DOES SRS PROVIDE ADEQUATE BALANCE?

William G. Cumberland, U.C.L.A.
Richard M. Royall, Johns Hopkins University

1. INTRODUCTION

There is considerable controversy regarding the appropriateness of the use of prediction models for inference in finite populations. The Randomization Principle states that the only probability distribution on which reliable inferences can be made is that provided by the sampling plan. Adherents to this principle reject the contention of prediction theory that inferences should be made conditional upon the observed sample. Others, while admitting that superpopulation models are useful still relegate their role in inference to a secondary position, after that of the sampling plan. One implication of prediction theory regarding inference is the possibility of a bias in the ratio estimator when the sample chosen is not balanced. Since a simple random sampling (SRS) plan provides balance in expectation and since for large sample sizes the chance of getting a severely unbalanced sample is small, an important question to ask is how well does this sampling plan provide balance. Is it true for large sample sizes that one can ignore the problem of bias that may result from imbalance? To answer this question we examined the effect of increasing sample size on bias in the ratio estimator in two real populations when using a simple random sampling plan.

2. IMPLICATIONS OF PREDICTION THEORY

The ratio estimator is defined when two numbers are available for each population unit, one a positive constant x, known for all units, and the other an unknown y. A sample s of n units is drawn from the population; the ratio estimator for the population total

$$T=\Sigma_1^N y_i \text{ is } \hat{T}=(\Sigma_1^N x_i)(\Sigma_s y_i)/(\Sigma_s x_i).$$

The prediction approach to estimation assumes that the y's are realizations of random variables $Y_1,Y_2,\ldots,Y_N$. After the sample is observed, n of these Y's are known and N-n remain unobserved. Since the sample total can be written as $T=\Sigma_s y_i+\Sigma_r y_i$, the sample sum plus the sum over the nonsample units r, estimating T from the sample is equivalent to predicting the value $\Sigma_r y_i$ of the unobserved random variable $\Sigma_r Y_i$.

Relationships among the variables are expressed in a model for the joint distribution of the Y's; a model frequently used with the ratio estimator (Cochran 1953) specifies that the regression of Y on x is a straight line passing through the origin with variance proportional to x:

$$\begin{aligned} E(Y_i)&=\beta x_i \\ \mathrm{Var}(Y_i)&=\sigma^2 x_i \\ \mathrm{Cov}(Y_i,Y_j)&=0 \quad i\neq j \qquad (1) \end{aligned}$$

Let $\bar{x}_r,\bar{x}_s$, and $\bar{x}$ denote the average x values in the nonsample units, the sample units, and the whole population respectively. Define $\bar{y}_r,\bar{y}_s$, and $\bar{y}$ similarly. Under the simple prediction model (1) the ratio estimator $\hat{T}$ has minimum variance among linear unbiased estimators of the population total. Royall and Herson (1973) investigated what happens to the ratio estimator when the regression function is not as specified. They showed that $\hat{T}$ can incur a serious bias when the regression is not a straight line passing through the origin. For example, if in fact $E(Y_i)=\beta_0+\beta_1 x$ then the ratio estimator has a bias $N\beta_0(\bar{x}-\bar{x}_s)/\bar{x}_s$ which can be large if $\bar{x}_s$ is much different from $\bar{x}$. Note that the bias does not depend directly on the sample size n. This bias vanishes if the sample s is balanced on x ($\bar{x}_s=\bar{x}$). If the sample is balanced on higher powers of x as well, say on x^j for $j=1,2,\ldots,J$ then the ratio estimator is unbiased under any Jth degree polynomial regression model. Royall and Cumberland (1981) carried out an empirical study of the ratio estimator with sample size n=32 using six real populations for which the model (1) was a reasonable description of the logical relationship between two measurements on each population unit. This study verified that the bias of the ratio estimator can be substantial for those samples where $\bar{x}_s$ is much larger (or smaller) than $\bar{x}$.

Royall and Herson (1973), continuing their discussion, showed that the argument regarding balance generalizes not only to moments of x but also to any arbitrary function h of x which might show up as a linear term in the regression equation. Thus, if the sample is balanced on x and on h(x) (i.e. $\bar{x}_s=\bar{x}$ and $(1/n)\Sigma_s h(x_i)=(1/N)\Sigma_1^N h(x_i)$) then the ratio estimator will be unbiased under the model

$$E(Y_i)=\beta_o+\beta_1 x_1+\gamma h(x_i). \qquad (2)$$

This model actually includes the case of overlooking an important regressor, say z, in specifying the model. Letting $h(x_i)=z_i$ we see that the ratio estimator is unbiased under (2) when the sample chosen is such that both $\bar{x}_s=\bar{x}$ and $\bar{z}_s=\bar{z}$.

