Journal of Official Statistics
Vol. 6. No. 2, 1990, pp. 115–131

Comparisons of Variance Estimators in Stratified Random and Systematic Sampling

Richard Valliant[1]

Abstract: Properties of stratified ratio and regression estimators and associated variance estimators can differ considerably under systematic as opposed to random sampling. Large sample properties under the sampling plans are compared theoretically and empirically. A simulation study of estimator performance is presented which examines the effect of different values of certain population parameters. The study contrasts properties under stratified random and systematic sampling and illustrates advantages of the separate regression estimator together with the jackknife variance estimator under a systematic plan. The separate and combined ratio estimators and the combined regression estimator, on the other hand, are less satisfactory. Each requires a more restrictive model in order to be conditionally unbiased than does the separate regression estimator. Even in populations in which the biases of the ratio estimators and the combined regression estimator can be controlled by systematic sampling, the variance estimators studied here can be substantial overestimates.

Key words: Jackknife variance estimator; separate ratio estimator; separate regression estimator; superpopulation model.

1. Introduction

Ratio and regression estimation in conjunction with stratification are familiar and well-studied methods in the survey sampling literature. Design-based variance estimators are summarized by Cochran (1977). Wu (1985) introduced a class of estimators, which included the standard ones, for the combined ratio estimator and obtained the member of the class optimal in terms of design mean squared error (MSE). In the unstratified case, design-based studies of the ratio estimator have been done by Rao and Rao (1971), Wu (1982), and Wu and Deng (1983). Deng and Wu (1987) also studied design-based properties of variance estimators for the unstratified regression estimator. Conditional model-based studies have been done by Royall and Cumberland (1981a, 1981b) and Royall and Eberhardt (1975) and have been extended to stratification by Valliant (1987a).

Most previous studies have been done in the context of simple random sampling (SRS) or stratified simple random sampling (STSRS) with relatively little attention given to stratified systematic sampling (STSYS) in ratio and regression estimation problems. Much of the literature on variance esti-

[1] Mathematical statistician, Office of Mathematical Statistics, Bureau of Labor Statistics, Room 2126, 441 G St. NW, Washington, D.C. 20212, U.S.A. The author thanks the associate editor and two referees whose comments led to substantial improvements in the presentation. Any opinions expressed are those of the author and do not constitute policy of the Bureau of Labor Statistics.

mation in systematic sampling deals only with the simple sample mean (e.g., Heilbron 1978; Wolter 1984). Iachan (1982) gives an extensive review of studies on systematic sampling and notes that there is a need for work on more complex estimators. This paper contrasts the effects of STSRS and STSYS on properties of variance estimators for ratio and regression estimators. Kott (1986) noted that systematic sampling is one method of protecting against certain kinds of model biases when estimating a mean. As illustrated here, systematic sampling can also have important effects on variance estimators.

The population is divided into H, a fixed number, of strata and within stratum h a sample of n_h units is selected from the total of N_h units. The sampling fraction in stratum h is $f_h = n_h/N_h$ and the set of sample units from stratum h is denoted as s_h. The total population size is $N = \Sigma_h N_h$ and the total sample size is $n = \Sigma_h n_h$. The proportion of the population in stratum h is $W_h = N_h/N$. Associated with unit (hi) is a random variable y_{hi} and an auxiliary x_{hi} with the latter known and positive for every unit in the population. Assume that there are bounds B_1 and B_2 such that $0 < B_1 \leqslant x_{hi} \leqslant B_2 < \infty$ for each h and i. As in Valliant (1987a,b), for model-based analyses we will consider a situation in which N_h, $n_h \to \infty$, $f_h \to 0$, and n_h/n and W_h converge to constants in all strata.

The finite population means of y and x are $\bar{y} = \Sigma_h^H \Sigma_i^{N_h} y_{hi}/N$ and $\bar{x} = \Sigma_h^H \Sigma_i^{N_h} x_{hi}/N$ and the stratum means are $\bar{y}_h = \Sigma_i^{N_h} y_{hi}/N_h$ and $\bar{x}_h = \Sigma_i^{N_h} x_{hi}/N_h$. The separate and combined ratio estimators are defined as

$$\bar{y}_{\mathrm{RS}} = \sum_h^H W_h \bar{y}_{hs} \bar{x}_h/\bar{x}_{hs}$$

and

$$\bar{y}_{\mathrm{RC}} = \bar{y}_s \bar{x}/\bar{x}_s,$$

where $\bar{y}_{hs} = \Sigma_{s_h} y_{hi}/n_h$, $\bar{x}_{hs} = \Sigma_{s_h} x_{hi}/n_h$, $\bar{y}_s$ is the stratified expansion estimator defined as $\bar{y}_s = \Sigma_h^H W_h \bar{y}_{hs}$, and $\bar{x}_s = \Sigma_h^H W_h \bar{x}_{hs}$. The separate and combined regression estimators are

$$\bar{y}_{\mathrm{LS}} = \sum_h^H W_h[\bar{y}_{hs} + b_{hs}(\bar{x}_h - \bar{x}_{hs})]$$

and

$$\bar{y}_{\mathrm{LC}} = \bar{y}_s + b(\bar{x} - \bar{x}_s),$$

where $b_{hs} = s_{xyhs}/s_{xxhs}$ and $b = \Sigma_h K_h s_{xyhs}/\Sigma_h K_h s_{xxhs}$ with $K_h = W_h^2(1 - f_h)/n_h$, $s_{xyhs} = \Sigma_{s_h}(x_{hi} - \bar{x}_{hs})y_{hi}/(n_h - 1)$, and $s_{xxhs} = \Sigma_{s_h}(x_{hi} - \bar{x}_{hs})^2/(n_h - 1)$.

We will study these estimators under some special cases of the model

$$y_{hi} = \alpha_h + \beta_h x_{hi} + \varepsilon_{hi},$$

$$E_\xi(\varepsilon_{hi}) = 0, \qquad (1)$$

and

$$\mathrm{var}_\xi(\varepsilon_{hi}) = v_{hi}$$

with the ε_{hi}'s uncorrelated. This model is often reasonable when strata are formed based on the size of x and a more complicated relationship between y and x may be approximated linearly within strata. Such populations are often encountered in surveys of business establishments or institutions such as hospitals conducted by national governments.

2. Properties of the Ratio and Regression Estimators

Theoretical properties of the ratio and regression estimators are sketched in this section. In order to make comparisons we employ both model and design-based calculations. Two results are useful in this

regard. First, under appropriate conditions, $\sqrt{n_h}(\bar{x}_{hs} - \bar{x}_h)$ converges in distribution to a normal random variable under simple random sampling without replacement as $n_h \to \infty$ (Scott and Wu 1981), i.e., $(\bar{x}_{hs} - \bar{x}_h) = O_d(n_h^{-1/2})$ where O_d denotes probabilistic order with respect to the sample design. The Lindeberg–Hájek condition under which this order result holds is somewhat technical to state here, but, to paraphrase Scott and Wu, the condition essentially requires that the contribution of gross outliers to the stratum total sum of squares $\Sigma_{i=1}^{N_h}(x_{hi} - \bar{x}_h)^2$ be relatively small. The second result is due to Kott (1986) and states that when a systematic sample is selected from a list ordered by x and x is bounded as in Section 1, then $(\bar{x}_{hs} - \bar{x}_h) = O(n_h^{-1})$ with the order being nonprobabilistic. Assuming that n_h/n converges to a constant in each stratum, we have $\bar{x}_h/\bar{x}_{hs} = 1 + O_d(n^{-1/2})$ under STSRS but $\bar{x}_h/\bar{x}_{hs} = 1 + O(n^{-1})$ under STSYS. It follows that under STSRS $\bar{y}_{\mathrm{RS}} = \bar{y}_s + O_d(n^{-1/2})$ while $\bar{y}_{\mathrm{RS}} = \bar{y}_s + O(n^{-1})$ under STSYS. These same relationships to the stratified expansion estimator $\bar{y}_s$ also hold for $\bar{y}_{\mathrm{RC}}$, $\bar{y}_{\mathrm{LS}}$, and $\bar{y}_{\mathrm{LC}}$. Thus, the differences among the four estimators are of small consequence in large systematic samples because that sampling plan is an effective way of achieving sample balance on x.

