BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number:CENSUS/SRD/RR-84-02

AN INVESTIGATION OF SOME ESTIMATORS OF VARIANCE FOR SYSTEMATIC SAMPLING

by Kirk M. Wolter

Chief, Statistical Research Division
U.S. Bureau of the Census

This paper was submitted for publication in Journal of the American Statistical Association in December 1983.

December 12, 1983

AN INVESTIGATION OF SOME ESTIMATORS OF VARIANCE FOR SYSTEMATIC SAMPLING

by

Kirk M. Wolter*

ABSTRACT

This paper provides the survey practitioner some guidance about the problem of variance estimation for equal probability single-start systematic samples. An investigation of eight alternative estimators of the variance of the sample mean is presented. In the first half of the investigation, the theoretical properties of the estimators are compared using several superpopulation models, while in the second half, the comparison is empirical, based on several real populations. Recommendations are made about the appropriateness of the various estimators.

KEY WORDS: Systematic sampling, finite population, variance estimation, superpopulation, autocorrelation.

*Kirk M. Wolter is Chief, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C. 20233 and Professorial Lecturer, The George Washington University, Washington, D.C. 20052. The author wishes to acknowledge useful commments received from Cary Isaki, Rod Little, C.F. Wu, Nash Monsour, and two referees.

AN INVESTIGATION OF SOME ESTIMATORS OF VARIANCE FOR SYSTEMATIC SAMPLING

by

Kirk M. Wolter*

1. Introduction

The method of systematic sampling, first studied by the Madows (1944), is used widely in surveys of finite populations. When properly applied, the method picks up any obvious or hidden stratification in the population, and thus can be more precise than random sampling. In addition, systematic sampling is implemented easily, thus reducing costs. Since a systematic sample can be regarded as a random selection of one cluster, however, it is not possible to give an unbiased, or even consistent, estimator of the design variance. Biased estimators of variance must be sought if we are to estimate the precision of our survey estimators from the sample itself.

The objective of this paper is to provide the survey practitioner some guidance about the specific problem of estimating the design variance of the systematic sampling mean, $\bar{y}$. We shall only consider equal probability systematic sampling with a single, random start. It is, of course, possible to produce an unbiased estimator of variance when we draw two or more random starts, though Gautschi (1957) shows that this practice may lead to inefficient estimates of the population mean.

Few guiding principles about variance estimation are available in the literature on systematic sampling, particularly for household and establishment surveys. In the 1940's several authors addressed this issue, including Osborne (1942), Cochran (1946), Matérn (1947), and Yates (1949). Recent references are Koop (1971), Heilbron (1978), Zinger (1980), and Wu (1981). One of the most comprehensive discussions is given by Cochran (1977). The topic may have received little attention because systematic sampling is often used at the last stage of sampling, where rigorous estimators of the total variance can be given.

Section 2 contains a description of eight alternative estimators of the variance of $\bar{y}$. Some theoretical results regarding the eight estimators are worked out in Section 3 using a superpopulation model. In Section 4 some empirical comparisons of the estimators are made. Section 5 closes the paper with a general summary and recommendations.

To aid in the reading of this article, it is useful to remember the following procedure for considering variance estimation issues.

(a) Gather as much prior information as possible about the nature and ordering of the population.

(b) If an auxiliary variable, closely related to the estimation variable, is available for all units in the population, then try several variance estimators on this variable. This investigation may provide information about which estimator will have the best

properties for estimating the variance of the estimation variable.

(c) Use the prior information in (a) to construct a simple model for the population. The results in Sections 3 and 4 may be used to select an appropriate estimator for the chosen model.

(d) Keep in mind that most surveys are multipurpose and it may be important to use different variance estimators for different characteristics.

Steps (a) - (d) essentially suggest that one know the population well before choosing a variance estimator, which is exactly the advice most authors since the Madows have suggested before using systematic sampling.

2. Description of the Estimators

To concentrate on essentials, we shall assume $N = nk$ where N is the population size, n is the desired sample size, and k is the sampling interval. We let Y_{ij} denote the value of the characteristic of interest for the j-th unit in the i-th systematic sample, where $i = 1, \ldots, k$ and $j = 1, \ldots, n$. We adopt the convention of using upper case Y's to denote the values of units in the population, and lower case y's to denote the values of the units in the sample. The systematic sampling mean $\bar{y}$ and its variance are

$$\bar{y} = \sum_{j}^{n} y_{ij}/n$$

and

$$\mathrm{Var}\{\bar{y}\} = (\sigma^2/n)[1+(n-1)\rho],$$

respectively, where

$$\sigma^2 = \sum_{i}^{k}\sum_{j}^{n} (Y_{ij} - \bar{Y}..)^2/nk$$

denotes the population variance,

$$\rho = \sum_{i}^{k}\sum_{j}^{n}\sum_{j'\neq j}^{n} (Y_{ij} - \bar{Y}..)(Y_{ij'} - \bar{Y}..)/kn(n-1)\sigma^2$$

denotes the intraclass correlation between pairs of units in the same sample, and the customary "dot" notation indicates a summation.

Eight alternative estimators of the variance $\mathrm{Var}\{\bar{y}\}$ will be compared in succeeding sections. They are defined in Table 1 for the i-th selected sample.

(Table 1 goes here)

There is an abundance of reasonable estimators which may be used to estimate $\mathrm{Var}\{\bar{y}\}$ and these eight estimators represent a cross-section of the various general classes of estimators. A class of estimators not considered here arises when one supplements the systematic sample with either a simple random sample or another systematic sample of smaller size. See Zinger (1980) or Wu (1981).

Table 1. Eight Estimators of Variance for Systematic Sampling

Form of Estimator	Comments
$v_1(i) = (1-f)s^2/n$	This estimator corresponds to simple random sampling without replacement. For systematic sampling, it tends to over or underestimate the variance as $\rho < -1/(N-1)$ or $\rho > -1/(N-1)$.
$v_2(i) = (1-f)(1/n)\sum_{j=2}^{n} a_{ij}^2/2(n-1)$	Estimators 2 and 3 are based on overlapping and nonoverlapping differences. v_3 corresponds to stratified sampling with 2 units in each of n/2 strata. v_2 is based on more "degrees of freedom" than v_3.
$v_3(i) = (1-f)(1/n)\sum_{j=1}^{n/2} a_{i,2j}^2/n$	
$v_4(i) = (1-f)(1/n)\sum_{j=3}^{n} b_{ij}^2/6(n-2)$	Estimators 4, 5, and 6 are based on higher order differences. v_6 was first suggested by Yates (1949). v_4 is based upon second differences, which annihilate a linear trend in the population values. The divisor is the sum of squares of the coeffients times the number of differences in the sum.
$v_5(i) = (1-f)(1/n)\sum_{j=5}^{n} c_{ij}^2/3.5(n-4)$	

