ALTERNATIVE ESTIMATORS OF VARIANCE FOR SYSTEMATIC SAMPLING

Kirk M. Wolter and Shana McCann
U.S. Bureau of the Census

1. Introduction

This paper is concerned with the problem of estimating the variance of the sample mean, say $\bar{y}_{sy}$, when the sample is drawn systematically from a finite population of size N. We shall only consider equal probability systematic sampling with a single, random start. Unequal probability systematic sampling or sampling with two or more random starts will not be treated here.

In the 1940's several authors addressed the issue of variance estimation for systematic samples, including Osborne (1942), Cochran (1946), Matérn (1947), and Yates (1949). One of the most comprehensive discussions is given by Cochran (1963). A more recent reference is Koop (1971). Little in the way of empirical comparisons of alternative estimators is available in this literature. In recent years, the topic appears to have received little attention, no doubt because systematic sampling is often used at the last stage of sampling, a case where rigorous estimates of the variance can be given. However, there remain many surveys where an estimate of $\mathrm{Var}\{\bar{y}_{sy}\}$ is required. In such cases we have noticed a tendency on the part of many researchers to regard the sample as random, and, in the absence of knowing what else to do, to estimate the variance using random sample formulae. This practice often leads to badly biased estimates of variance, and to incorrect inferences concerning the population mean.

In the remainder of this paper we shall empirically investigate eight estimators of the variance of $\bar{y}_{sy}$. Our goal is to provide some guidance about when a given estimator may be more appropriate than other estimators. The estimators are defined in Section 2. In Section 3, the various populations used in our study are described. The results of the comparison are then summarized in Sections 4 and 5.

2. Description of the Estimators

Throughout our investigation we assume N=nk where n is the sample size and k is an integer. We let Y_{ij} denote the value of the y-variable for the j-th unit in the i-th systematic sample, where i=1, ...,k and j=1,...,n. Then, the eight estimators of variance for the i-th selected sample are defined as follows:

1. $v_{sy1}(i) = \frac{N-n}{Nn} \frac{1}{n-1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{sy})^2$.

2. $v_{sy2}(i) = \frac{N-n}{Nn} \sum_{j=1}^{n-1} (y_{ij} - y_{i,j+1})^2/2(n-1)$.

3. $v_{sy3}(i) = \frac{N-n}{Nn} \sum_{j=1}^{n/2} (y_{i,2j-1} - y_{i,2j})^2/n$.

4. $v_{sy4}(i) = \frac{N-n}{N} \frac{n'}{n^2} \sum_{j=1}^{n-2} (y_{ij} - 2y_{i,j+1} + y_{i,j+2})^2/6(n-2)$,

where $\frac{n'}{n^2} = \left(\frac{1}{n} + \frac{2i-k-1}{2(n-1)k}\right)^2 + \frac{n-2}{n^2} + \left(\frac{1}{n} - \frac{2i-k-1}{2(n-1)k}\right)^2$.

5. $v_{sy5}(i) = \frac{1}{4}(\bar{y}_A - \bar{y}_B)^2$, where $\bar{y}_A$ is the is the mean of the even numbered members of the sample and $\bar{y}_B$ is the mean of the odd numbered members.

6. $v_{sy6}(i) = \frac{N-n}{Nn} \sum_{j=1}^{n-4} c_{ij}^2/3.5(n-4)$, where

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

7. $v_{sy7}(i) = \frac{N-n}{Nn} \sum_{j=1}^{n-8} d_{ij}^2/7.5(n-8)$, where

$$d_{ij} = \frac{1}{2}y_{ij} - y_{i,j+1} + y_{i,j+2} - y_{i,j+3} + y_{i,j+4} - y_{i,j+5} + y_{i,j+6} - y_{i,j+7} + \frac{1}{2}y_{i,j+8} .$$

8. $v_{sy8}(i) = \frac{N-n}{Nn} s^2 \left\{1 + \frac{2}{\ln\hat{\rho}_k} + \frac{2}{\left(\hat{\rho}_k^{-1}-1\right)}\right\}$,

$= \frac{N-n}{Nn} s^2$ if $\hat{\rho}_k > 0$
if $\hat{\rho}_k \leq 0$,

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

$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{sy})^2$.

v_{sy1} is the estimate of variance for simple random sampling. v_{sy2} and v_{sy3} are based on overlapping and nonoverlapping differences, respectively. v_{sy4}, v_{sy6}, and v_{sy7} are based on higher order contrasts. Koop's (1971) estimator, v_{sy5}, is obtained by splitting the systematic sample into equal halves. v_{sy8} was devised from an assumption about the correlogram (cf. Cochran (1946)).