Turning to the model bias and variance of $\bar{y}_{\mathrm{RS}}$ under (1), Valliant (1987a) noted that

$$E_\xi(\bar{y}_{\mathrm{RS}} - \bar{y}) = \sum_h W_h\alpha_h(\bar{x}_h - \bar{x}_{hs})/\bar{x}_{hs} \tag{2}$$

and

$$\mathrm{var}_\xi(\bar{y}_{\mathrm{RS}} - \bar{y}) \approx \sum_h W_h^2 D_{xh}^2 \frac{\bar{v}_{hs}}{n_h}, \tag{3}$$

where $D_{xh} = \bar{x}_h/\bar{x}_{hs}$, $\bar{v}_{hs} = \Sigma_{s_h} v_{hi}/n_h$, and $\approx$ denotes "asymptotically equivalent." The model variance has order n^{-1}, assuming $\bar{v}_{hs}$ and n_h/n converge to constants as $n_h \to \infty$. The model bias (2) is a random variable with respect to the sample design. Since, under STSRS $(\bar{x}_{hs} - \bar{x}_h) = O_d(n^{-1/2})$, the square of the bias (2) has order n^{-1} under STSRS which is the same order as the model variance (3). On the other hand, under STSYS the square of the bias is order n^{-2}. The results of Kott (1986) on systematic sampling also can be applied more generally when, for example, $E_\xi(y_{hi})$ is a polynomial in x_{hi}. In summary, when an STSYS is selected, the dominant term of the model mean squared error is (3) with the square of the model bias being asymptotically much less important than under STSRS.

Similar arguments lead to the same conclusions for the combined ratio and combined regression estimators. Defining $D_x = \bar{x}/\bar{x}_s$, the model bias and approximate model variance of $\bar{y}_{\mathrm{RC}}$ are

$$E_\xi(\bar{y}_{\mathrm{RC}} - \bar{y}) = (D_x - 1)\sum_h W_h\alpha_h + \sum_h W_h\beta_h(D_x\bar{x}_{hs} - \bar{x}_h), \tag{4}$$

and

$$\mathrm{var}_\xi(\bar{y}_{\mathrm{RC}} - \bar{y}) \approx D_x^2 \sum_h W_h^2 \bar{v}_{hs}/n_h. \tag{5}$$

For the combined regression estimator the model bias and approximate variance are

$$E_\xi(\bar{y}_{\mathrm{LC}} - \bar{y}) = \sum_h W_h\beta_h(\bar{x}_{hs} - \bar{x}_h) + \frac{(\bar{x} - \bar{x}_s)}{S_{xx}}\sum_h \beta_h K_h s_{xxhs}$$

and

$$\mathrm{var}_\xi(\bar{y}_{\mathrm{LC}} - \bar{y}) \approx \sum_h W_h^2 \frac{\bar{v}_{hs}}{n_h} + \frac{(\bar{x} - \bar{x}_s)}{S_{xx}}\sum_h W_h \frac{K_h}{n_h} s_{1hs},$$

where $S_{xx} = \Sigma_h K_h s_{xxhs}$ and $s_{1hs} = \Sigma_{s_h}(x_{hi} - \bar{x}_{hs})v_{hi}/(n_h - 1)$. When stratum samples are large, $D_x = 1 + O_d(n^{-1/2})$ under STSRS and $1 + O(n^{-1})$ under STSYS. These facts and the aforementioned properties of $(\bar{x}_{hs} - \bar{x}_h)$ imply that the squares of the model biases of $\bar{y}_{RC}$ and $\bar{y}_{LC}$ both have order n^{-1} under STSRS but n^{-2} under STSYS. As was the case for $\bar{y}_{RS}$, the parts of the MSE's accounted for by the squares of the biases of $\bar{y}_{RC}$ and $\bar{y}_{LC}$ are far less important under STSYS than under STSRS.

The separate regression estimator is model unbiased under (1), as is well known, and has approximate model variance

$$\mathrm{var}_\xi(\bar{y}_{LS} - \bar{y}) \approx \sum_h W_h^2 \frac{\bar{v}_{hs}}{n_h} + 2\sum_h \frac{W_h^2}{n_h}(\bar{x}_h - \bar{x}_{hs})\frac{s_{1hs}}{s_{xxhs}}. \quad (6)$$

The first term on the right-hand side of (6) has order n^{-1}. The second term in (6) has order $n^{-3/2}$ under STSRS and order n^{-2} under STSYS. Thus, little difference between the STSRS and STSYS variances of $\bar{y}_{LS}$ is expected in large samples.

3. Variance Estimators

The fact that estimating repeated sampling variances from systematic samples may present special problems not encountered with random samples has long been recognized (e.g., Cochran 1946; Osborne 1942; Wolter 1984). These special problems are often not accounted for in practice. Wolter (1985, ch. 7) notes that common practice in applied survey work is to regard a systematic sample as random and estimate design variances using random sampling formulae. In a population with linear trend, computed variances are often considered to be overestimates because the random sampling formulae do not appropriately reflect the effect of the trend which is picked up by systematic selection (see e.g., Hansen, Hurwitz, and Madow 1953, §11.8; Wolter 1984).

A variety of variance estimators have been studied for $\bar{y}_{RC}$ and $\bar{y}_{RS}$. This paper examines a number of the choices that have been proposed for use under STSRS plans with emphasis on contrasting the properties that obtain under stratified simple random and stratified systematic plans. For $\bar{y}_{RS}$ we include

$$v_{RSg} = \sum_h \frac{K_h}{n_h - 1} D_{xh}^g \sum_{s_h} r_{1hi}^2$$

and

$$v_{RSJ} = \sum_h K_h \frac{n_h - 1}{n_h^2} D_{xh}^2 \times \sum_{s_h}\left[\frac{r_{1hi}}{1 - k_{1hi}} - \frac{1}{n_h}\sum_{s_h}\frac{r_{1hj}}{1 - k_{1hj}}\right]^2,$$

where $r_{1hi} = y_{hi} - x_{hi}\bar{y}_{hs}/\bar{x}_{hs}$ and $k_{1hi} = x_{hi}/(n_h\bar{x}_{hs})$. For the combined ratio estimator we consider

$$v_{RCg} = D_x^g \sum_h^H \frac{K_h}{n_h - 1}\sum_{s_h} r_{2hi}^2$$

and

$$v_{RCJ} = D_x^2 \sum_h^H \frac{K_h}{n_h - 1} \times \sum_{s_h}\left[\frac{r_{2hi}}{1 - k_{2hi}} - \frac{1}{n_h}\sum_{s_h}\frac{r_{2hj}}{1 - k_{2hj}}\right]^2,$$

where $r_{2hi} = (y_{hi} - \bar{y}_{hs}) - (\bar{y}_s/\bar{x}_s)(x_{hi} - \bar{x}_{hs})$, $k_{2hi} = W_h(x_{hi} - \bar{x}_{hs})/\{(n_h - 1)\bar{x}_s\}$.