Table 1. Eight Estimators of Variance for Systematic Sampling continued

Form of Estimator	Comments
$v_6(i) = (1-f)(1/n) \sum_{j=9}^{n} d_{ij}^2/7.5(n-8)$	
$v_7(i) = (1-f) \frac{1}{p(p-1)} \sum_{\alpha}^{p} (\bar{y}_\alpha - \bar{y})^2$	Based on splitting the sample into p systematic subsamples. When f is negligible, Bias $\{v_7\} = (\mathrm{Var}\{\bar{y}_\alpha\} - p\mathrm{Var}\{\bar{y}\})/(p-1)$, implying that v_7 is unbiased when the variance is inversely proportional to sample size. See Koop (1971) regarding the case p = 2.
$v_8(i) = (1-f)(s^2/n)[1+2/\ell n(\hat{\rho}_k)+2/(\hat{\rho}_k^{-1}-1)]$, if $\hat{\rho}_k > 0$ $= (1-f)s^2/n$ if $\hat{\rho}_k < 0$	Devised from a superpopulation model where the correlation between two units in the population depends only on the distance between them. See e.g., Cochran (1946), Osborne (1942), Matérn (1947). In v_8 $\hat{\rho}_k$ is an estimator of the correlation between units k units apart. Heilbron (1978) gives three estimtors that are similar to v_8.

Table 1. Eight Estimators of Variance for Systematic Sampling continued

NOTE:

$$s^2 = \sum_{j=1}^{n} (y_{ij}-\bar{y})^2/(n-1)$$

$$\hat{\rho}_k = \sum_{j=2}^{n} (y_{ij}-\bar{y})(y_{i,j-1}-\bar{y})/(n-1)s^2$$

$$a_{ij} = y_{ij} - y_{i,j-1}$$

$$b_{ij} = y_{ij} - 2y_{i,j-1} + y_{i,j-2}$$

$$c_{ij} = y_{ij}/2 - y_{i,j-1} + y_{i,j-2} - y_{i,j-3} + y_{i,j-4}/2$$

$$d_{ij} = y_{ij}/2 - y_{i,j-1} + - \cdots + y_{i,j-8}/2$$

$\bar{y}_\alpha$ = sample mean of the α-th systematic subsample of size n/p (where p and n/p are integers)

$f = n/N = 1/k.$

3. Some Comparisons Based on Simple Models

In this section, we shall introduce the notion of model bias and use it as a criterion for comparing the various estimators of variance. We assume the finite population is generated according to the superpopulation model

$$Y_{ij} = \mu_{ij} + e_{ij}, \tag{3.1}$$

where the μ_{ij} denote fixed constants and the errors e_{ij} are $(0, \sigma^2)$ random variables. Our main goal is the determination of conditions on μ_{ij} and e_{ij} under which the eight estimators of variance perform well with respect to model bias.

The expected bias and expected relative bias of an estimator v_α (α=1, ..., 8) are defined by

$$\mathcal{B}\{v_\alpha\} = \mathcal{E}E\{v_\alpha\} - \mathcal{E}Var\{\bar{y}\}$$

and

$$\mathcal{R}\{v_\alpha\} = \mathcal{B}\{v_\alpha\}/\mathcal{E}Var\{\bar{y}\},$$

respectively. In Sections 3.1 - 3.4 we present expressions for $\mathcal{R}\{v_\alpha\}$ for five useful models of the form (3.1). These results extend simply to the model with heterogeneous error variances.

In addition to this analytical work, we describe in Section 3.5 the results of a small Monte Carlo study that was made to investigate properties of the estimators. Seven models were chosen for the study and they are described in Table 5. For each model, 200 finite populations of size N = 1000 were generated, and in each population, the bias and MSE of each estimator were computed, as well as the proportion of confidence intervals that contained the true population mean. These quantities were

averaged over the 200 populations, giving the expected bias, the expected MSE, and the expected confidence level for each of the eight estimators. The multiplier used in forming the confidence intervals was the 0.025 point of the standard normal distribution. Estimator v_7 was studied with $p = 2$.

3.1 Linear Trend

Populations with linear trend may be represented by

$$\mu_{ij} = \beta_0 + \beta_1[i+(j-1)k], \tag{3.2}$$

where β_0 and β_1 denote fixed (but unknown) constants and the errors e_{ij} are independent and identically distributed (iid) $(0, \sigma^2)$ random variables. It is easily seen that the expected variance for this model is

$$\mathcal{E}\mathrm{Var}\{\bar{y}\} = \beta_1^2(k^2-1)/12 + (1-f)\sigma^2/n. \tag{3.3}$$

The expectations of the eight estimators of variance are given in column 2 of Table 2. The expression for $\mathcal{E}E\{v_8\}$ was derived by approximating the expectation of the function $v_8(s^2, \hat{\rho}_k s^2)$ by the same function of the expectations $\mathcal{E}E\{s^2\}$ and $\mathcal{E}E\{\hat{\rho}_k s^2\}$, where we have used an expanded notation for v_8. In deriving this result it was also assumed that $\hat{\rho}_k > 0$ with probability one. This assumption guarantees that terms involving the operator $\ell n(\cdot)$ are well defined.

From Table 2 and (3.3), we conclude that the value of the intercept β_0 has no effect on the relative biases of the variance estimators, while the error variance σ^2 has only a slight effect. Similarly, the value of the slope β_1 has little

effect on the relative bias, unless β_1 is extraordinarily small. For populations where k is large and β_1 is not extremely close to 0, we have the following useful approximations:

$$\begin{aligned}
\mathcal{R}\{v_1\} &\doteq n \\
\mathcal{R}\{v_2\} &\doteq -(n-6)/n \\
\mathcal{R}\{v_3\} &\doteq -(n-6)/n \\
\mathcal{R}\{v_4\} &\doteq -1 \\
\mathcal{R}\{v_5\} &\doteq -1 \\
\mathcal{R}\{v_6\} &\doteq -1 \\
\mathcal{R}\{v_7\} &\doteq p.
\end{aligned}$$

Thus, from the point of view of relative bias, the estimators v_2 and v_3 are preferred.

The reader will recall that these results differ from Cochran (1977), who suggests v_4 for populations with linear trend. The contrasts defining v_4, v_5, and v_6 eliminate the linear trend, whereas v_2, v_3, and v_8 do not. Eliminating the linear trend is not a desirable property here because the variance is a function of the trend.

3.2 Stratification Effects

We now view the systematic sample as a selection of one unit from each of n strata. This situation may be represented by the model

$$\mu_{ij} = \mu_j, \tag{3.4}$$

for all i and j, where the errors e_{ij} are iid $(0,\sigma^2)$ random variables. That is, the unit means μ_{ij} are constant within a stratum of k units.