3. Description of the Populations

3.1 The Artificial Populations

Sixteen artificial populations, each of size N=1000, were generated according to the simple model

$$Y_{ij} = \mu_{ij} + u_{ij}, \qquad (2.1)$$

where the μ_{ij} denote fixed constants and the errors u_{ij} are drawn from some infinite superpopulation. The reader will recognize (2.1) as the model employed by Cochran (1963). The eight estimators were evaluated using the sixteen populations and four sampling fractions: $f=k^{-1}$=.01,.02,.1 and .25. Due to limited space, only seven populations and two sampling fractions, f=.01 and .02, will be discussed here.

The seven populations for which results will be presented and the specific assumptions about the μ_{ij} and u_{ij} are described in Table 1. For notational convenience, we shall employ the

Table 1. Description of the Artificial Populations

Code	Description	μ_{ij}	u_{ij}
A1	Random	0	u_{ij} iid $\Gamma(2,11.32)$
A2	Linear Trend	$i+(j-1)k$	u_{ij} iid N(0,2.25)
A3	Stratification Effects	$\mu_{.j}$	u_{ij} iid N(0,9)
A4	Stratification Effects	$\mu_{.j}$	u_{ij} iid N(0,9)
A5	First Order Autocorrelated	0	$u_{ij}=\rho u_{i-1,j}+e_{ij}$; $u_{11}\sim N\left(0,\frac{55.43}{(1-\rho^2)}\right)$; e_{ij} iid N(0,55.43); ρ = .9
A6	First Order Autocorrelated	0	$u_{ij}=\rho u_{i-1,j}+e_{ij}$; $u_{11}\sim N\left(0,\frac{190.84}{(1-\rho^2)}\right)$; e_{ij} iid N(0,190.84); ρ = .5
A7	Periodic	$2\sin\frac{\pi i}{2}$	u_{ij} iid N(0,.07)

population codes in future references to these populations. Population A3 was only used with the f=.01 sampling fraction and the $\mu_{.j}$'s took the values 8, 42, 70, 90, 99, 96, 81, 57, 24, and 8. Similarly, population A4 was only used with the f=.02 sampling fraction and the $\mu_{.j}$'s took the values 0, 17, 34, 50, 64, 76, 86, 94, 98, 100, 98, 94, 86, 76, 64, 50, 34, 17, 0, and 17. Each of the remaining populations was studied for both f=.01 and .02.

Of the 9 populations for which results are not being presented, three were random, three had a linear trend, two had stratification effects, and one was autocorrelated.

3.2 The Real Populations

The estimators of variance were also compared on the basis of six real populations obtained from Census Bureau files. The first two populations, R1 and R2, were comprised of 6900 fuel oil dealers from the 1972 Economic Census. The y-variable was annual sales in both cases. R1 was sorted by multi- versus single-unit firms, by State, and by ID number. The nature of the ID number was such that within a given class of firms within a given State, the sort was essentially random. R2 was sorted by annual payroll.

The remaining four populations were from the Income Supplement to the March, 1975 Current Population Survey (CPS). A one-in-five sample of persons in the civilian labor force and living in SMSA's of 250,000 population or more was the basis for these populations. For R3 and R4 the y-variable was the unemployment indicator

$$y = \begin{cases} 1, & \text{if unemployed} \\ 0, & \text{if employed} \end{cases}$$

while in R5 and R6 the y-variable was total income. R3 and R5 were in sort by two census tract characteristics: "non-whites as a percent of the total population" and "persons with four or more years of high school as a percent of all persons 25 years old or older." R4 and R6 were in sort by the census tract characteristic "median family income." Populations R3, R4, R5, and R6, were each of size N=11300.

4. Empirical Results

Some of the results of our investigation are presented in Tables 2, 3, and 4. Tables 2 and 3 give the relative biases and relative mean square errors (MSE) of the eight estimators, respectively. Table 4 presents the actual proportion of confidence intervals which contained the true population mean, where the confidence interval for the α-th estimator is of the form

$$\left(\bar{y}_{sy} - t_{n-1,.025}v_{sy\alpha}(i)\ ,\ \bar{y}_{sy} + t_{n-1,.025}v_{sy\alpha}(i)\right)$$

and $t_{n-1,.025}$ denotes the .025 percentage point of Student's t distribution with n-1 degrees of freedom. As noted in Section 3, the populations and sampling fractions reported in the tables comprise less than half of those actually studied. When describing the results, however, our remarks shall apply to all of the study populations, not merely the illustrative ones.