The estimators v_{RSg} and v_{RCg} define classes studied by Wu (1985) who found values of g that were optimal in the sense of minimizing the approximate design MSE's of the variance estimators. For the separate estimators

we treat the case of the same value of g in all strata although Wu proposed that g be allowed to vary among strata. Cases of special interest are $g = 0, 1, 2$ which have been studied by a number of authors. The estimators v_{RSJ} and v_{RCJ} are computational forms for the stratified delete-one jackknife estimator whose general form was defined by Jones (1974). For some estimator $\hat{\theta}$ the general form is $v_J = \Sigma_h(1 - f_h)\{(n_h - 1)/n_h\}$ $\Sigma_{s_h}\{\hat{\theta}_{(hi)} - \hat{\theta}_{(h)}\}^2$ where $\hat{\theta}_{(hi)}$ has the same form as $\hat{\theta}$ but omits the $(hi)^{th}$ sample unit and $\hat{\theta}_{(h)} = \Sigma\hat{\theta}_{(hi)}/n_h$. Since all x_{hi} are bounded, k_{1hi} and k_{2hi} are both $o(1)$ and it is clear from the computational forms above that v_{RSJ} is asymptotically equivalent to v_{RS2}, and v_{RCJ} is asymptotically equivalent to v_{RC2}. Wu (1985) earlier showed that under STSRS v_{RC2} is the closest approximation to v_{RCJ} within the class v_{RCg}. Royall and Cumberland (1978, § 6) also showed that the general jackknife v_J is asymptotically equivalent to a variance estimator, G_1 in their notation, which was derived to be robust against failure of the variance specification in a linear model.

Variance estimators we consider for the separate regression estimator are in the class

$$v_{LSg} = \sum_h \frac{K_h}{n_h - 2} D_{xh}^g \sum_{s_h} d_{1hi}^2$$

where $d_{1hi} = (y_{hi} - \bar{y}_{hs}) - b_{hs}(x_{hi} - \bar{x}_{hs})$. For the combined regression estimator consider

$$v_{LCg} = D_x^g \sum_h^H \frac{K_h}{n_h - 1} \sum_{s_h} d_{2hi}^2$$

where $d_{2hi} = (y_{hi} - \bar{y}_{hs}) - b(x_{hi} - \bar{x}_{hs})$. The classes defined by v_{LSg} and v_{LCg} were studied by Deng and Wu (1987) for the unstratified case and by Wu (1985). In the empirical study we additionally include the jackknife variance estimators for $\bar{y}_{LS}$ and $\bar{y}_{LC}$, computational forms of which are included in Valliant (1987a).

In the case of the sample mean, Wolter (1984) has studied a number of estimators involving contrasts and other functions of the sample y's which are designed to address the peculiarities produced by systematic samples. The focus here will not be to develop new variance estimators but to study the consequences of the common practice of using random sampling estimators when the sample is actually systematic.

4. Properties of Variance Estimators

First, consider variance estimators for the separate ratio estimator. Since, for a fixed value of g, $D_{xh}^g = 1 + O_d(n^{-1/2})$ under STSRS, we have $v_{RSg} = v_{RS0} + O_d(n^{-3/2})$ under that plan. However, under systematic sampling $D_{xh}^g = 1 + O(n^{-1})$ and $v_{RSg} = v_{RS0} + O(n^{-2})$. Thus, the choice of g is of less consequence when an STSYS plan is used. Under model (1)

$$E_\xi(v_{RSg}) \approx \sum_h W_h^2 \frac{D_{xh}^g}{n_h}\left[\bar{v}_{hs} + \alpha_h^2 \frac{s_{xxhs}}{\bar{x}_{hs}^2}\right]. \tag{7}$$

Recalling (3), v_{RS2} is approximately model unbiased when $\alpha_h = 0$ while other choices of g lead to a bias. When $\alpha_h \neq 0$, all v_{RSg} are biased estimators of the model MSE. The bias may be substantial and positive under STSYS because systematic sampling from a list sorted by x prevents small values of s_{xxhs} but reduces the importance of the bias (2). This observation is similar to the findings of Royall and Cumberland (1978, § 5.2) on the overestimation by certain variance estimators for the unstratified ($H = 1$) ratio estimator in balanced samples ($\bar{x}_{hs} = \bar{x}_h$). On the other hand, if y is extremely variable for a given x so that $\bar{v}_{hs} \gg \alpha_h^2 s_{xxhs}/\bar{x}_{hs}^2$, then the model bias of v_{RSg} can be negligible under STSYS.

Turning to the jackknife, we noted in

Section 3 that v_{RSJ} was approximately equal to v_{RS2} in large samples so that (7) with $g = 2$ also applies to the jackknife for the separate ratio estimator. Thus, v_{RSJ} is approximately unbiased when $\alpha_h = 0$. When $\alpha_h \neq 0$, the jackknife, like v_{RS2}, is likely to be a considerable overestimate under STSYS in populations where the variance of y given x is small. Generally, v_{RSJ} will be larger than v_{RS2} because of the denominator factors, $1 - k_{1hi}$, which are less than one.

Similar theory can be worked out for v_{RCg}. An approximation to $E_\xi(v_{RCg})$ is given by the right-hand side of (7) with $\bar{x}_{hs}$ replaced by $\bar{x}_s$. Consequently, the same remarks given above on the model bias of v_{RSg} under STSYS also apply to v_{RCg} and to v_{RCJ} because of its large sample equivalence to v_{RC2}.

Next, consider the regression estimators. Using the approximation $D_{xh}^g \approx 1 - g(\bar{x}_{hs} - \bar{x}_h)/\bar{x}_h$ and results from Valliant (1987a, §3.3), the approximate model bias of v_{LSg} is

$$\text{bias}_\xi(v_{LSg}) \approx \sum_h \frac{W_h^2}{n_h}(\bar{x}_{hs} - \bar{x}_h) \times \left[-g\frac{\bar{v}_{hs}}{\bar{x}_{hs}} + 2\frac{s_{1hs}}{s_{xxhs}} \right]$$

which has order $n^{-3/2}$ under STSRS but only n^{-2} under STSYS. In either case, the bias of v_{LSg} has a lower order than the variance of $\bar{y}_{LS}$ which is $O(n^{-1})$. Similar findings apply to v_{LCg} if $\beta_h = \beta$ in all strata. However, if the slope parameter is not the same in all strata, v_{LCg} has a model bias of order n^{-1} as do v_{RSg} and v_{RCg}. Using the computational forms of v_{LSJ} and v_{LCJ} in Valliant (1987a, §4), it can be shown that, in large samples, v_{LSJ} is approximately equal to v_{LS0} and that v_{LCJ} is approximately equal to v_{LC0}. Thus, the above comments on v_{LS0} and v_{LC0} also apply to the corresponding jackknife estimators.

5. Simulation Results

The earlier theory was tested in a simulation study using six artificial populations. Use of generated rather than real populations has some advantages in allowing certain population parameters to be systematically varied in order to study their effect on estimator performance. In particular, we controlled (1) curvature of the regression of y on x and (2) the conditional variance of y given x. In each of the six populations 2000 (x, y) pairs were generated. Each x was generated as $x = 150 + 600w$ where w was a standardized chi square random variable with six degrees of freedom (df), i.e. $w = (\chi_6^2 - 6)/\sqrt{12}$. Given x, y was generated as

$$y = a + bx + cx^2 + dx^g z$$

where a, b, c, and d were constants and z was a standardized chi square random variable with six df. Values of x were constrained to be in the interval [1, 1500] while y was restricted to [50, 2500]. Table 1 lists the parameter values used for each population and Figure 1 shows scatterplots of systematic samples of 200 units from each population. Populations 1 and 2 both have the same specification for $E_\xi(y)$; population 1 has the

Table 1. Parameters used in generating study populations

Population	b	c	g
1	1.5	0	0.75
2	1.5	0	1.00
3	1.8	−0.0008	0.75
4	1.8	−0.0008	1.00
5	−0.3	0.0009	0.75
6	−0.3	0.0009	1.00

Note: In all six populations $a = 100$ and $d = 0.5$.