Table 2. Expected Values of Eight Estimators of Variance

Estimator	Population Type		
	Linear Trend	Stratification Effects	Autocorrelated
v_1	$(1-f)[\beta_1^2 k^2(n+1)/12+\sigma^2/n]$	$(1-f)[\sum_j^n(\mu_j-\bar{\mu})^2/n(n-1)+\sigma^2/n]^b$	$(1-f)(\sigma^2/n)\{1-\frac{2}{n-1}\frac{(\rho^k-\rho^N)}{(1-\rho^k)}+\frac{2}{n(n-1)}[\frac{(\rho^k-\rho^N)}{(1-\rho^k)^2}-(n-1)\frac{\rho^N}{(1-\rho^k)}]\}$
v_2	$(1-f)[\beta_1^2 k^2/2n+\sigma^2/n]$	$(1-f)[\sum_j^{n-1}(\mu_j-\mu_{j+1})^2/2n(n-1)+\sigma^2/n]$	$(1-f)(\sigma^2/n)(1-\rho^k)$
v_3	$(1-f)[\beta_1^2 k^2/2n+\sigma^2/n]$	$(1-f)[\sum_j^{n/2}(\mu_{2j-1}-\mu_{2j})^2/n^2+\sigma^2/n]$	$(1-f)(\sigma^2/n)(1-\rho^k)$
v_4	$(1-f)\sigma^2/n$	$(1-f)[\sum_j^{n-2}(\mu_j-2\mu_{j+1}+\mu_{j+2})^2/6n(n-2)+\sigma^2/n]$	$(1-f)(\sigma^2/n)[1-4\rho^k/3+\rho^{2k}/3]$
v_5	$(1-f)\sigma^2/n$	$(1-f)[\sum_j^{n-4}(\mu_j/2-\mu_{j+1}+\mu_{j+2}-\mu_{j+3}+\mu_{j+4}/2)^2/3.5n(n-4)+\sigma^2/n\}]$	$(1-f)(\sigma^2/n)[1-12\rho^k/7+8\rho^{2k}/7-4\rho^{3k}/7+\rho^{4k}/7]$
v_6	$(1-f)\sigma^2/n$	$(1-f)[\sum_j^{n-8}(\mu_j/2-\mu_{j+1}+-\ldots+\mu_{j+8}/2)^2/7.5n(n-8)+\sigma^2/n]$	$(1-f)(\sigma^2/n)[1-28\rho^k/15+24\rho^{2k}/15-20\rho^{3k}/15+16\rho^{4k}/15-12\rho^{5k}/15+8\rho^{6k}/15-4\rho^{7k}/15+\rho^{8k}/15]$
v_7	$(1-f)[\beta_1^2 k^2(p+1)/12+\sigma^2/n]$	$(1-f)[\rho^{-1}(p-1)^{-1}\sum_\alpha^p(\bar{\mu}_\alpha-\bar{\mu})^2+\sigma^2/n]^c$	$(1-f)(\sigma^2/n)\{1+[2/(p-1)][p(\rho^{pk}-\rho^N)/(1-\rho^{pk})-(\rho^k-\rho^N)/(1-\rho^k)]-[2/(p-1)][\{p^2/n\}\{(\rho^{pk}-\rho^N)/(1-\rho^{pk})^2-(n/p-1)\rho^N/(1-\rho^{pk})\}-n^{-1}\{(\rho^k-\rho^N)/(1-\rho^k)^2-(n-1)\rho^N/(1-\rho^k)\}]]$

Table 2. Expected Values of Eight Estimators of Variance continued

Estimator	Population Type		
	Linear Trend	Stratification Effects	Autocorrelation
v_8	$(1-f)[\gamma(0)/n][1+\frac{2}{\ell n[\gamma(1)/\gamma(0)]}+\frac{2}{\gamma(0)/\gamma(1)-1}]$ [a]	$(1-f)n^{-1}(\kappa(0)+\sigma^2)\,[1+\frac{2}{\ell n\frac{\kappa(1)}{\kappa(0)+\sigma^2}}+\frac{2}{\frac{\kappa(0)+\sigma^2}{\kappa(1)}-1}]$ [e]	$(1-f)(\sigma^2/n)[1+2/\ell n(\rho^k)+2\rho^k/(1-\rho^k)]+0(n^{-2})$ [d]

[a] $\gamma(1) = E\{\hat{\rho}_k s^2\} = \beta_1^2 k^2 (n-3)(n+1)/12 - \sigma^2/n$

$\gamma(0) = E\{s^2\} = \beta_1^2 k^2 n(n+1)/12 + \sigma^2$

[b] $\bar{\mu} = \sum_j^n \mu_j/n$

[c] $\bar{\mu}_\alpha$ = mean of a systematic subsample of size n/p of the μ_j.

[d] The approximation follows from elementary properties of the estimated autocorrelation function for stationary time series and requires bounded sixth moments.

[e] $\kappa(0) = (n-1)^{-1} \sum_j^n (\mu_j-\bar{\mu})^2$

$\kappa(1) = (n-1)^{-1} \sum_j^{n-1} (\mu_j-\bar{\mu})(\mu_{j+1}-\bar{\mu})$.

For this class of populations, we note that

$$\gamma_{i.} - \gamma_{..} = \bar{e}_{i.} - \bar{e}_{..},$$

and it follows that the expected variance of $\bar{y}$ is

$$\text{Var}\{\bar{y}\} = (1-f)\sigma^2/n. \qquad (3.5)$$

The expectaions of the eight estimators of variance are given in column 3 of Table 2. Once again, the expression for the expectation of v_8 is an approximation, and will be valid when n is large and $\Pr\{\hat{\rho}_k > 0\} = 1$.

From Table 2 and (3.5) we see that each of the first seven variance estimators has small and roughly equal relative bias when the stratum means μ_j are approximately equal. When the stratum mean are not equal, there can be striking differences between the estimators and v_1 and v_8 often have the largest absolute relative biases. This point is demonstrated in Table 3 which gives the expected biases for $\mu_j = j$, $\ell n(j) + \sin(j)$ with $n = 20$ and $p = 2$.

Based on these simple examples, we conclude that v_4, v_5, and v_6 provide the most protection against stratification effects. The contrasts used in these estimators tend to eliminate a linear trend in the stratum means, which is desirable here because the expected variance is not a function of such a trend. Conversely, v_2, v_3, and v_7 do not eliminate the trend. Estimators v_5 and v_6

Table 3. Expected Relative Bias Times σ^2 for Eight Estimators of Variance for Populations with Stratification Effects

Estimator	μ_j	
	j	$\ell n(j)+\sin(j)$
v_1	35.00	0.965
v_2	0.50	0.235
v_3	0.50	0.243
v_4	0.00	0.073
v_5	0.00	0.034
v_6	0.00	0.013
v_7	5.00	0.206
v_8	-0.67	-0.373

NOTE: $n = 20$, $p = 2$, $\sigma^2 = 100$.

will be preferred when there is a nonlinear trend in the stratum means. When the means μ_j are equal in adjacent nonoverlapping pairs of strata, estimator v_3 will have smallest expected bias. Estimator v_7 will have smallest expected bias when the μ_j are equal in adjacent nonoverlapping groups of p strata. For p = 2, v_3 and v_7 are comparable in terms of bias.