An important observation regarding these results is that our sample of populations is far too small to conclusively demonstrate estimator behavior. As a result, we shall not try to claim too much from our results. Our remarks will be limited to instances where, in our view, a reasonably consistent pattern of behavior was established.

Many additional commentaries could be given beyond those presented in this section. For example, one may wish to observe certain patterns of bias depending on the value of the intraclass correlation coefficient. However, such analyses will be left to the reader as our space is limited.

Table 2. Relative Bias of Eight Estimators of $V\{\bar{y}_{sy}\}$

Population Code	Sampling Fraction f	Estimator of Variance								Intraclass Correlation
		v_{sy1}	v_{sy2}	v_{sy3}	v_{sy4}	v_{sy5}	v_{sy6}	v_{sy7}	v_{sy8}	
R1	0.01	0.239	-0.416	-0.991	-0.587	-0.719	-0.582	-0.726	-0.705	-0.00292
R1	0.02	0.713	-0.321	-0.984	-0.513	0.792	-0.564	-0.552	-0.696	-0.00311
R2	0.01	0.505	0.218	-0.257	0.146	0.042	-0.034	-0.182	-0.253	-0.00429
R2	0.02	0.313	0.155	-0.328	0.142	-0.105	0.096	-0.065	-0.403	-0.00184
R3	0.01	-0.094	-0.121	0.122	-0.123	-0.218	-0.129	-0.138	-0.376	0.00082
R3	0.02	-0.146	-0.152	0.100	-0.144	-0.096	-0.134	-0.122	-0.388	0.00065
R4	0.01	0.106	0.111	0.109	0.114	0.269	0.105	0.086	-0.167	-0.00093
R4	0.02	0.114	0.101	0.140	0.096	0.035	0.084	0.073	-0.196	-0.00053
R5	0.01	0.381	0.065	-0.461	-0.007	0.245	-0.057	-0.113	-0.404	-0.00251
R5	0.02	0.349	0.073	-0.453	0.016	-0.023	-0.064	-0.123	-0.540	-0.00121
R6	0.01	-0.068	-0.069	-0.109	-0.063	-0.032	-0.072	-0.076	-0.307	0.00055
R6	0.02	-0.041	-0.048	-0.076	-0.050	0.078	-0.052	-0.040	-0.329	0.00010
A1	0.01	0.022	0.056	0.012	0.099	0.101	0.169	0.230	-0.204	-0.00317
A1	0.02	-0.068	-0.036	-0.008	-0.017	-0.147	-0.034	-0.090	-0.190	0.00255
A2	0.01	9.901	-0.405	-0.404	-1.000	2.008	-1.000	-1.000	-0.353	-0.10001
A2	0.02	19.652	-0.705	-0.705	-1.000	2.010	-1.000	-1.000	-0.441	-0.05001
A3	0.01	133.928	27.361	26.435	2.310	1.613	2.804	2.869	11.446	-0.11021
A4	0.02	146.639	9.616	9.824	1.401	4.627	0.787	0.712	3.875	-0.05226
A5	0.01	0.866	0.762	0.896	0.816	0.982	0.865	0.408	0.216	-0.04926
A5	0.02	1.011	0.532	0.318	0.360	0.226	0.126	-0.022	-0.235	-0.02361
A6	0.01	0.118	0.137	0.107	0.195	0.184	0.181	0.011	-0.144	-0.01161
A6	0.02	0.222	0.166	0.138	0.140	0.211	0.088	0.020	-0.229	-0.01000
A7	0.01	-0.996	-0.996	-0.996	-0.996	-0.996	-0.996	-0.997	-0.997	0.96502
A7	0.02	32.668	61.868	61.891	83.143	631.178	140.949	262.807	32.444	-0.05102

Table 3. Relative Mean Square Error (MSE) of Eight Estimators of $V\{\bar{y}_{sy}\}$