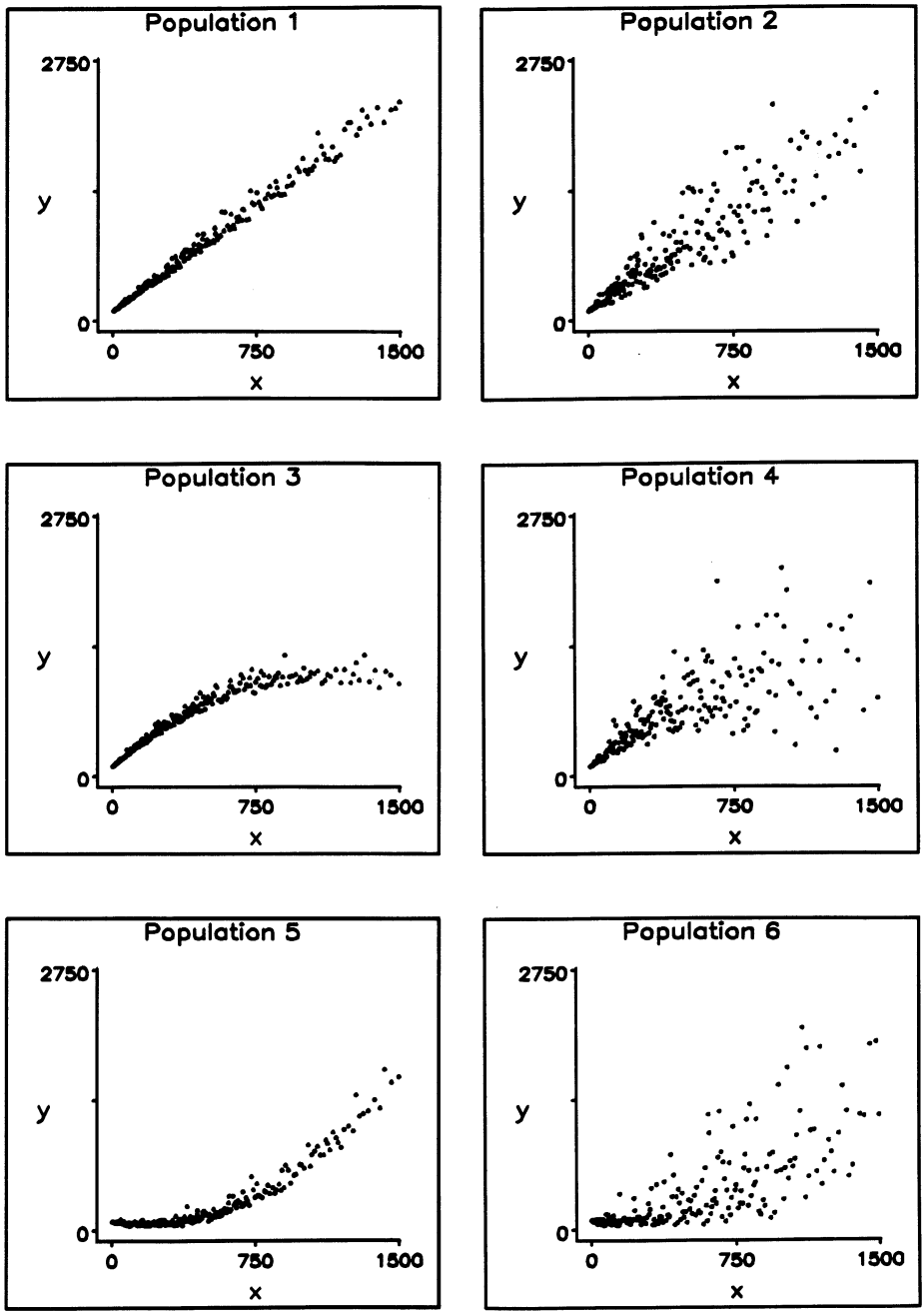

Fig. 1. Scatterplots of 200 units from each of the six simulation study populations.

variance of y proportional to $x^{3/2}$ while population 2 has $\text{var}_\xi(y) \propto x^2$. The remaining populations are similarly paired.

Each population was divided into five strata, after sorting units in ascending order on x, with $N_h = 400$ ($h = 1, \ldots, 5$). The main goal in strata formation here was to create enough strata so that the piecewise

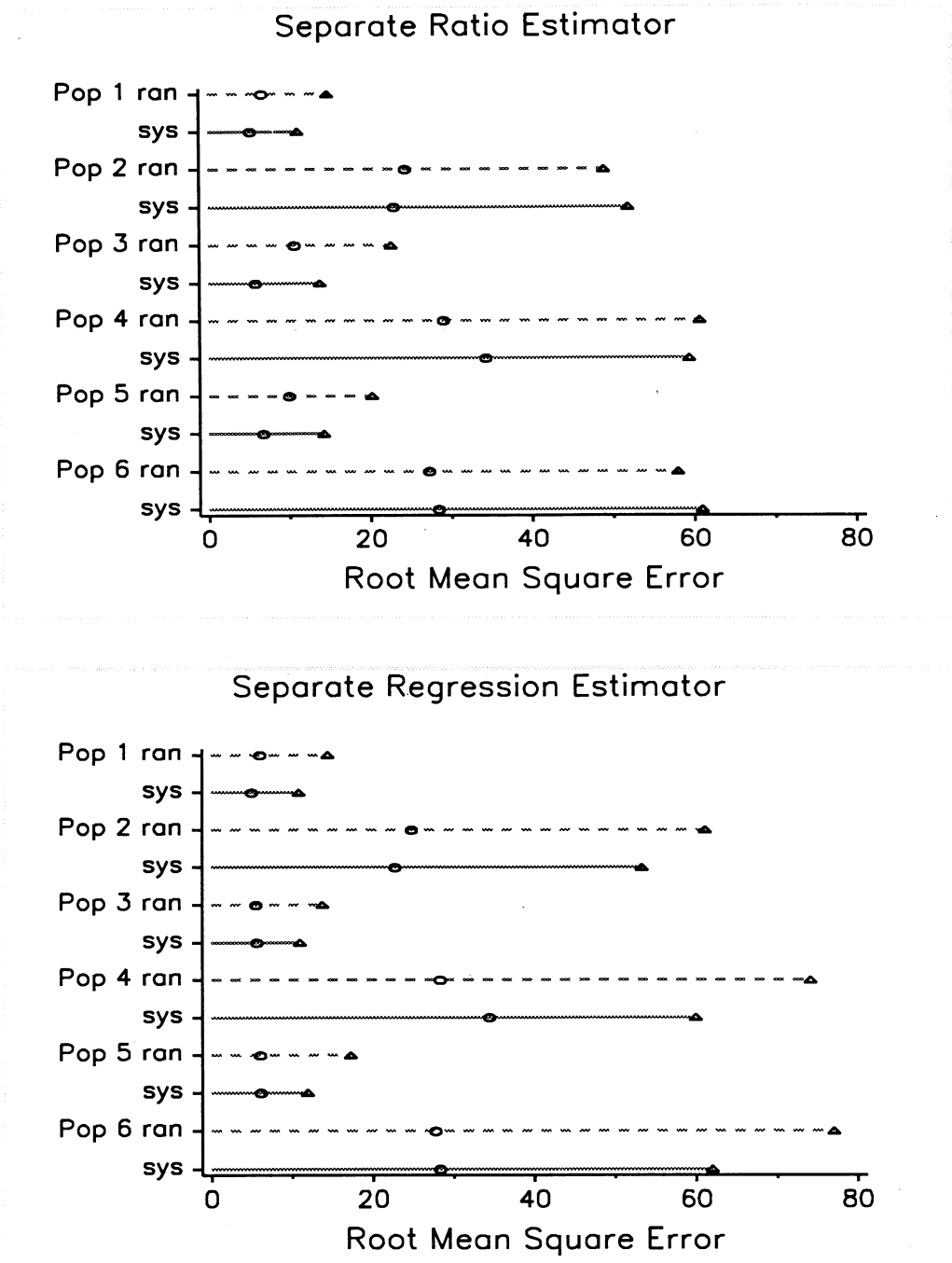

Fig. 2. Square roots of empirical mean squared errors for the separate ratio and regression estimators for 1000 samples from each of six populations. Separate lines are shown for stratified simple random samples (RAN) and stratified systematic samples (SYS). Triangles denote samples of n = 25. Ovals denote samples of n = 100.