We note that the random model is a special case of the stratifiction effects model with $\mu_{ij} = \mu$ for all i and j. For this special case, the expected bias of the first seven estimators of variance is zero.

3.3 Correlated Populations

Another important class of populations occurs where the unit values are correlated. We may study such populations by assuming the y-variable has the time series specification

$$Y_t - \mu = \sum_{j=-\infty}^{\infty} \alpha_j \varepsilon_{t-j} \qquad (3.6)$$

for t = 1, ..., kn, where the sequence $\{\alpha_j\}$ is absolutely summable, and the ε_t are uncorrelated $(0,\sigma^2)$ random variables. The expected variance for this model is

$$\mathcal{E}\,\mathrm{Var}\{\bar{y}\} = (1-f)(1/n)\{\gamma(0) - \frac{2}{kn(k-1)} \sum_{h=1}^{kn-1} (kn-h)\gamma(h)$$

$$+ \frac{2k}{n(k-1)} \sum_{h=1}^{n-1} (n-h)\gamma(kh)\}, \qquad (3.7)$$

where

$$\gamma(h) = \mathcal{E}\{(Y_t-\mu)(Y_{t-h}-\mu)\} = \sum_{-\infty}^{\infty} \alpha_j \alpha_{j-h} \sigma^2.$$

By assuming that (3.6) arises from a low order autoregressive, moving average process, we may construct estimators of $\mathrm{Var}\{\bar{y}_{sy}\}$ and study their properties.

For example, we consider the model that motivates v_8. A representation for this model is the first order autoregressive process

$$Y_t-\mu = \rho(Y_{t-1}-\mu) + \varepsilon_t, \qquad (3.8)$$

where ρ is the first order autocorrelation coefficient (to be distinguished from the intraclass correlation coefficient) and $0 < \rho < 1$. By (3.7) the expected variance for this model is

$$\mathcal{E}\mathrm{Var}\{\bar{y}\} = (1-f)(\sigma^2/n)\{1 - \frac{2}{(k-1)} \frac{(\rho-\rho^{kn})}{(1-\rho)} + \frac{2}{kn(k-1)}[\frac{(\rho-\rho^{kn})}{(1-\rho)^2} - (kn-1)\frac{\rho^{kn}}{(1-\rho)}$$

$$+ \frac{2k}{(k-1)} \frac{(\rho^k-\rho^{kn})}{(1-\rho^k)} - \frac{2k}{n(k-1)} [\frac{(\rho^k-\rho^{kn})}{(1-\rho^k)^2} - (n-1) \frac{\rho^{kn}}{(1-\rho^k)}]\}. \qquad (3.9)$$

Letting n index a sequence with k fixed we obtain the following approximation to the expected variance:

$$\mathcal{E}\mathrm{Var}\{\bar{y}\} = (1-f)(\sigma^2/n)\{1 - \frac{2}{(k-1)} \frac{\rho}{(1-\rho)} + \frac{2k}{(k-1)} \frac{\rho^k}{(1-\rho^k)}\} + O(n^{-2}). \qquad (3.10)$$

The expectations of the eight estimators of variance are presented in column 4 of Table 2. The expression for v_8 is a large-n approximation, as in (3.10), whereas the other expressions are exact. Large-n approximations to the expectations of v_1 and v_7 are given by

$$\mathcal{E}E\{v_{sy1}\} = (1-f)\sigma^2/n + O(n^{-2}) \tag{3.11}$$

$$\mathcal{E}E\{v_{sy7}\} = (1-f)(\sigma^2/n)\{1+[2/(p-1)][p\rho^{pk}/(1-\rho^{pk})-\rho^k/(1-\rho^k)]\} + O(n^{-2}). \tag{3.12}$$

The expectations of the remaining estimators (v_2 to v_6) do not involve terms of lower order than $O(n^{-1})$.

From Table 2 and (3.10) - (3.12), it is apparent that each of the eight estimators has small bias for ρ near zero. If k is reasonably large, then v_1 is only slightly biased regardless of the value of ρ, provided ρ is not very close to 1. This is also true of estimators v_2 through v_8. The expectation of the first estimator tends to be larger than those of the other estimators since, e.g.,

$$\mathcal{E}E\{v_{y1}\} - \mathcal{E}E\{v_{y4}\} \doteq (1-f)(\sigma^2/n)\{(4/3)\rho^k - (1/3)\rho^{2k}\} > 0.$$

As Cochran (1946) noticed, a good approximation to $-2\,\rho/k(1-\rho)$ is given by $2/\ell n(\rho^k)$. On this basis, v_8 should be a very good estimator since the expectation $\mathcal{E}E\{v_8\}$ is nearly identical with the expected variance in (3.10).

Exact statements about the comparative biases of the various estimators depend on the values of ρ and k. In Table 4 we see that differences between the estimator biases are negligible for small ρ, and increase as ρ increases. For a given value of ρ, the differences decline with increasing sampling interval k. Estimator v_8 tends to underestimate the variance, while the remaining estimators (most notably v_1) tend towards an overestimate. Further, v_8 tends to have the smallest absolute bias, except when ρ is small. When ρ is small, the $\ell n(\rho^k)$ approximation is evidently not very satisfactory.

3.4 Periodic Populations

A simple periodic population is given by

$$\mu_{ij} = \beta_0 \sin\{\beta_1[i+(j-1)k]\}, \qquad (3.13)$$

with e_{ij} iid $(0, \sigma^2)$. As is well known, such populations are the nemesis of systematic sampling, and we study them here only to display that fact. When the sampling interval is equal to a multiple of the period, the variance of $\bar{y}$ tends to be enormous while all of the estimators of variance tend to be very small. Conversely, when the sampling interval is equal to an odd multiple of the half period, $\mathrm{Var}\{\bar{y}\}$ tends to be extremely small while the estimators of variance tend to be large.

Table 4, Expected Relative Biases of Eight Estimators for Autocorrelated Populations

First Order Autocorrelation Coefficient ρ	Sampling Interval k	Estimator							
		v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
0.01	4	0,678-02	0,678-02	0,678-02	0,678-02	0,678-02	0,678-02	0,678-02	-0.103-00
	10	0.225-02	0.225-02	0.225-02	0.225-02	0,225-02	0,225-02	0.225-02	-0.413-01
	30	0.697-03	0,697-03	0.697-03	0.697-03	0.697-03	0.697-03	0.697-03	-0.138-01
0.10	4	0.797-01	0.796-01	0.796-01	0.795-01	0.795-01	0,795-01	0.795-01	-0.155-00
	10	0,253-01	0.253-01	0.253-01	0.253-01	0.253-01	0.253-01	0.253-01	-0.637-01
	30	0.772-02	0,772-02	0,772-02	0.772-02	0.772-02	0,772-02	0,772-02	-0.215-01
0.50	4	0.957+00	0.834+00	0.834+00	0.796+00	0.755+00	0.740+00	0.841+00	-0.194+00
	10	0.282+00	0.281+00	0.281+00	0.280+00	0.280+00	0.280+00	0.281+00	-0.853-01
	30	0.741-01	0.741-01	0.741-01	0.741-01	0.741-01	0.741-01	0,741-01	-0.292-01
0.90	4	0.104+02	0.293+01	0.293+01	0.207+01	0.165+01	0.150+01	0.590+01	-0.200+00
	10	0.427+01	0.243+01	0.243+01	0.204+01	0.174+01	0.163+01	0.291+01	-0.907-01
	30	0.112+01	0.103+01	0.103+01	0.100+01	0.974+00	0.961+00	0,104+01	-0,321-01
0.99	4	0.118+03	0.370+01	0.370+01	0.220+01	0.174+01	0.156+01	0.599+02	-0.200+00
	10	0.533+02	0.419+01	0.419+01	0.263+01	0.211+01	0.190+01	0.275+02	-0.909-01
	30	0.183+02	0.402+01	0.402+01	0.278+01	0.225+01	0.205+01	0.101+02	-0.323-01

NOTE: Results ignore terms of order n^{-2}.

