model based nonparametric variance estimation in systematic sampling.
- [2] Journal of Contemporary Research in Business Vol 5 No 12, 2014.
- [3] F. Breidt and J. Opsomer. Local polynomial regression estimators in survey sampling. Annals of Statistics 28, 1026-1053, 2000.
- [4] Cleveland W. S and Devlin S. Local weighted regression: An approach to regression analysis by local fitting. Journal of American Statistical Association, vol 85, 596-610, 1988.
- [5] W. G. Cochran. Sampling Techniques. Wiley eastern, 1977
- [6] J. Fan. Design-adaptive nonparametric regression. Journal of American Statistical Association 87, 998-1004, 1992.
- [7] J. Fan. Local linear regression smoother and their minimax efficiencies. Annals of statistics 21, 196-216, 1993.
- [8] J. Fan and I. Gijbels. Data-driven bandwidth and local linear regression smoothers. Journal of the royal statistical society, series B 57, 371-394, 1995.
- [9] Fan and I. Gijbels. Local polynomial model and its application. CRS press, 1996.
- [10] Madow and L. Madow. On the theory of systematic sampling. Annals of Mathematical Statistics, 25, pp.1-24, 1944.
- [11] Montanari. G and Bartolucci. F. Estimating the variance of systematic sample mean. Journal of Italian statistical society, 7: 185-196, 1998.

- [12] Montanari. G and Bartolucci. F. A new class of variance estimators of variance of the systematic sample means. Journal of Statistical planning and inference, 136 pg 1512-1525, 2006.
- [13] Opsomer J. D and Ruppert. Fitting a bivariate additive model by local polynomial regression. Annals of statistics 25, 186-211, 1997.
- [14] Rana and R. Singh. mean on systematic sampling with supplementary observation. The Indian Journal of Statistics pg 205-211, 1989.
- [15] Stone. Optimal rate of convergence for nonparametric estimators. Annals of statistics 8, 1348-1360, 1980,
- [16] M. Wand and M. Jones. Kernel smoothing. Chapman and Hall, London, 1995.
- [17] K. Wolter. Introduction to variance estimation. Springer, 2007.
- [18] C. J. Wu. Estimation in systematic sampling with supplementary observation. The Indian Journal of Statistics Vol 46,3 pg 306-315, 1984.
- [19] X. Li. Application of nonparametric regression in survey statistics. Restrospective theses and Desertations, 2006.
- [20] X. Li and J. Opsomer. Model-based variance estimation for systematic sampling. Department of Statistics, Iowa State University, Ames, IA 50011, 2010
- [21] Zinger. Variance estimation in partially systematic sampling. Journal of American
- [22] Statiscal Association, vol 75,206-211, 1980.