Although these results were derived without reference to the sampling plan used to select s, they might be interpreted as supporting the use of simple random sampling (or stratified random sampling) with the ratio estimator. An argument supporting this contention follows from Chebychev's inequality. Under simple random sampling, the probability of obtaining a sample s for which $|\bar{x}_s-\bar{x}|$ exceeds any fixed constant, say δ, is bounded above by a quantity proportional to $n^{-1}(1-nN^{-1})$. Thus as n increases, the chance of drawing a sample which is badly balanced ($|\bar{x}_s-x|>\delta$) decreases. When we apply this argument to the case of an overlooked regressor z it appears even more appealing. We can recognize samples that are badly balanced on x (and possibly reject them or adjust them) but we cannot do the same for unknown quantities $\bar{z}_s$ and $\bar{z}$. One could perhaps draw some comfort from the knowledge that although he cannot be certain that his sample is well-balanced on z, he did use a sampling procedure which rarely produces badly balanced samples.

The problem with this argument lies in the precise meaning of the expression "badly balanced" samples. Although the difference $|\bar{x}_s-\bar{x}|=\delta$ might represent negligible imbalance for a small sample size n, the same difference could signify severe imbalance for a larger sample size. The reason for this is that the severity of imbalance (and hence of the bias) must be compared to the variability of the estimator which also decreases with increasing n. Thus it is the rate at which balance improves (and bias vanishes) as n increases that it is critical and not simply the fact that as the sample size grows balance improves. Royall and Herson pointed out that the balance must improve faster than at the rate provided by simple random sampling if the bias is to become negligible relative to the standard error of the estimate.

Royall and Cumberland (1978) elaborated on this point in their evaluation of a collection of statistics for estimating the variance of the ratio estimator. Studying a group of variance estimators which includes the usual formula (derived under a random sampling plan) as well as the jackknife estimator, they showed that the coverage probability of the approximate 95% confidence interval $\hat{T}\pm1.95v^{\frac{1}{2}}$ converges to a value at least as great as .95 if balance improves faster than the random sampling rate.

As an illustration of these ideas consider the following. Suppose the true model is as follows:

$$\begin{aligned}E(Y_i)&=\beta_o+\beta_1 x_i\\ Var(Y_i)&=\sigma^2 x_i\\ Cov(Y_i,Y_j)&=0 \qquad i\neq j\end{aligned} \tag{3}$$

Under this model, the bias is $N(\bar{x}/\bar{x}_s-1)\beta_o$, and the error-variance $Var(\hat{T}-T)$ equals $N(N-n)\bar{x}_r\bar{x}(n\bar{x}_s)^{-1}\sigma^2$. We describe the bias as asymptotically negligible if the relative bias,

$$E(\hat{T}-T)[Var(\hat{T}-T)]^{-\frac{1}{2}}=n^{\frac{1}{2}}(1-nN^{-1})^{-\frac{1}{2}}(\bar{x}-\bar{x}_s)\beta_o(\bar{x}\bar{x}_s\bar{x}_r\sigma^2)^{-\frac{1}{2}} \tag{4}$$

converges to zero as n grows and f approaches 0. Regularity conditions which ensure that key population parameters remain stable as N grows must also be imposed. (For a discussion of these restrictions see Royall and Cumberland (1978).) Under these conditions, when the units are selected using simple random sampling, $n^{\frac{1}{2}}(1-nN^{-1})^{-\frac{1}{2}}(\bar{x}-\bar{x}_s)$ converges in law to a normally distributed random variable, not to zero. Thus in order that the relative bias (4) converge to zero, balance must improve faster than at the rate provided by simple random sampling. When balance improves at the simple random sampling rate the bias and the standard error are of the same order of magnitude. For example, if $\bar{x}_s=\bar{x}-kn^{-\frac{1}{2}}(1-nN^{-1})^{\frac{1}{2}}$ for some fixed k then the relative bias converges to a constant.

Also of importance is the behavior of variance estimators as the sample size grows. In particular, the effect of increasing n on the size of the relative bias in a variance estimator must be investigated.