3.5 Monte Carlo Results

The Monte Carlo results for the random population are presented in the row labeled A1 of Tables 6, 7, and 8. On the basis of this investigation, estimator v_1 seems the best choice in terms of both minimum MSE and the ability to produce 95 percent confidence intervals. Estimator v_8 is the only estimator that is seriously biased. The variance of the variance estimators is related to the number of "degrees of freedom", and on this basis v_1 is the preferred estimator. The actual confidence levels are lower than the nominal rate in all cases.

For the linear trend population (see row labeled A2), all of the estimators are seriously biased. v_2, v_3, and particularly v_8 seem to be more acceptable than the remaining estimators, although each is downward biased and actual confidence levels are lower than the nominal rate of 95 percent. The good performance of v_8 is surprising because this estimator was constructed specifically for autocorrelated populations. Because of large bias, v_1 and v_7 are particularly unattractive for populations with linear trend.

The Monte Carlo results for the stratification effects populations are presented in rows labled A3 and A4. Population A4 is essentially the same as A3, except truncated so as not to permit negative values. Estimators v_2, v_3, and v_4 are clearly preferred here; they have smaller absolute bias and MSE than the remaining estimators. v_5 and v_6 have equally small bias but

Table 5. Description of the Artificial Populations

Population	Description	n	k	μ_{ij}	e_{ij}
A1	Random	20	50	0	e_{ij} iid N(0,100)
A2	Linear Trend	20	50	$1 + (j-1)k$	e_{ij} iid N(0,100)
A3	Stratification Effects	20	50	j	e_{ij} iid N(0,100)
A4	Stratification Effects	20	50	$j + 10$	$e_{ij} = \varepsilon_{ij}$ if $\varepsilon_{ij} \geq -(j+10)$ $= -(j+10)$ otherwise ε_{ij} iid N(0,100)
A5	Autocorrelated	20	50	0	$e_{ij} = \rho e_{i-1,j} + \varepsilon_{ij}$ $e_{11} \sim N(0,100/(1-\rho^2))$ e_{ij} iid N(0,100) $\rho = 0.8$
A6	Autocorrelated	20	50	0	same as A5 with $\rho = 0.4$
A7	Periodic	20	50	$20 \sin\{(2\pi/50)[i+(j-1)k]\}$	e_{ij} iid N(0,100)

larger variance, presumably because of a deficiency in the "degrees of freedom." Primarily because of large bias, estimators v_1, v_7, and v_8 are unattractive for populations with stratification effects.

Results for the autocorrelated populations are in rows A5 and A6. Estimator v_8 performs well in the highly autocorrelated population (A5), but not as well in the moderately autocorrelated population (A6). Even in the presence of high autocorrelation, the actual confidence level associated with v_8 is low. Any one of the first four estimators could be recommended for low autocorrelation.

Row A7 gives the results of the Monte Carlo study of the periodic population. As was anticipated (because the sampling interval $k = 50$ is equal to the period) all of the eight estimators are badly biased downward, and the associated confidence intervals are completely unusable.

All simulations presented here were performed on UNIVAC 1100 series computers, using the EXEC 8 operating system. The programming language was FORTRAN V. All random numbers were generated by IMSL subroutines. Computations were made in single precision, giving about 8-9 decimal place accuracy.

4. Some Numerical Comparisons

We now compare the eight estimators of variance using eight real populations. The first six populations were taken from the Income Supplement to the March, 1981 Current Population Survey (CPS). The populations consisted of all persons age 14+, in the

U.S. civilian labor force, and living in one of the ten largest Standard Metropolitan Statistical Areas (SMSA). For three of the populations, EMPINC, EMPRSA, and EMPNOO, the y-variable was the unemployment indicator

$$y = 1, \text{ if unemployed}$$
$$= 0, \text{ if employed}$$

while for the remaining three populations, INCINC, INCRSA, and INCNOO, the y-variable was total income. EMPINC and INCINC were ordered by the median income of the census tract in which the person resided. EMPRSA and INCRSA were ordered by the person's race by sex by age (white before black before other, male before female, age in natural ascending order). EMPNOO and INCNOO were in the customary CPS file order, essentially a geographic ordering. These CPS populations were each of size N = 13,000.

The last two populations, FUELID and FUELAP, were comprised of 6,500 fuel oil dealers from the 1972 Economic Censuses. The y-variable was 1972 annual sales in both cases. FUELID was ordered by State by identification number. The nature of the identification number was such that within a given State, the order was essentially random. FUELAP was ordered by 1972 annual payroll.

The populations INCINC, INCRSA, and INCNOO are depicted in Figures A, B, and C (these figures actually depict a 51-term centered moving average of the data). The ordering by median income (INCINC) results in an upward trend, possibly linear at first and then sharply increasing at the upper tail of the income distribution. There are rather distinct stratification effects

for the population INCRSA, where the ordering is by race by sex by age. The geographical ordering displays characteristics of a random population.

The unemployment populations EMPINC, EMPRSA, and EMPNOO are similar in appearance to INCINC, INCRSA, and INCNOO, respectively, except that they display negative relationships between the y variable and the sequence number whenever the income populations display positive relationships, and vice-versa.

The fuel oil population FUELAP is similar in appearance to INCINC, except the trend is much stronger in FUELAP than in INCINC. FUELID appears to be a random population, or possibly a population with weak stratification effects (due to a State or regional effect).

The results of our investigation of bias are presented in Tables 6, 7, and 8, where the sampling fraction is f = 0.02 for all eight populations. In general, the results for these real populations are similar to the Monte Carlo results presented in the last section.

Populations with a Trend

Any of the five estimators $v_2, \ldots, v_6$ may be recommended for INCINC. For FUELAP, (which has stronger trend than INCINC) v_2 and v_3 are the least biased estimators and also provide confidence levels closest to the nominal rate. The estimator v_1 was shatteringly bad for both of these populations. For EMPINC (which has much weaker trend than INCINC), however, the first estimator v_1 peformed as well as any of the estimators $v_2, \ldots, v_6$.