linear model in (1) was reasonable even though $E_\xi(y)$ was generated as a quadratic function of x. Creation of strata with equal numbers of units is one choice considered by Royall and Herson (1973, Theorem 2) for achieving optimality under a particular model when stratified balanced sampling is used and the same number of sample units is allocated to each stratum. Other methods of stratification, based on design variance considerations, such as the cumulative square root rule (Cochran 1977, p. 127), are also often used. The preceding theory relies on model (1) being a useful approximation, a condition that would hold for many stratification algorithms in the populations studied here.

From each population four sets of 1000 samples were selected: (1) 1000 stratified simple random samples of size $n = 25$ ($n_h = 5$ for all h), (2) 1000 STSRS's of $n = 100$ ($n_h = 20$), (3) 1000 STSYS's of $n = 25$ ($n_h = 5$), and (4) 1000 STSYS's of $n = 100$ ($n_h = 20$). All simple random samples were selected without replacement and all systematic samples were selected with separate random starts in each stratum after sorting units in ascending order on x.

Figure 2 is a plot of root mean squared errors (RMSE's) for the separate ratio and regression estimators over the sets of 1000 samples. The RMSE of the separate ratio estimator, for example, is defined as $[\sum_{i=1}^{S}(\bar{y}_{\mathrm{RS}i} - \bar{y})^2/S]^{1/2}$ where $\bar{y}_{\mathrm{RS}i}$ is the estimate from sample i and $S = 1000$. For each population separate lines are given in Figure 2 for stratified random samples and for stratified systematic samples. Triangles represent the RMSE's for samples of 25 and ovals give the RMSE's for $n = 100$. Results for the combined estimators are omitted to conserve space. We emphasize unconditional comparisons, i.e., ones over all 1000 samples, because conditional properties under STSRS have been examined elsewhere (Valliant 1987a) and because systematic sampling virtually eliminates conditional differences in the estimators studied here.

Based on the theory for model (1), the conditional (model) bias of $\bar{y}_{\mathrm{RS}}$ can make a substantial contribution to the RMSE over all samples when an STSRS plan is used. Stratified systematic sampling should remove the model bias component of the RMSE. This effect is clearly manifested in Figure 2 in the lower variance populations (populations 1, 3, 5) where the separate ratio estimator has a considerably lower RMSE at either sample size under systematic sampling than under random sampling. For populations 1, 3, and 5, for example, the exact ratios, (RMSE under STSRS) ÷ (RMSE under STSYS), for $\bar{y}_{\mathrm{RS}}$ when $n = 100$ are 1.27, 1.83, and 1.48. In the higher variance populations (2, 4, 6), on the other hand, differences in the RMSE's of $\bar{y}_{\mathrm{RS}}$ are smaller under the two sampling plans. For populations 2, 4, and 6 the ratios, (RMSE under STSRS) ÷ (RMSE under STSYS), for $\bar{y}_{\mathrm{RS}}$ when $n = 100$ are 1.06, 0.85, and 0.96.

The separate regression estimator is unbiased under model (1) and, based on approximation (6), little difference is anticipated theoretically between STSRS and STSYS in large samples. When $n = 100$ in Figure 2, the RMSE's of $\bar{y}_{\mathrm{LS}}$ conform to theory, having similar values under random and systematic sampling with the exception of population 4 where STSRS is actually more precise. However, when $n = 25$, the separate regression estimator is more precise for all populations under STSYS than under STSRS, indicating some small sample differences not apparent in (6). There are noticeable differences between the RMSE's of $\bar{y}_{\mathrm{RS}}$ and $\bar{y}_{\mathrm{LS}}$ under random sampling, particularly for $n = 25$ in the higher variance

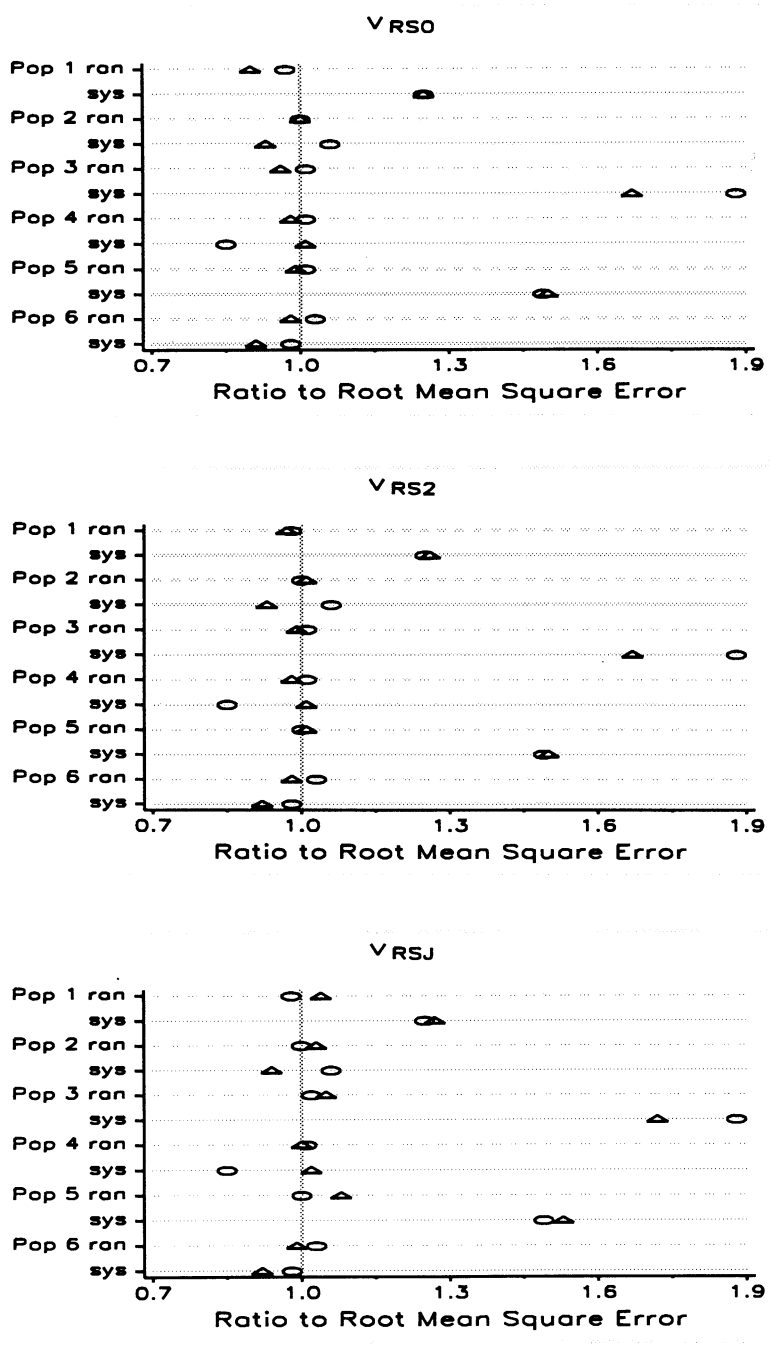

Fig. 3. Ratios of square roots of average values of variance estimators to empirical root mean squared errors, $\bar{v}^{I\,2}_{RSj}/RMSE(\bar{y}_{RS})$ $(j = 0, 2, J)$, for the separate ratio estimator for sets of 1000 samples from each of six populations. Separate lines are shown for stratified simple random samples (RAN) and stratified systematic samples (SYS). Triangles denote samples of $n = 25$. Ovals denote samples of $n = 100$.