Population	Sampling Fraction f	v_{sy1}	v_{sy2}	v_{sy3}	v_{sy4}	v_{sy5}	v_{sy6}	v_{sy7}	v_{sy8}
R1	0.01	4.349	2.002	0.982	1.482	1.235	1.638	0.988	1.109
R1	0.02	4.123	1.435	0.969	1.094	12.722	1.080	1.160	0.809
R2	0.01	3.598	2.988	3.004	2.764	5.162	2.086	1.865	1.960
R2	0.02	1.391	1.298	1.389	1.364	1.245	1.431	1.131	0.823
R3	0.01	0.078	0.082	0.185	0.088	1.528	0.105	0.144	0.242
R3	0.02	0.053	0.054	0.078	0.055	1.596	0.061	0.076	0.215
R4	0.01	0.092	0.102	0.228	0.112	2.246	0.132	0.195	0.169
R4	0.02	0.055	0.070	0.161	0.079	2.842	0.099	0.132	0.201
R5	0.01	0.557	0.206	0.496	0.160	3.235	0.184	0.265	0.494
R5	0.02	0.275	0.137	0.343	0.115	1.963	0.103	0.119	0.398
R6	0.01	0.212	0.241	0.395	0.274	1.677	0.343	0.376	0.358
R6	0.02	0.138	0.143	0.166	0.142	2.146	0.149	0.213	0.235
A1	0.01	0.561	0.736	0.830	1.046	3.037	2.603	4.133	0.601
A1	0.02	0.238	0.339	0.394	0.449	1.624	0.684	1.183	0.343
A2	0.01	98.034	0.164	0.164	1.000	4.033	1.000	1.000	0.125
A2	0.02	386.213	0.497	0.497	1.000	4.043	1.000	1.000	0.194
A3	0.01	17990.670	751.789	724.542	6.066	9.285	8.959	11.640	131.725
A4	0.02	21532.965	93.282	99.893	2.476	45.284	1.266	1.942	15.102
A5	0.01	1.391	1.365	2.127	1.892	4.266	3.000	2.118	0.721
A5	0.02	1.359	0.593	0.398	0.501	1.418	0.454	0.582	0.393
A6	0.01	0.303	0.460	0.578	0.711	2.344	1.275	2.036	0.397
A6	0.02	0.209	0.299	0.250	0.380	3.596	0.563	1.190	0.413
A7	0.01	0.993	0.993	0.993	0.992	0.992	0.993	0.993	0.994
A7	0.02	2132.668	7678.381	7693.850	13882.034	799450.766	39904.320	138758.197	2132.783

Table 4. Proportion of Times that the True Population Mean Fell within the Confidence Interval formed Using One of Eight Estimators of Variance

Population Code	Sampling Fraction f	Estimator of Variance								$V\{\bar{y}_{sy}\}$
		v_{sy1}	v_{sy2}	v_{sy3}	v_{sy4}	v_{sy5}	v_{sy6}	v_{sy7}	v_{sy8}	
R1	0.01	0.99	0.75	0.19	0.71	0.47	0.67	0.56	0.59	$6.135 \cdot 10^{3}$
R1	0.02	1.00	0.84	0.18	0.84	0.82	0.80	0.78	0.64	$2.197 \cdot 10^{3}$
R2	0.01	0.91	0.88	0.79	0.87	0.65	0.86	0.84	0.74	$5.423 \cdot 10^{3}$
R2	0.02	0.90	0.86	0.80	0.86	0.68	0.86	0.84	0.76	$2.862 \cdot 10^{3}$
R3	0.01	0.93	0.93	0.92	0.93	0.58	0.93	0.93	0.83	$7.8 \cdot 10^{-4}$
R3	0.02	0.90	0.90	0.98	0.90	0.64	0.92	0.90	0.84	$4.1 \cdot 10^{-4}$
R4	0.01	0.96	0.96	0.89	0.96	0.69	0.95	0.94	0.92	$6.4 \cdot 10^{-4}$
R4	0.02	0.92	0.92	0.92	0.92	0.74	0.94	0.94	0.82	$3.1 \cdot 10^{-4}$
R5	0.01	0.97	0.97	0.83	0.96	0.73	0.96	0.93	0.83	$4.247 \cdot 10^{5}$
R5	0.02	0.88	0.88	0.82	0.86	0.74	0.86	0.84	0.76	$2.148 \cdot 10^{5}$
R6	0.01	0.92	0.92	0.89	0.92	0.74	0.91	0.90	0.85	$6.275 \cdot 10^{5}$
R6	0.02	0.92	0.90	0.90	0.90	0.70	0.90	0.92	0.84	$3.020 \cdot 10^{5}$
A1	0.01	0.97	0.94	0.91	0.94	0.74	0.89	0.76	0.87	$2.435 \cdot 10^{1}$
A1	0.02	0.92	0.90	0.88	0.88	0.62	0.82	0.74	0.82	$1.314 \cdot 10^{1}$
A2	0.01	1.00	1.00	1.00	0.02	1.00	0.01	0.00	1.00	$8.323 \cdot 10^{2}$
A2	0.02	1.00	0.66	0.66	0.02	1.00	0.02	0.02	0.90	$2.075 \cdot 10^{2}$
A3	0.01	1.00	1.00	1.00	1.00	0.89	1.00	1.00	1.00	$9.328 \cdot 10^{-1}$
A4	0.02	1.00	1.00	1.00	1.00	0.94	0.98	0.98	1.00	$4.071 \cdot 10^{-1}$
A5	0.01	0.99	0.97	0.97	0.98	0.90	0.94	0.80	0.93	$1.393 \cdot 10^{1}$
A5	0.02	1.00	0.98	1.00	0.96	0.84	0.88	0.86	0.84	$6.259 \cdot 10^{0}$
A6	0.01	0.95	0.93	0.90	0.93	0.71	0.89	0.73	0.90	$2.251 \cdot 10^{1}$
A6	0.02	0.96	0.92	0.94	0.94	0.70	0.92	0.90	0.86	$1.018 \cdot 10^{1}$
A7	0.01	0.49	0.48	0.45	0.48	0.36	0.43	0.38	0.45	$2.005 \cdot 10^{0}$
A7	0.02	1.00	1.00	0.96	0.96	0.92	0.94	0.94	0.96	$3.17 \cdot 10^{-3}$