Populations with Stratification Effects

Any of the three estimators v_2, v_3, v_4 may be recommended for the populations INCRSA and EMPRSA. The absolute bias of v_1 tends to be somewhat larger than the biases of these preferred estimators. All of the preferred estimators are downward biased for INCRSA and, thus, actual confidence levels are too low. Estimator v_6 has larger MSE than the preferred estimators.

Random Populations

Any of the first six estimators may be recommended for INCNOO, EMPNOO, and FUELID. The last estimator also performs quite well for these populations, except for FUELID where it has a larger downward bias and corresponding confidence levels are too low.

Table 6. Relative Bias of Eight Estimators of $\mathrm{Var}\{\bar{y}\}$

Population	Estimator of Variance							
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
A1	0.047	0.046	0.043	0.046	0.049	0.053	0.060	-0.237
A2	19.209	-0.689	-0.688	-0.977	-0.977	-0.977	1.910	-0.449
A3	0.419	0.051	0.049	0.046	0.050	0.054	0.116	-0.443
A4	0.416	0.051	0.047	0.048	0.057	0.067	0.116	-0.441
A5	0.243	0.236	0.234	0.230	0.234	0.243	0.263	-0.095
A6	0.073	0.071	0.069	0.070	0.073	0.075	0.084	-0.217
A7	-0.976	-0.976	-0.976	-0.976	-0.976	-0.976	-0.976	-0.983
FUELID	-0.191	-0.220	-0.212	-0.223	-0.234	-0.256	-0.517	-0.437
FUELAP	1.953	-0.251	0.104	-0.544	-0.641	-0.698	0.693	-0.601
EMPINC	-0.184	-0.195	-0.193	-0.191	-0.188	-0.208	-0.158	-0.402
EMPRSA	0.316	0.241	0.239	0.234	0.235	0.234	0.100	-0.280
EMPNOO	0.121	0.123	0.119	0.134	0.151	0.148	0.707	-0.155
INCINC	0.398	0.279	0.290	0.279	0.268	0.219	0.214	-0.256
INCRSA	0.210	-0.139	-0.148	-0.143	-0.156	-0.171	-0.450	-0.748
INCNOO	0.662	0.659	0.650	0.658	0.668	0.660	0.547	0.272

Table 7. Relative Mean Square Error (MSE) of Eight Estimators of $Var\{\bar{y}\}$

Population	Estimator of Variance							
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
A1	0.158	0.212	0.262	0.272	0.467	0.954	02.322	0.294
A2	369.081	0.476	0.479	0.957	0.957	0.957	03.923	0.204
A3	0.417	0.213	0.263	0.272	0.467	0.954	02.549	0.442
A4	0.386	0.200	0.249	0.261	0.464	0.973	02.555	0.441
A5	0.377	0.439	0.505	0.509	0.765	1.446	03.363	0.430
A6	0.180	0.236	0.286	0.296	0.491	0.982	02.367	0.307
A7	0.955	0.955	0.955	0.955	0.955	0.955	00.955	0.967
FUELID	1.173	1.186	1.150	1.199	1.163	1.109	00.746	0.943
FUELAP	16.761	1.969	7.229	0.547	0.513	0.544	14.272	1.455
EMPINC	0.060	0.067	0.068	0.068	0.072	0.095	00.897	0.241
EMPRSA	0.142	0.104	0.115	0.109	0.144	0.206	03.706	0.196
EMPNOO	0.051	0.059	0.065	0.066	0.084	0.112	04.846	0.153
INCINC	0.267	0.185	0.192	0.200	0.199	0.200	02.620	0.247
INCRSA	0.120	0.084	0.087	0.091	0.109	0.132	00.601	0.569
INCNOO	.554	0.574	0.563	0.585	0.613	0.654	04.865	0.383

Table 8. Proportion of Times that the True Population Mean Fell within the Confidence Interval formed Using One of Eight Estimators of Variance

Population	Estimator of Variance							
	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
A1	94	93	93	93	91	86	70	85
A2	100	64	64	17	17	16	100	85
A3	97	93	93	93	91	86	71	77
A4	97	93	93	93	91	86	71	77
A5	96	95	94	94	92	88	73	88
A6	94	94	93	93	91	86	71	86
A7	14	14	14	13	13	13	11	11
FUELID	88	86	82	84	82	80	60	74
FUELAP	100	90	88	86	84	82	80	76
EMPINC	90	90	88	90	88	88	64	84
EMPRSA	96	94	94	94	94	96	74	88
EMPNOO	98	98	98	98	98	98	76	92
INCINC	98	94	94	94	94	96	74	88
INCRSA	94	90	90	90	90	88	76	70
INCNOO	98	98	98	98	100	100	76	92