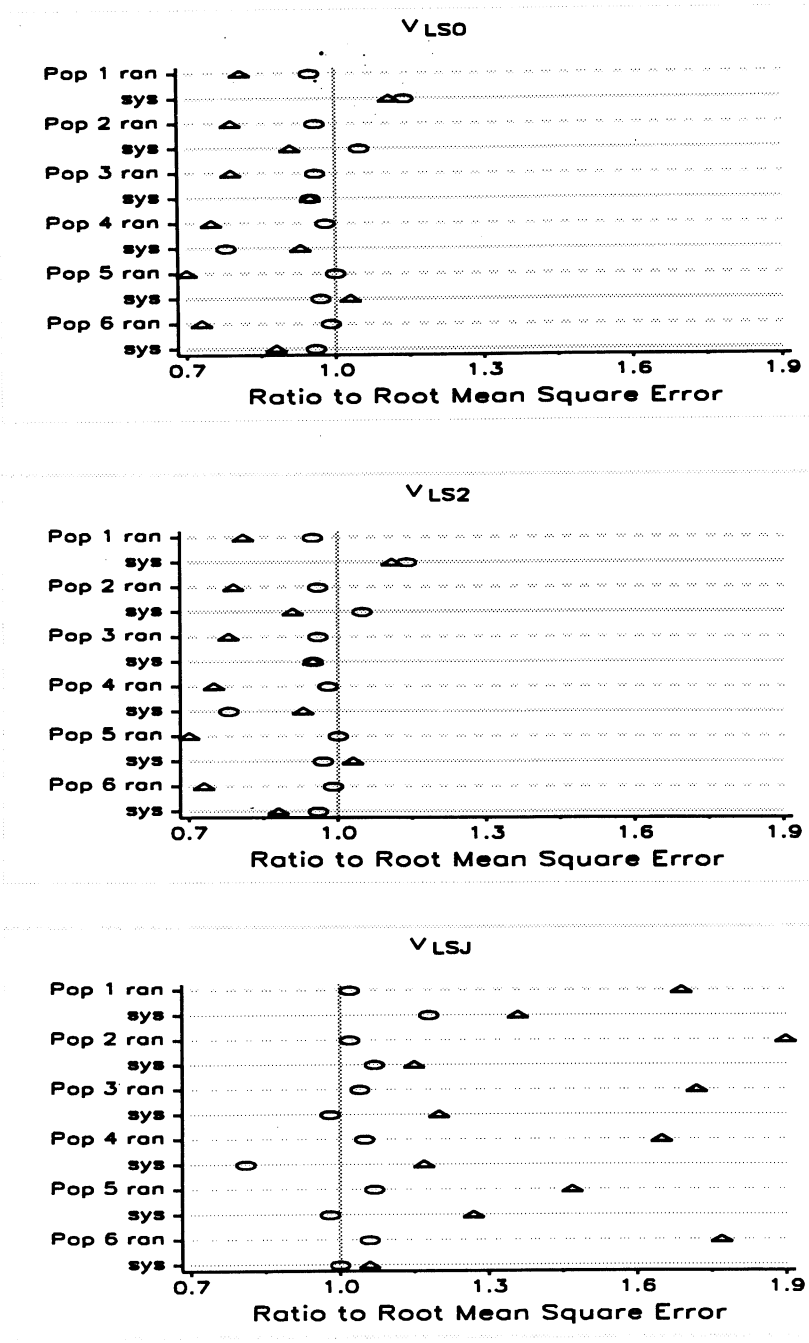

Fig. 4. Ratios of square roots of average values of variance estimators to empirical root mean squared errors, $\bar{v}_{LSj}^{1/2}/RMSE(\bar{y}_{LS})$ *(*$j = 0, 2, J$*), for the separate regression estimator for sets of 1000 samples from each of six populations. Separate lines are shown for stratified simple random samples (RAN) and stratified systematic samples (SYS). Triangles denote samples of* $n = 25$*. Ovals denote samples of* $n = 100$*.*

populations where $\bar{y}_{RS}$ is more precise. However, in the systematic samples the RMSE's of the separate ratio and regression estimates are similar, especially at the larger sample size. This is in accord with the theoretical observation in Section 2 that $\bar{y}_{RS}$ and $\bar{y}_{LS}$ differ from each other only by a term of order n^{-1} under STSYS.

Figures 3 and 4 plot the ratio of the square roots of average variance estimates to the RMSE's for the separate ratio and regression estimates. The ratios plotted in Figure 3, for example, are $\bar{v}_{RSj}^{1/2}/\text{RMSE}(\bar{y}_{RS})$, $(j = 0, 2, J)$, where $\bar{v}_{RSj} = \Sigma_{i=1}^{S} v_{RSji}/S$ and v_{RSji} is the value of v_{RSj} in sample i. As in Figure 2, separate lines are given in the figures for each population and selection method with triangles and ovals denoting results for the two sample sizes. Ratios of 1 indicate unbiasedness; ratios less than 1 correspond to underestimation; and ratios greater than 1 are cases of overestimation. Figure 3 includes ratios for v_{RS0}, v_{RS2}, and v_{RSJ} while Figure 4 includes the corresponding variance estimators for $\bar{y}_{LS}$. The choice $v_{RS1}(v_{LS1})$ had ratios intermediate between v_{RS0} and $v_{RS2}(v_{LS0}$ and $v_{LS2})$ and is omitted.

First, consider the STSRS results. The usual design-based theory predicts that in large STSRS's, all choices in Figures 3 and 4 should be approximately unbiased over all samples. In these STSRS simulations each of v_{RSg} $(g = 0, 2)$ are generally moderate to small underestimates at either sample size in Figure 3. The jackknife v_{RSJ} is somewhat of an overestimate in STSRS. For the regression estimator $\bar{y}_{LS}$ the asymptotic properties of variance estimators in random sampling do not appear to apply as quickly as for the variance estimators for $\bar{y}_{RS}$. Each v_{LSg} $(g = 0,2)$ is a severe underestimate in Figure 4 for (STSRS, $n = 25$) with the problem being less severe but still present at $n = 100$. (For STSRS, $n = 25$) the jackknife v_{LSJ} has especially wild behavior, overestimating in all populations with some of the worst cases being the high variance populations (2, 4, 6). For (STSRS, $n = 100$) the jackknife for $\bar{y}_{LS}$ is the best performer, being a slight overestimate in all populations while the other choices tend to be underestimates.

With systematic sampling the picture changes. Recalling (7), v_{RS0}, v_{RS2}, and v_{RSJ} are expected to be overestimates in the low variance populations under STSYS. This is clearly illustrated in Figure 3 for populations 1, 3, and 5 where each variance estimator is an overestimate in STSYS at both sample sizes. Results for populations 3 and 5, where the ratios for all variance estimates are about 1.9 and 1.5 when $n = 100$, are especially striking. On the other hand, in the high variance populations (2, 4, 6) the pattern of consistent overestimation of the RMSE of $\bar{y}_{RS}$ does not hold. For population 4 with (STSYS, $n = 100$) the RMSE is underestimated by about 15%. In Figure 4, the performance of the variance estimators for the separate regression estimator is substantially better under STSYS than STSRS. Except in population 4, the degree of underestimation by v_{LS0} and v_{LS2} is reduced or eliminated at $n = 25$ and at $n = 100$ is relatively minor where present. For (STSYS, $n = 100$) the best performer in Figure 4 in terms of bias is v_{LSJ} by a slight margin.