4.1 Random Populations (A1)

Estimators v_{sy1}, v_{sy2}, v_{sy3}, and v_{sy4} were comparable and each displayed acceptable properties. v_{sy5} had a larger bias than $v_{sy1},\ldots,v_{sy4}$; its mean square error was extremely large; and it led to unacceptable confidence intervals. v_{sy8} tended to have the largest bias, but one of the smaller MSE's. v_{sy8} also produced slightly low confidence levels. v_{sy6} and v_{sy7} behaved similarly, with larger mean square error than $v_{sy1},\ldots,v_{sy4}$ and similar confidence levels to v_{sy8}.

4.2 Populations with Linear Trend (A2)

Remarkably, estimator v_{sy8} always produced the smallest bias, the smallest MSE, and the best confidence intervals (in the sense that confidence levels were nearest to 95 percent). v_{sy2} and v_{sy3} were comparable, producing lower confidence levels and larger bias and MSE than v_{sy8}. Estimators v_{sy1} and v_{sy5} had particularly bad properties.

4.3 Populations with Stratification Effects (A3,A4)

Estimators v_{sy2}, v_{sy3}, and particularly v_{sy1} were consistently bad both in terms of bias and MSE. This was undoubtedly due to our construction of the populations, with very large differences between successive values of $\mu_{\cdot j}$. The behavior of v_{sy5} and v_{sy8} was not firmly established with both estimators displaying relatively large and small biases for various populations. Estimators v_{sy4}, v_{sy6}, and v_{sy7} were comparable, usually having smaller bias and MSE than the other estimators. Confidence levels tended to be too high for all estimators except v_{sy5}, and all estimators tended to overestimate the true variance.

4.4 Autocorrelated Populations (A5, A6)

The results for autocorrelated populations depended largely on the value of ρ, i.e. the first order autocorrelation coefficient, and on the sampling fraction. For small to moderate values of ρ, the estimators behaved as they did for the random populations. For large ρ, v_{sy8} tended to have the smallest absolute bias and by far the smallest MSE. There was very little to choose between v_{sy1}, v_{sy2}, v_{sy3}, and v_{sy4} for large ρ: their MSE's were about twice that of v_{sy8} and likewise their biases. Confidence levels for intervals formed from v_{sy1}, v_{sy2}, v_{sy3}, v_{sy4} were very near to the nominal level of 95 percent, however. The differences between the estimators seemed to decrease as the sampling fraction increased.

4.5 Periodic Populations (A7)

As one would expect of periodic populations, the behavior of the estimators depended exclusively on the sampling fraction. However, for all sampling fractions studied, the eight estimators possessed nearly identical bias and MSE. In one case v_{sy1} and v_{sy8} had smaller MSE than the other estimators, but the MSE's were so large that it would make little practical difference which estimator was used.