Figure A. Plot of Total Income Versus Sequence Number for Population INCINC

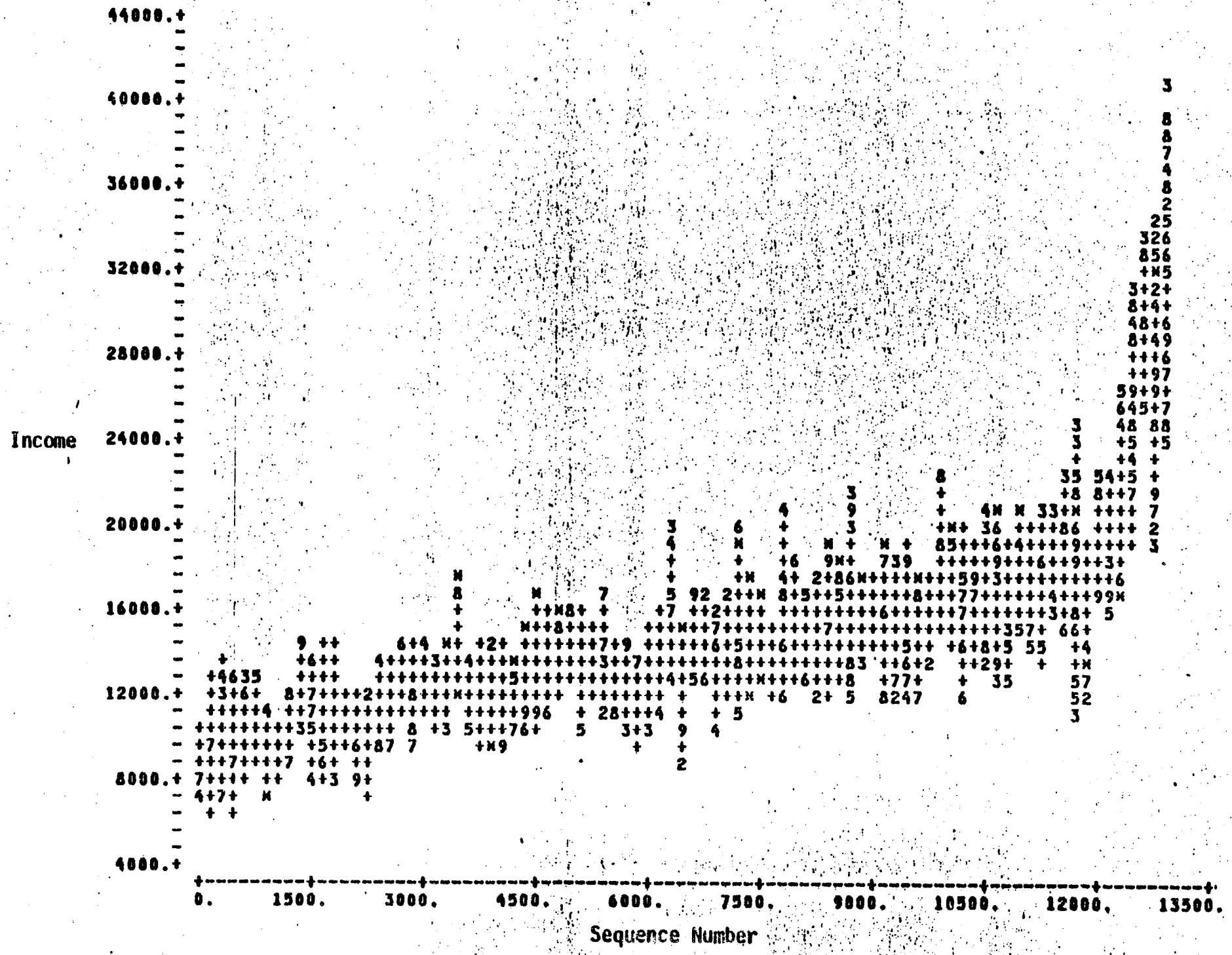


Figure B. Plot of Total Income Versus Sequence Number for Population INCRSA

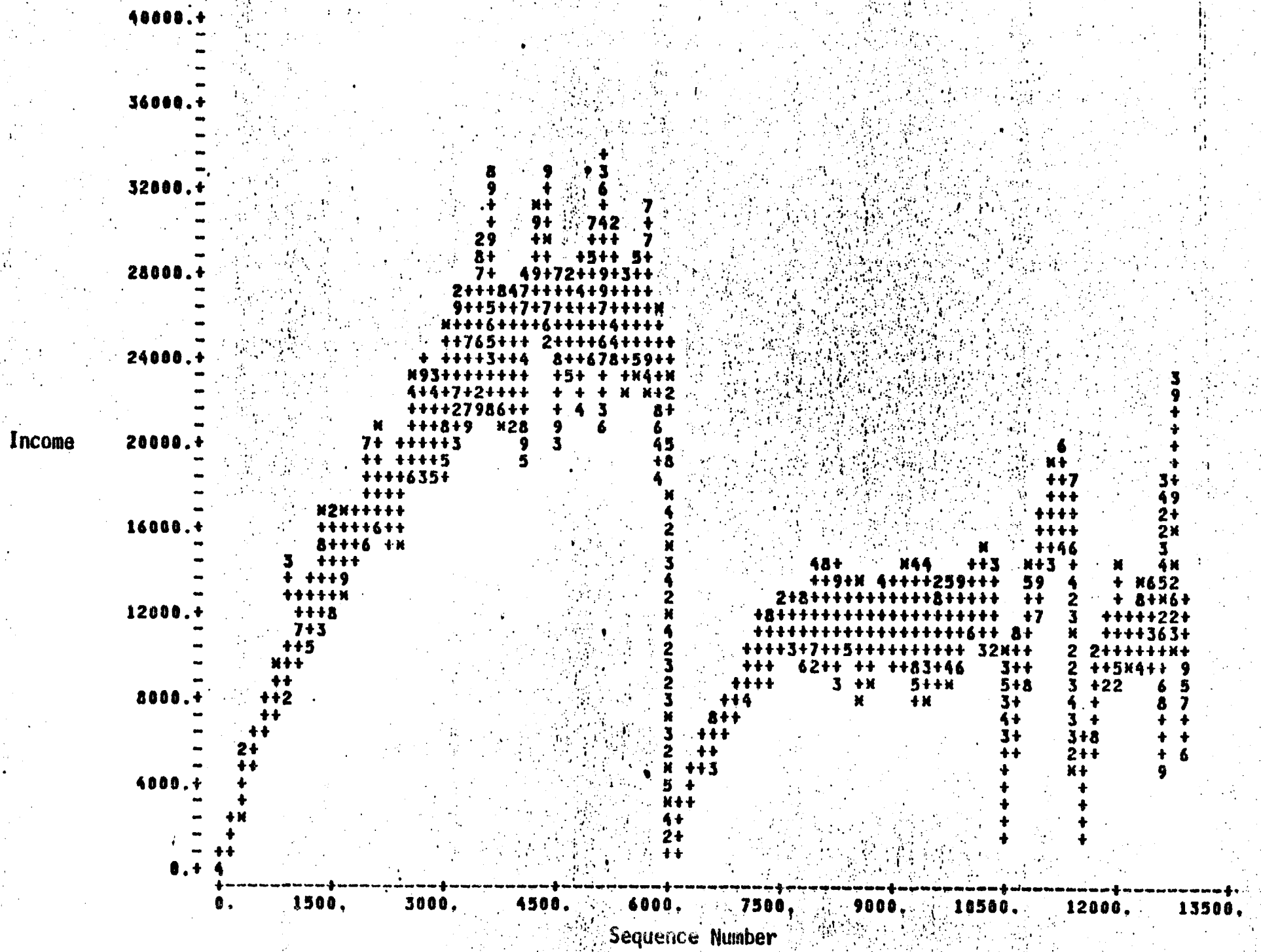
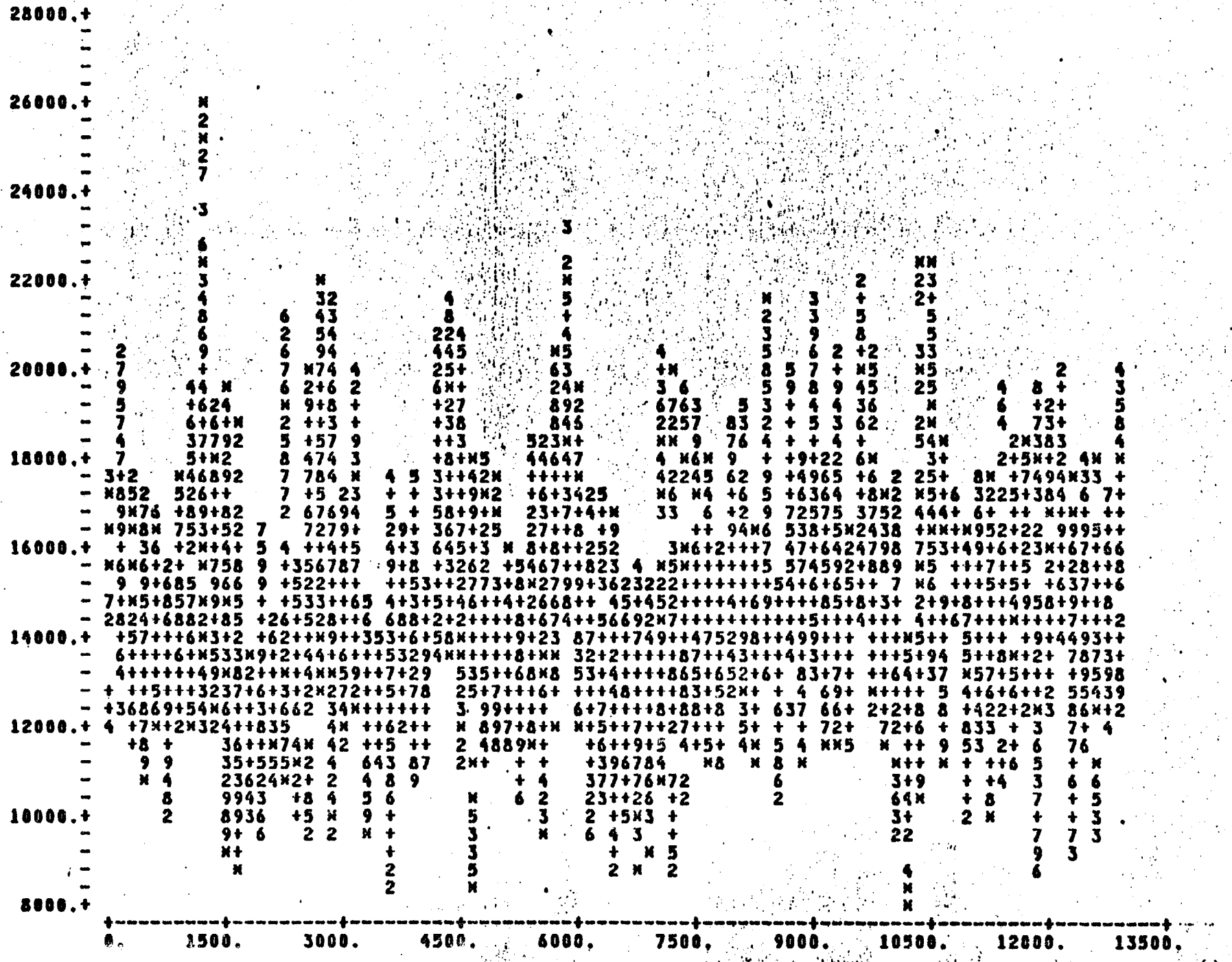