Table 2 gives empirical standard deviations (s.d.'s) of the variance estimates. In either random or systematic sampling there are differences in precision among the v_{RSg} and among the v_{LSg} but the differences are of no great consequence. The most dramatic numbers in Table 2 are for the jackknife for separate regression estimator which has enormous s.d.'s for (STSRS, $n = 25$) a finding similar to that of Andersson, Forsman, and Wretman (1987) in the context of price

Table 2. Standard deviations of variance estimates for the separate ratio and separate linear regression estimators in sets of 1000 stratified simple random and systematic samples from six populations

Population	Sample Type	n	Standard deviations in 1000 samples							
			v_{RS0}	v_{RS1}	v_{RS2}	v_{RSJ}	v_{LS0}	v_{LS1}	v_{LS2}	v_{LSJ}
1	ran	25	82.7	89.8	122	231	85.2	85.9	88.0	1387
		100	9.2	9.4	10.0	10.1	8.8	8.8	8.8	12.9
	sys	25	85.2	86.0	87.4	88.7	89.8	89.9	90.4	161
		100	9.0	9.1	9.1	9.1	8.9	9.0	9.0	9.7
2	ran	25	1028	1033	1060	1169	1165	1173	1202	41971
		100	114	114	115	115	109	109	111	152
	sys	25	1033	1026	1023	1034	1182	1182	1185	2081
		100	126	127	128	128	119	120	121	122
3	ran	25	224	223	240	345	70.9	67.7	66.1	1675
		100	24.8	24.5	24.9	25.5	7.3	7.1	7.1	10.8
	sys	25	173	169	167	176	53.3	53.3	53.5	115
		100	17.7	17.7	17.8	18.0	7.4	7.4	7.5	8.2
4	ran	25	1725	1708	1725	1789	1691	1675	1691	31915
		100	181	178	178	179	151	149	149	241
	sys	25	1799	1783	1775	1831	1837	1833	1836	3746
		100	174	174	173	174	156	155	155	161
5	ran	25	193	196	223	621	93.6	92.9	94.0	1330
		100	22.7	22.4	22.4	22.8	11.5	11.5	11.5	17.6
	sys	25	186	178	172	180	108	105	101	137
		100	21.2	21.0	20.7	20.9	9.4	9.4	9.4	9.4
6	ran	25	1548	1543	1562	1627	1748	1743	1761	66736
		100	163	162	163	163	163	162	163	256
	sys	25	1503	1506	1514	1546	1595	1603	1615	2592
		100	157	157	157	157	157	156	156	160

index estimation. The potential for high variability of the jackknife was also noted by Wu (1986) in linear model analysis. The extreme variability of the jackknife is reduced by using systematic sampling, particularly for $n = 100$.

Figure 5 gives empirical coverage probabilities of normal approximation confidence intervals which have nominal coverage rates of 95%. The upper half of the figure includes results using $\bar{y}_{RS}$ together with v_{RS0} and v_{RSJ}. The lower half shows results for $\bar{y}_{LS}$ together with v_{LS0} and v_{LSJ}. The choices v_{RS1} and v_{RS2} (v_{LS1} and v_{LS2}) gave coverage percentages similar to v_{RS0} (v_{LS0}) and are not shown. First, examine the results for the separate ratio estimator. For populations 1, 3, and 5 the contrasts between random and systematic sampling when using $\bar{y}_{RS}$ are evident. Under STSRS coverage probabilities in Figure 5 using $\bar{y}_{RS}$ and its variance estimates are consistently too low when $n = 25$ but are nearer the nominal values when $n = 100$, though still slightly low. The overestimation under STSYS by the variance estimates for $\bar{y}_{RS}$ in

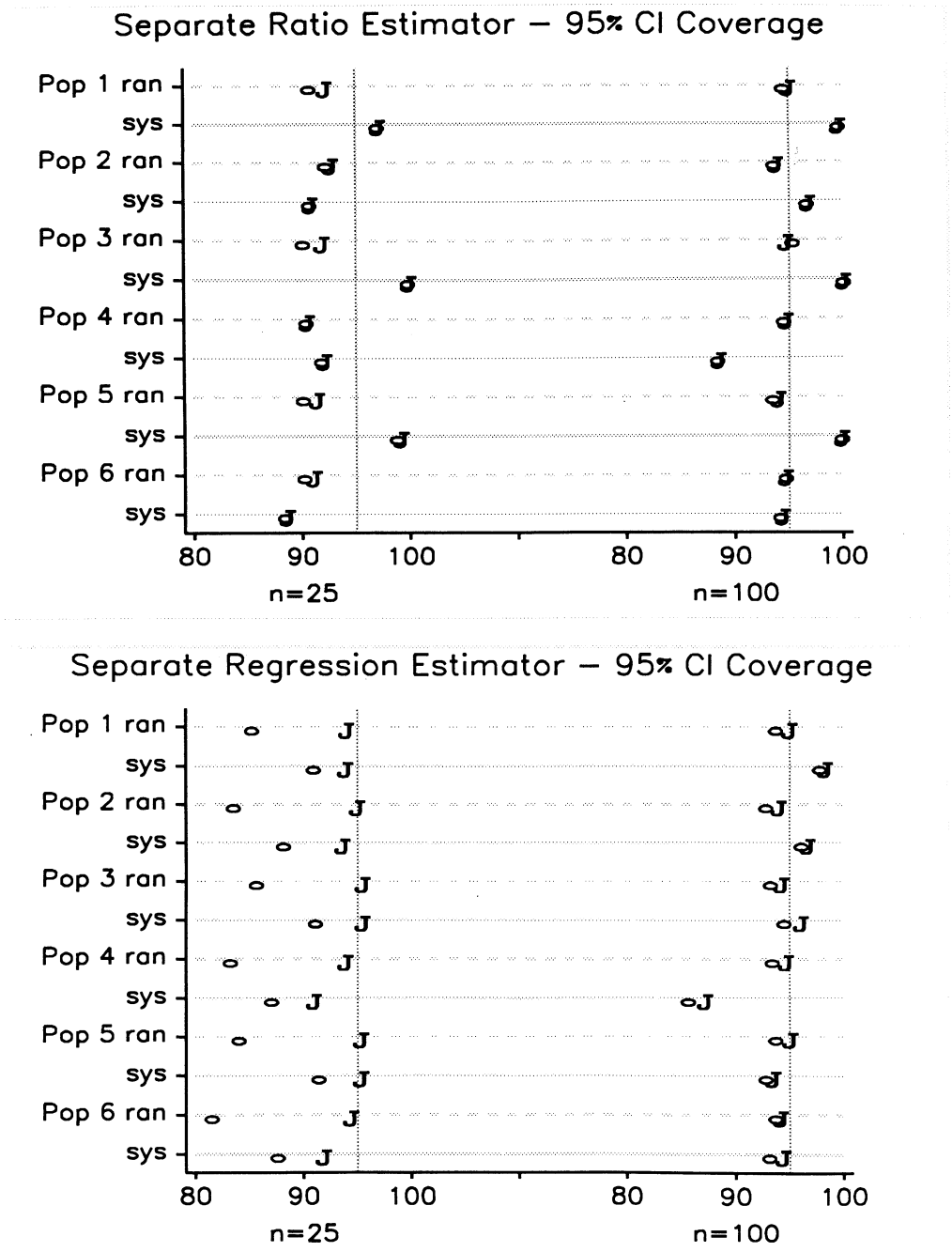

Fig. 5. Empirical coverage probabilities in sets of 1000 samples of nominal 95% confidence intervals based on the separate ratio estimator and two estimators of its variance and on the separate regression estimator and two estimators of its variance. Separate lines are shown for stratified simple random samples (RAN) and stratified systematic samples (SYS). Here 0 denotes v_{RS0} or v_{LS0} and J denotes v_{RSJ} or v_{LSJ}. Vertical reference lines are drawn at 95% for samples of $n = 25$ and $n = 100$.