4.6 Fuel Oil Dealers Sales (R1, R2)

Sort by Multi- versus Single-Unit by State and by ID Number: Estimator v_{sy1} overestimated the true variance, while the remaining estimators possessed a negative bias. v_{sy1} or v_{sy2} had the

smallest absolute bias, and v_{sy3} and v_{sy5} the largest. v_{sy3}, v_{sy7}, and v_{sy8} tended to have the smallest MSE. v_{sy4} and v_{sy6} had MSE's which were comparable to those of v_{sy3}, v_{sy7}, and v_{sy8} for large sampling fractions. The MSE's of the remaining estimators were larger. In spite of its small MSE, v_{sy3} led to extremely poor confidence intervals owing to its large bias. Except for v_{sy1}, whose MSE was too large, each of the estimators led to lower confidence levels than the anticipated 95 percent. v_{sy2}, v_{sy4}, and v_{sy6} seemed to give the best confidence intervals.

Sort by Annual Payroll: Estimator v_{sy6} tended to have the smallest absolute bias; v_{sy4}, v_{sy5}, and v_{sy7} also had relatively small absolute bias, but larger than v_{sy6}. The MSE's of v_{sy6}, v_{sy7} and v_{sy8} tended to be smaller than those of the other estimators. In particular, v_{sy5} had a very large MSE when f=.01. Among those estimators with relatively small bias, v_{sy4}, v_{sy6}, and v_{sy7} produced good confidence intervals, though the coverage rate was lower than expected. The population in this sort seemed to follow the linear model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij} \text{ , where } E\{u_{ij}|x_{ij}\} = 0,$$

$E\{u_{ij}^2|x_{ij}\} = \sigma^2 x_{ij}^g$, $g \in [\frac{1}{2}, 2]$, and x_{ij} denotes the annual payroll of the (i,j)-th unit.

4.7 CPS Unemployment (R3, R4)

Sort by % Nonwhite Etc. of Census Tract: The absolute biases of v_{sy1}, v_{sy2}, v_{sy3}, v_{sy4}, v_{sy6}, and v_{sy7} were comparable and relatively small, usually less than around 15%. The bias of v_{sy5} was also small when f=.02, but for f=.01 it exceeded 20%. The absolute bias of v_{sy8} was larger. Most estimators tended to underestimate $\mathrm{Var}\{\bar{y}_{sy}\}$. v_{sy1}, v_{sy2}, and v_{sy4} had the smallest MSE's, closely followed by v_{sy3}, v_{sy6}, and v_{sy7}. The MSE's of v_{sy8} and particularly v_{sy5} were larger. Most estimators led to acceptable confidence intervals except v_{sy5} and v_{sy8}, where the confidence levels were very low and slightly low, respectively.

Sort by Median Family Income of Census Tract: Estimator v_{sy8} tended to have larger bias than the other estimators. Also the bias of v_{sy8} was negative, while all other estimators tended to overestimate $\mathrm{Var}\{\bar{y}_{sy}\}$. v_{sy1}, v_{sy2}, v_{sy6} and v_{sy4} tended to have the smallest MSE, followed by v_{sy8}, v_{sy7}, and v_{sy3}. The MSE of v_{sy5} was much larger. All estimators produced acceptable confidence intervals, with the exception of v_{sy5} whose confidence level was too low.

4.8 CPS Income (R5, R6)

Sort by % Nonwhite Etc. of Census Tract: v_{sy2}, v_{sy4}, v_{sy6}, and v_{sy7} tended to have the smallest bias, though v_{sy5} also had small bias when f=.02. v_{sy1}, v_{sy3}, and v_{sy8} had larger biases. The MSE's of v_{sy2}, v_{sy4}, v_{sy6}, and v_{sy7} were comparable and relatively small. v_{sy1}, v_{sy3}, v_{sy8}, and particularly v_{sy5} had larger MSE's. Confidence intervals formed from v_{sy3}, v_{sy5}, and v_{sy8} had low coverage rates.