Figure C. Plot of Total Income Versus Sequence Number for Population INCNOO



5. Summary

In this article we have studied some of the theoretical and empirical properties of eight reasonable estimators of the variance of the sample mean, attempting to gain some understanding of their range of applicability. Based on this work we offer the practicing statistician the following general advice:

> Reiterating our introduction, acquire as much prior information as possible about the population of interest. Use the information to construct plausible models of the population and to try out different reasonable variance estimators. Subsequently, choose a variance estimator(s) for implementation. Intelligent use of prior information provides the best hope for finding a variance estimator with good statistical properties. If all else fails, we like v_2, v_3, and possibly v_4. Based upon the work in Sections 3 and 4, these estimators seem to be broadly useful for a variety of kinds of populations.

The specific findings of our investigation are as follows:

(1) The bias and MSE of the simple random sampling estimator v_1 are reasonably small for all populations which have approximately constant mean u_{ij}. This excludes populations with a strong trend in the mean or stratification effects. Confidence intervals formed from v_1 are relatively good overall, though are often

too wide and lead to true confidence levels exceeding the nominal level.

(2) In relation to v_1, the estimators v_4, v_5, v_6 based on higher order differences provide protection against a trend, autocorrelation, and stratification effects. They are often good for the approximate random populations as well. v_4 often has the smallest MSE of these three, because the variances of v_5 and v_6 are large when the sample size (and thus the number of differences) is small. In larger samples and in samples with nonlinear trend or complex stratification effects, these estimators should perform relatively better than they did in this study. Confidence intervals are basically good, except when there is a pure linear trend in the mean.

(3) The bias of Koop's estimator v_7 is unpredictable, and its variance is generally too large to be useful. This estimator cannot be recommended on the basis of the work done here. An issue for further work, however, concerns the possibility of increasing p (cf. Section 2.1). This may reduce the variance of v_7 enough to make it useful in real applications.

(4) Estimator v_8 has remarkably good properties for the artificial populations with linear trend or autocorrelation, otherwise it is quite mediocre. Its bias is usually negative, and consequently, confidence intervals formed from v_8 can fail to cover the true

population mean a sufficient proportion of the time. This estimator seems too sensitive to the form of the model to be broadly useful in real applications.

(5) The estimators v_2 and v_3 based on simple differences afford the user considerable protection against most model forms studied in this article. They are susceptible to bias for populations with strong stratification effects. They are also biased for the linear trend population, but even then the other estimators are more biased. Stratification effects and trend did occur in the real populations, but they were not sufficiently strong to defeat the good properties of v_2 and v_3. In the real populations these estimators performed, on average, as well as any of the estimators. Estimators v_2 and v_3 (more degrees of freedom) often have smaller variance than estimators v_4, v_5, and v_6 (fewer degrees of freedom). In very small samples, v_2 might be the preferred estimator.

Finally, we note that the findings presented in this article apply primarily to surveys of establishments and people. Stronger correlation patterns are likely to exist in surveys of land use, forestry, geology and the like, and the properties of the estimators may be somewhat different in these applications.

References

Cochran, W.G. (1946), "Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations," Annals of Mathematical Statistics, 17, 164-177.

_____ (1977), Sampling Techniques, New York: John Wiley.

Gautschi, W. (1957), "Some Remarks on Systematic Sampling," Annals of Mathematical Statistics, 28, 385-394.

Heilbron, D.C. (1978), "Comparison of the Estiators of the Variance of Systematic Sampling," Biometrika, 65, 429-433.

Koop, J.C. (1971), "On Splitting a Systematic Sample for Variance Estimation," Annals of Mathematical Statistics, 42, 1084-1087.

Madow, W.G. and L. H. (1944), "On the Theory of Systematic Sampling," Annals of Mathematical Statistics, 15, 1-24.

Matérn, B. (1947), "Methods of Estimating the Accuracy of Line and Sample Plot Surveys," Medd. Fr. Statens Skogsforskinings-institut, 36, 1-138.

Osborne, J.G. (1942), "Sampling Errors of Systematic and Random Surveys of Cover-Type Areas," Journal of the American Statistical Association, 37, 256-264.

Wu, C. (1981), "Estimation in Systematic Sampling with Supplementary Observations," University of Wisconsin, Madison, unpublished manuscript.

Yates, F. (1949), Sampling Methods for Censuses and Surveys, London: Griffin.

Zinger, A. (1980), "Variance Estimation in Partially Systematic Sampling," Journal of the American Statistical Association, 75, 206-211.