Figure 2 leads to wasteful overcoverage in Figure 5 in populations 1, 3, and 5 at both $n = 25$ and 100. On the other hand, in the high variance populations 2, 4, and 6 coverage probabilities under STSYS are most often somewhat less than the nominal levels.

The separate regression estimator together with its variance estimators generally performs less erratically across the different populations for confidence interval construction under STSYS than does the separate ratio estimator. The extreme overcoverage under STSYS noted for $\bar{y}_{RS}$ in populations 1, 3, and 5 is not present for $\bar{y}_{LS}$ in Figure 5. With (STSYS, $n = 100$) the empirical coverages are reasonably close to the nominal values except in population 4 where the undercoverage problem persists. When $n = 100$ and the variance estimates are more stable, v_{LSJ} appears to be the best choice in both STSRS and STSYS. For the smaller sample size, $n = 25$, the severe overestimation by v_{LSJ} in Figure 4 does not lead to the same degree of overcoverage in Figure 5.

General summary findings and recommendations for (estimator, design) pairs, based on the simulation and the theory presented here and on previous work by Kott (1986), Royall and Cumberland (1981a, b), and Valliant (1987a), are as follows:

- For ($\bar{y}_{RS}$, STSRS)
  - In populations where a straight line through the origin is a reasonable model, either v_{RSJ} or v_{RS2} perform well both conditionally or unconditionally.
  - $\bar{y}_{RS}$ can have a substantial conditional bias in populations where the relationship of y to x is not well approximated by a straight line through the origin. If the number of strata is large, this problem will be reduced, but $\bar{y}_{RS}$ is very sensitive to departures from the model $E_\xi(y_{hi}) = \beta_h x_{hi}$.
- For ($\bar{y}_{RS}$, STSYS)
  - The STSYS plan will often reduce conditional bias.
  - In low variance populations all variance estimators will be overestimates if the model for y contains an intercept.
  - In high variance populations the jackknife may be preferable, but results here are inconclusive.
- For ($\bar{y}_{LS}$, STSRS or STSYS)
  - Conditional bias is less of a problem than with ($\bar{y}_{RS}$, STSRS).
  - All variance estimators are poor in small samples.
  - In larger samples v_{LSJ} is preferable.

Although no combination studied here is without flaw, the one which has the most to recommend itself is ($\bar{y}_{LS}$, STSYS) as long as stratum sample sizes are moderately large. The STSYS plan helps guard $\bar{y}_{LS}$ against conditional bias due to failure of model (1). In sufficiently large samples v_{LSJ}, in most cases, successfully estimates the variance even in systematic samples. The variance estimators in the classes studied by Royall and Cumberland (1978) should also perform similarly to the jackknife for ($\bar{y}_{LS}$, STSYS).

6. Conclusion

In populations where there is a reasonably smooth relationship between a target variable y and an auxiliary x, systematic sampling is a defensive strategy. Systematic sampling within strata from a frame sorted by x protects stratified ratio and regression estimators against certain kinds of model biases by producing samples which are more likely to be balanced on moments of x than are simple random samples. However, that bias protection does not always extend to

variance estimators. The dispersion of y about the regression line of y on x has a major effect on the performance of variance estimators. In populations where $\text{var}_\xi(y \mid x)$ is relatively low, variance estimators for the separate ratio estimator are subject to severe overestimation in systematic samples which persists even in large samples. In cases in which strata are formed based on the size of x and the regression of y on x can be approximated as a straight line within each stratum, the separate regression estimator is a good choice for controlling model bias. Systematic sampling will further protect the separate regression estimator against bias caused by departures from the straight line model within each stratum. Additionally, in the types of populations studied here, the jackknife variance estimator for the separate regression estimator performs well in systematic samples as long as stratum sample sizes are moderately large.

7. References

Andersson, C., Forsman, G., and Wretman, J. (1987). Estimating the Variance of a Complex Statistic: A Monte Carlo Study of Some Approximate Techniques. Journal of Official Statistics, 3, 251–265.

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

Cochran, W.G. (1977). Sampling Techniques. New York: John Wiley.

Deng, L.Y. and Wu, C.F.J. (1987). Estimation of Variance of the Regression Estimator. Journal of the American Statistical Association, 82, 568–576.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory, Vol. I. New York: John Wiley.

Heilbron, D.C. (1978). Comparison of Estimators of the Variance of Systematic Sampling. Biometrika, 65, 429–433.

Iachan, R. (1982). Systematic Sampling: A Critical Review. International Statistical Review, 50, 293–303.

Jones, H.L. (1974). Jackknife Estimation of Functions of Stratum Means. Biometrika, 61, 343–348.

Kott, P.S. (1986). Some Asymptotic Results for the Systematic and Stratified Sampling of a Finite Population. Biometrika, 73, 485–491.

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

Rao, P.S.R.S. and Rao, J.N.K. (1971). Small Sample Results for Ratio Estimators. Biometrika, 58, 625–630.

Royall, R.M. and Cumberland, W.G. (1978). Variance Estimation in Finite Population Sampling. Journal of the American Statistical Association, 73, 351–358.

Royall, R.M. and Cumberland, W.G. (1981a). An Empirical Study of the Ratio Estimator and Estimators of Its Variance. Journal of the American Statistical Association, 76, 66–77.

Royall, R.M. and Cumberland, W.G. (1981b). The Finite Population Linear Regression Estimator and Estimators of Its Variance – An Empirical Study. Journal of the American Statistical Association, 76, 924–930.

Royall, R.M. and Eberhardt, K.R. (1975). Variance Estimates for the Ratio Estimator. Sankhyā, ser. C, 37, 43–52.

Royall, R.M. and Herson, J. (1973). Robust Estimation in Finite Populations II: Stratification on a Size Variable. Journal of the American Statistical Association, 68, 890–893.

Scott, A.J. and Wu, C.F.J. (1981). On the Asymptotic Distribution of Ratio and Regression Estimators. Journal of the American Statistical Association, 76, 98–102.

Valliant, R. (1987a). Conditional Properties of Some Estimators in Stratified Sampling. Journal of the American Statistical Association, 82, 509–519.

Valliant, R. (1987b). Some Prediction Properties of Balanced Half-Sample Variance Estimators in Single-Stage Sampling. Journal of the Royal Statistical Society, ser. B, 49, 68–81.

Wolter, K. (1984). An Investigation of Some Estimators of Variance for Systematic Sampling. Journal of the American Statistical Association, 79, 781–790.

Wolter, K. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Wu, C.F.J. (1982). Estimation of Variance of the Ratio Estimator. Biometrika, 69, 183–189.

Wu, C.F.J. (1985): Variance Estimation for the Combined Ratio and Combined Regression Estimators. Journal of the Royal Statistical Society, ser. B, 47, 147–154.

Wu, C.F.J. (1986). Jackknife, Bootstrap, and Other Resampling Methods in Regression Analysis (and Rejoinder). Annals of Statistics, 14, 1261–1295 and 1343–1350.

Wu, C.F.J. and Deng, L.Y. (1983). Estimation of Variance of the Ratio Estimator: An Empirical Study. In Scientific Inference, Data Analysis, and Robustness, eds. G.E.P. Box, T. Leonard, and C.F. Wu. New York: Academic Press, 245–277.

Received September 1988
Revised November 1989