Sort by Median Family Income of Census Tract: Most of the estimators tended to underestimate $\mathrm{Var}\{\bar{y}_{sy}\}$. The biases of all estimators were of the same order of magnitude except v_{sy8}, which was larger. v_{sy1}, v_{sy2}, and v_{sy4} tended to have the smallest MSE. v_{sy3}, v_{sy6}, v_{sy7}, and v_{sy8} also displayed consistently small MSE, while the MSE of v_{sy5} was much larger. Each of the estimators except v_{sy5} gave acceptable confidence intervals, though the confidence levels were lower than 95 percent.

5. Detailed Analysis of Populations With Linear Trend

One of the interesting aspects of the results in Section 4 was the performance of the estimator v_{sy8}. In a variety of circumstances this estimator had relatively small bias and MSE and gave useable confidence intervals. This was particularly true of the populations with linear trend, even though v_{sy8} was constructed for another purpose (i.e. autocorrelated populations). This led us to question whether the behavior observed was a unique attribute of the particular populations studied, or was a more general result characteristic of all populations with linear trend. A partial answer to this question can be provided by obtaining the expected bias of each estimator of variance.

Towards this end, we assume the finite population is generated according to (2.1), with

$$\mu_{ij} = \beta_0 + \beta_1(i + (j-1)k)$$

and

$$u_{ij} \text{ iid } (0, \sigma^2).$$

If we let E denote the expectation with respect to the superpopulation, then the expected bias and expected relative bias of the α-th estimator are defined by

$$B\{v_{sy\alpha}\} = E\,E\{v_{sy\alpha}\} - EV\{\bar{y}_{sy}\}$$

and

$$R\{v_{sy\alpha}\} = B\{v_{sy\alpha}\} \,/\, EV\{\bar{y}_{sy}\} \ ,$$

respectively. It can then be shown that

$$R\{v_{sy1}\} = \frac{\beta_1^2(N-1)}{\beta_1^2(k+1) + 12\sigma^2/N} \ , \qquad (5.1)$$

$$R\{v_{sy2}\} = \frac{\beta_1^2(6k-N-n)/n}{\beta_1^2(k+1) + 12\sigma^2/N} \ , \qquad (5.2)$$

$$R\{v_{sy3}\} = R\{v_{sy2}\} \ , \qquad (5.3)$$

$$R\{v_{sy4}\} = \frac{-\beta_1^2(k+1)}{\beta_1^2(k+1) + 12\sigma^2/N} \ , \qquad (5.4)$$

$$R\{v_{sy5}\} = \frac{\beta_1^2(2k^2+1)}{\beta_1^2(k^2-1) + 12(k-1)\sigma^2/N} \ , \qquad (5.5)$$

$$R\{v_{sy6}\} = R\{v_{sy4}\} \ , \qquad (5.6)$$

$$R\{v_{sy7}\} = R\{v_{sy4}\} \ , \qquad (5.7)$$

and

$$R\{v_{sy8}\} \doteq \left\{\beta_1^2 k(n+1)\left[1 + \frac{2}{\ln\gamma(1)/\gamma(0)} + \frac{2}{\gamma(0)/\gamma(1)-1}\right] - \beta_1^{2(k+1)}\right\} \Big/ \left\{\beta_1^2(k+1) + 12\sigma^2/N\right\} \quad , \quad (5.8)$$

where

$$\gamma(1) = \beta_1^2 k^2 (n-3)(n+1)/12 - \sigma^2/n$$

$$\gamma(0) = \beta_1^2 k^2 n(n+1)/12 + \sigma^2 \quad .$$

The expression for $R\{v_{sy8}\}$ was derived by approximating the expectation, E, of the function v_{sy8} $(s^2, \hat{\rho}_k s^2)$ by the same function of the expectations $E\{s^2\}$ and $E\{\hat{\rho}_k s^2\}$, where we have used an expanded notation for v_{sy8}. In deriving this result it was also assumed that $\hat{\rho}_k > 0$ with probability one. This assumption is quite modest and guarantees that terms involving the operator ln (•) are well defined.