Five variance estimators for $\hat{T}$ were studied in Royall and Cumberland (1981). One of these, the least squares variance estimator, proved to be an unreliable estimator. Another, commonly recommended in many sampling texts, was

$$v_C=Nn^{-1}(N-n)\Sigma_s d_i^2/(n-1)$$

where $d_i=y_i-(\bar{y}_s/\bar{x}_s)x_i$. This estimator showed a serious bias in the empirical study in badly balanced samples. An approximately unbiased estimator (under general variance models) that was also studied, $v_D=Nn^{-2}(N-n)(x_r x/x_s^2)\Sigma_s d_i^2/[1-(x_i/n\bar{x}_s)]$ performed better than v_C. Its performance was generally very similar to that of v_H, introduced by Royall and Eberhardt (1975). The last estimator studied the jackknife,

$$v_J=N(N-n)\bar{x}^2(n-1)\Sigma_s D(j)^2/n$$

where for every j in s, D(j) is the difference between the ratio $(n\bar{y}_s-y_j)/(n\bar{x}_s-x_j)$ and the average of these n ratios. Although asymptotically equivalent to v_D and v_H, v_J did perform differently from these two in tracking the mean square error of $\hat{T}$.

Failure of the condition $E(Y_i)=\beta x_i$ in (1) increases the expected value of each of the variance estimators and the mean squared error. If the sample is balanced so that $\hat{T}$ is unbiased then v_D and v_H are conservative in that their expected values exceed the actual mean square error. The relative bias in v_C is approximately $\bar{x}_s^2(\bar{x}\bar{x}_r)^{-1}-1$ which vanishes only in samples balanced on x.

3. EMPIRICAL STUDY

These implications were studied empirically in two of the six real populations used in previous empirical studies of prediction theory in finite population sampling. Detailed descriptions of the populations including scatter plots appear in Royall and Cumberland (1981). The two populations chosen for this study (Hospitals and Counties 60) were the two in which the ratio estimator showed the clearest bias as a function of $\bar{x}_s$ in an earlier study with n=32 of that estimator (Royall and Cumberland (1981). The other populations show results similar to those presented here and are omitted for brevity.

From each population, 1000 simple random samples of sizes 16, 32, and 64 were drawn. For each sample we calculated the ratio estimate $\hat{T}$ and the actual error $\hat{T}$-T as well as three of the five variance estimates which were studied in Royall and Cumberland (1981): v_C, v_D, and v_J. For each sample size the 1000 samples were arranged in order of increasing values of $\bar{x}_s$, and then grouped into 20 sets of 50 each. For each group we calculated the average value of $\bar{x}_s$, the average error, the mean square error (MSE) and the average value of the three variance estimators. We then plotted the average errors and values of $(MSE)^{\frac{1}{2}}$, $(\bar{v}_C)^{\frac{1}{2}}$, and $(\bar{v}_J)^{\frac{1}{2}}$ against the average values of $\bar{x}_s$.

Figures 1 through 6 show plots of the five trajectories. The one showing average error is labelled error, the ones showing $(\bar{v}_C)^{\frac{1}{2}}$, $(\bar{v}_D)^{\frac{1}{2}}$, and $(\bar{v}_J)^{\frac{1}{2}}$ are labelled C, D, and J. The population mean $\bar{x}$ is shown on the abscissa. For the different sample sizes the scales were changed to best accomodate the decreasing bias and error as n increased.

The most noticable feature of the three figures for each population is the bias curve. Regardless of the sample size, there is a clear bias when $\bar{x}_s\neq\bar{x}$. Furthermore, the sign and magnitude of the bias at the extremes relative to the MSE remains nearly the same for each sample size. The bias curves all cross the axis in the vicinity of $\bar{x}_s=\bar{x}$, and for each population they have a slope which remains relatively unaffected by the change in sample size. Thus the extreme values observed in the bias decrease as n increases only because the range of the average values of $\bar{x}_s$ becomes smaller as the sample size increases. The value of the MSE at balance decreases as n increases by an amount commensurate with the increase in sample size. What is clear from these figures is that increasing the sample size is not sufficient to guarantee that $\bar{x}_s$ will sufficiently close to $\bar{x}$ so as to make the relative bias in $\hat{T}$ resulting from imbalance on x negligible.

The bias in v_C persists as well, as n increases four-fold although the relative bias at the extremes becomes less severe as n grows. Increasing n has helped v_C somewhat, but not enough to consider it as a serious competitor to the bias-robust estimators v_D and v_J which are much better in tracking the MSE than v_C. However, it is also clear that these bias-robust variance estimators are still only dependable as an estimate of the variability when the sample is balanced on x. As expected, as n increases, v_D and v_J have trajectories that become more and more alike.