From (5.1),...,(5.8), it can be seen that the value of the intercept, β_0, has no effect on the relative biases, while the error variance has only slight impact since terms in σ^2 are of lower order than the remaining terms. Similarly, the value of β_1 has little effect on the relative bias, unless β_1 is extraordinarily small. Note that $R\{v_{sy1}\},\ldots,R\{v_{sy7}\}$ converge to zero as $\beta_1 \to 0$. Thus, v_{sy1} through v_{sy7} are unbiased when the population is random. As $\beta_1 \to 0$, the assumption that $\Pr\{\hat{\rho}_k > 0\} = 1$ will not hold, and the expression for $R\{v_{sy8}\}$ in (5.8) will not be valid. For large populations where β_1 is not extremely close to 0, (5.1), (5.4), (5.5), (5.6), and (5.7) suggest the following useful approximations:

$$R\{v_{sy1}\} \doteq n$$
$$R\{v_{sy4}\} \doteq -1$$
$$R\{v_{sy5}\} \doteq 2$$
$$R\{v_{sy6}\} \doteq -1$$
$$R\{v_{sy7}\} \doteq -1 \quad .$$

We have also derived expressions for the relative biases under the more general assumption that the u_{ij} are mutually independent with zero mean and heterogeneous variance σ_{ij}^2. The observations made in the previous paragraph also apply to this model.

The results for population A2 in Table 2 agree well with the expressions for the expected relative biases. For example, letting N=1000, n=10, k=100, σ=1.5, β_1=1, and β_0=0, we find that equations (5.1),...,(5.8) take the values 9.888, -0.406, -0.406, -1.000, 2.000, -1.000, -1.000, and -0.355, respectively.

As further confirmation of the expressions for the relative biases, 100 populations of size N=1000 were generated according to the superpopulation model for each of the following values of $(\beta_0, \beta_1, \sigma)$: (0,.5,1.5), (0,1,1.5), (0,2,1.5), and (0,1,5). The bias, MSE, and significance level (associated with confidence intervals which used the multiplier $t_{n-1,.025}$) of each estimator of variance was then found for both f=.01 and .02 for each population. To illustrate, the results for the case $(\beta_0, \beta_1, \sigma)$=(0,1,5) with f=.01 are summarized in Table 5. The second and third columns give the quantities

$$\left(\Sigma(E\{v_{sy\alpha}\} - V\{\bar{y}_{sy}\})/100\right) \Big/ \left(\Sigma\, V\{\bar{y}_{sy}\}/100\right)$$

and

$$\left(\Sigma\{E(v_{sy\alpha} - V\{\bar{y}_{sy}\})^2\}/100\right) \Big/ \left(\Sigma(V\{\bar{y}_{sy}\})^2/100\right)$$

respectively, where α=1,...,8 and the summations are taken over the 100 populations. The fourth column gives the average significance level for each estimator. Note that the bias results agree well with the expressions in (5.1),...,(5.8). Clearly, the only estimators with acceptable properties are v_{sy2}, v_{sy3}, and v_{sy8}: the remaining estimators have either large MSE or lead to unacceptably low confidence levels. And among these estimators, v_{sy8} has the smallest bias and MSE. The results for the other values of $(\beta_0, \beta_1, \sigma)$ are similar.

Table 5. Monte Carlo Estimates of Expected Bias, Expected MSE, and Expected Confidence Levels

Estimator	Expected Relative Bias	Expected Relative MSE	Expected Confidence Level
v_{sy1}	9.856	97.168	100.00
v_{sy2}	-0.405	0.164	99.11
v_{sy3}	-0.405	0.164	99.06
v_{sy4}	-0.997	0.994	6.62
v_{sy5}	1.993	4.007	100.00
v_{sy6}	-0.997	0.994	6.14
v_{sy7}	-0.997	0.994	5.54
v_{sy8}	-0.355	0.126	99.94

It would be hazardous at this point for the reader to draw very general conclusions about the eight estimators, since the investigation in this section assumed a very specific model which may not be obtained in practice. In the future, we will be investigating models with a higher order polynomial trend and other alternative specifications. Our continuing goal in this work will be to establish conditions under which the various estimators have acceptable properties.

References

[1] Cochran, W.G. "Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations" Ann. Math. Statist. 17 (1946): 164-177.

[2] Cochran, W.G. Sampling Techniques. New York: John Wiley and Sons, 1963.

[3] Koop, J.C. "On Splitting a Systematic Sample for Variance Estimation." Ann. Math. Statist. 42 (1971): 1084-1087.

[4] Matérn, B. "Methods of Estimating the Accuracy of Line and Sample Plot Surveys." Medd. fr. Statens. Skogsforskningsinstitut 36 (1947); 1-138

[5] Osborne, J.G. "Sampling Errors of Systematic and Random Surveys of Cover-Type Areas." J. Amer. Statist. Assoc. 37 (June 1942): 256-264.

[6] Yates, F. Sampling Methods for Censuses and Surveys . London: Griffin, 1949.