4. DISCUSSION

These results demonstrate again the importance of making inferences regarding finite populations based on prediction models, not on the probability sampling distribution. It is the property of the sample (balance) not the sampling plan that is important. Although a simple random sampling plan produces samples which are approximately balanced it does not force balance fast enough to guarantee that we can ignore properties of the sample when making inferences even for large sample sizes.

REFERENCES

[1]Cochran, W. G. (1953), Sampling Techniques, New York: John Wiley.

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

[3]Royall, R. M. and Cumberland, W. G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of Its Variance", Journal of the American Statistical Association, 76, 66-88.

[4]Royall, R. M. and Eberhardt, K. R. (1975), "Variance Estimates for the Ratio Estimator", Sankhya, Ser. C, 37, 43-52.

[5]Royall, R. M. and Herson, J. H. (1973), "Robust Estimation in Finite Populations I", Journal of the American Statistical Association, 68, 890-893.

FIGURE 1

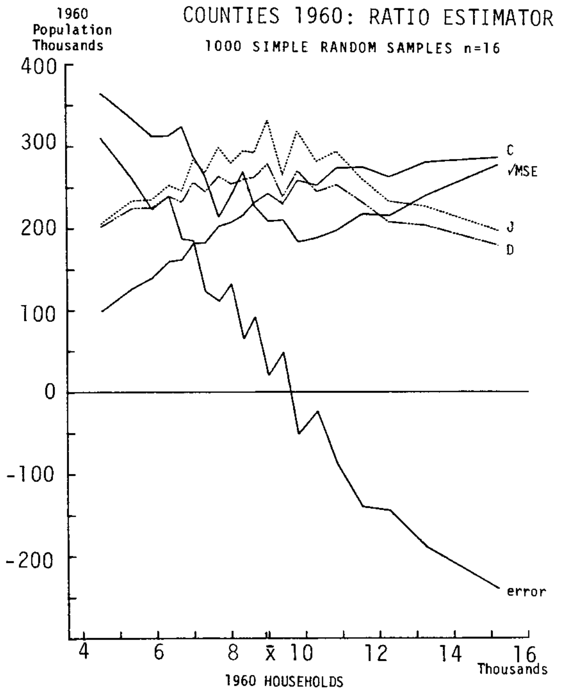


FIGURE 2

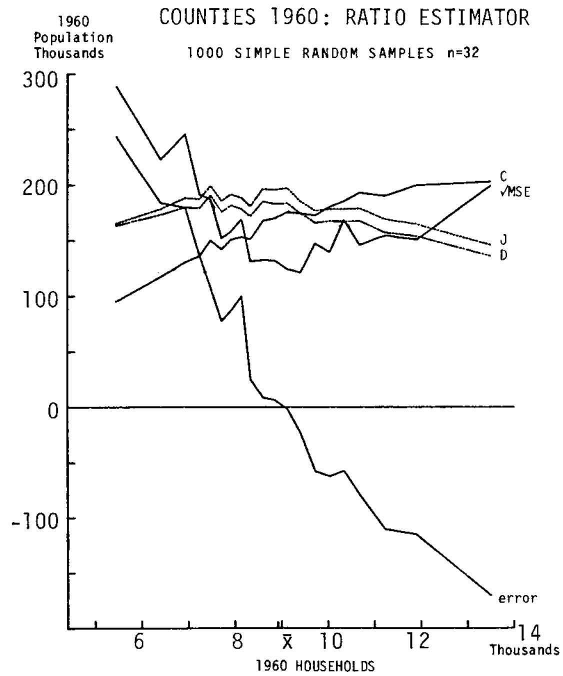


FIGURE 3

COUNTIES 1960: RATIO ESTIMATOR

1000 SIMPLE RANDOM SAMPLES n=64

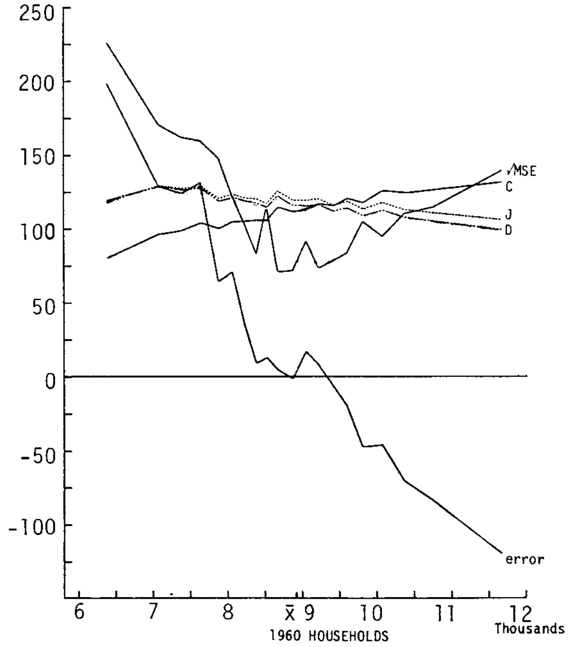


FIGURE 4

HOSPITALS: RATIO ESTIMATOR

1000 SIMPLE RANDOM SAMPLES n=16

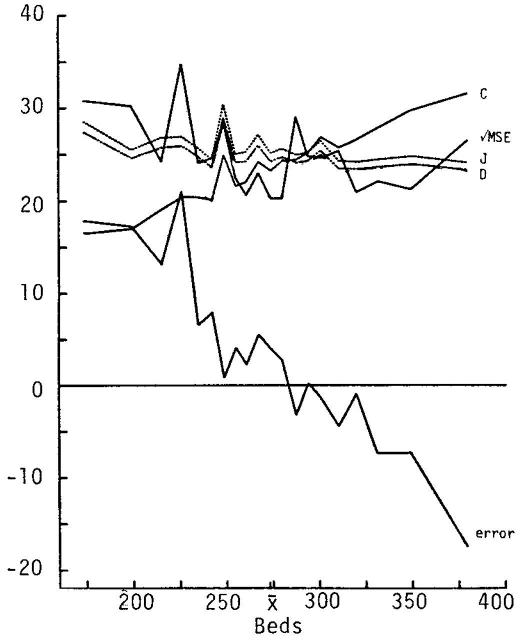


FIGURE 5

HOSPITALS: RATIO ESTIMATOR

1000 SIMPLE RANDOM SAMPLES n=32

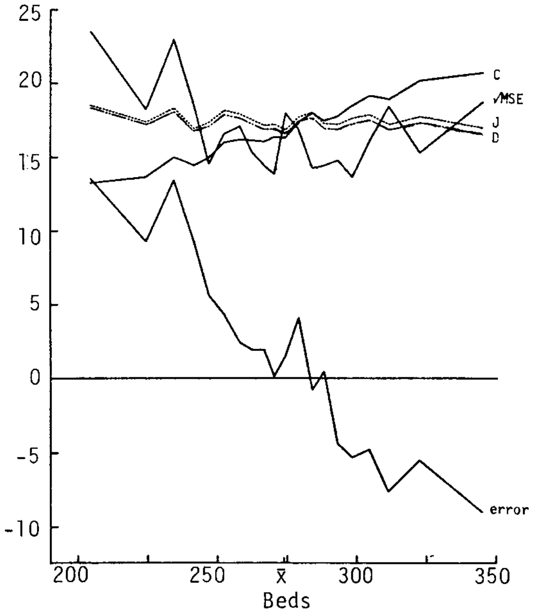
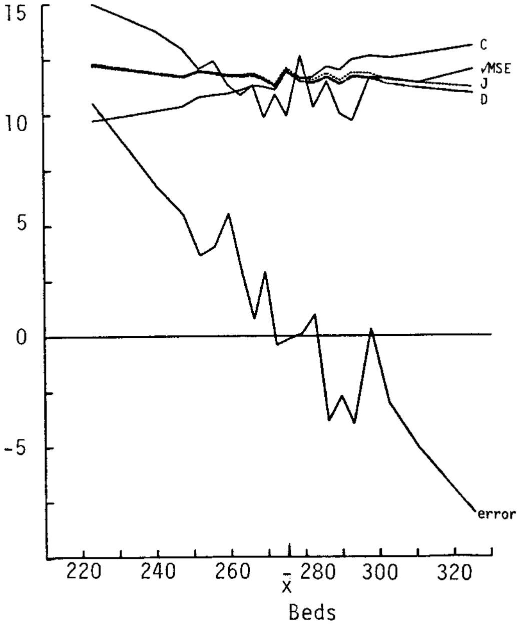


FIGURE 6

HOSPITALS: RATIO ESTIMATOR

1000 SIMPLE RANDOM SAMPLES n=64


Patients Thousands
C
√MSE
J
D
error
x̄
Beds