INTERVAL ESTIMATION IN UNEQUAL PROBABILITY SAMPLING

S.H. Biyani, East Carolina University, C.K. Midha, University of Florida

1. INTRODUCTION

When information on some auxiliary variable x related to the variable of interest y is available for all units of a finite population, unequal probability sampling is frequently used for increasing the efficiency of the estimation, generally in conjunction with the Horvitz-Thompson estimator of the population total (or mean). Several optimality results for the Horvitz-Thompson estimator of total are well known. The choice of a variance estimator for the Horvitz-Thompson estimator is, however, not clear. Some theoretical problems with the Yates-Grundy and other design-unbiased estimators were pointed out in Biyani (1980a). Some new estimators were derived using a Random Permutation Model and the performance of several estimators was compared empirically in Biyani (1980b), using the mean squared error as criterion. Cumberland & Royall (1980) have compared several estimators of the model-variance of the Horvitz-Thompson estimator, conditional on the sample, under a certain superpopulation model.

Using mean square error as a criterion may be most appropriate, if the variance estimate is to be used for survey design purposes. If, however, a variance estimate is to be used primarily for the interval estimation of the population total, then the performance of the interval estimators based on different variance estimators must be considered. In this paper, we have empirically compared the following aspects of interval estimation,

(a) the actual coverage percentages of nominal 95% confidence intervals based on the percentiles of student's t distribution

(b) the average widths of true 95% confidence intervals, based on the knowledge of the actual distribution of "Studentized estimates" of the form $(\hat{T} - T)/[\hat{V}(\hat{T})]^{1/2}$.

2. THE ESTIMATORS AND DESIGN

Let T denote the population total and e_{HT} denote the Horvitz-Thompson estimator of T. We consider the following estimators of the variance of e_{HT} over all possible samples: the Yates-Grundy estimator (v_{YG}) a weighted ratio estimator (v_F) due to Fuller (1970), an unweighted ratio estimator (v_R), a "best" estimator[1] (v_B), and "best" model-unbiased estimator[1] (v_{BMU}). The last three estimators are defined in Biyani (1980b). For a stratified sampling design, and a corresponding within strata random permutation model, the corresponding estimators can be obtained simply by adding the within strata estimators, except in case of v_B. For the latter, the expression in the stratified case becomes considerably more complicated and requires additional model assumptions. Hence v_B was not considered in the stratified case.

Several natural populations were used in the study, including some small unstratified populations and larger stratified ones. These are listed in Table 1. From each population one thousand independent samples were drawn by computer, using Sampford's (1967) scheme of sampling with probability proportional to X. The stratification was based on the X values and the allocation of sample size was proportional to the sums of X within strata.

Let $t_{(p)}$ denote the (100p)th percentile of Student's t distribution with (n-1) degrees of freedom and let $t_{i(p)}$ denote the corresponding percentiles of the actual distribution of the "Studentized estimators" $(e_{HT} - T)/v_i^{1/2}$, $i =$ YG, F, R, B, BMU. For the "95% Confidence Intervals" based on Student's t Distribution, (given by $e_{HT} \pm t_{(p)}v_i^{1/2}$), the observed coverage proportions are given in Table 2. Table 3 compares the average widths of the true 95% confidence intervals of the form $(e_{HT} + t_{i(p)}v_i^{1/2}$, $e_{HT} + t_{i(1-p)}v_i^{1/2})$, where the percentiles $t_{i(p)}$ are obtained from the empirical distribution of the t_i's over one thousand samples. These comparisons are made only for the unstratified populations.

TABLE 1. Populations used in the study

Pop. No.	Source	x	y	pop. size	no. of strata
1	Cochran (1963)	1920 population	1930 population	20	1
2	Jessen (1978)	area of farm	area under corn	20	1
3	Scheaffer, et al. (1979)	real estate value 2 yrs. ago	current value	20	1
4	Kish (1965)	No. of housing units per block	renter occupied housing units	270	3
5	Fortune (June 1981)	assets of corporation	sales	480[2]	6

TABLE 2. Estimated coverage probabilities of 95 percent confidence intervals based on Student's t distribution

Pop. No.	Sample Size	Variance Estimator Used				
		YG	F	R	B	BMU
1	5	.76	.79	.75	.70	.94
2	5	.95	.95	.97	.94	.95
3	5	.95	.95	.95	.93	.95
4	30	.95	.95	.95	not computed	.95
5	50	.94	.94	.94	not computed	.94

TABLE 3. Average widths of confidence intervals based on the empirical distribution of the Studentized estimates with actual coverage probability .95

Pop. No.	Sample Size	Variance Estimator Used YG	F	R	B	BMU
1	5	1396	1321	951	1014	1024
2	5	863	886	893	893	898
3	5	21.2	21.1	21.5	21.2	21.2

3. THE RESULTS

In sampling with probability proportional to x, the Horvitz-Thompson estimator and its variance estimators are functions of the ratios y_i/x_i. Thus the behavior of the estimators depends on the distribution of these ratios. In population 1, the distribution of y_i/x_i is considerably skewed, with skewness coefficient $\gamma_1 = 3.1$, and not too surprisingly, we find that the coverage probabilities of the confidence intervals (based on sample size 5) are much below the nominal values, ranging from 70 to 79% for the different variance estimator. In all other populations, we find the actual coverage to be surprisingly close to the nominal value, especially for the small samples, since the sampling design and the forms of the variance estimators are considerably different from those assumed in the t distribution. The distributions of the ratios y_i/x_i also differ from the normal distribution in varying degrees. In particular, two of the strata in the "Fortune 500" population have kurtosis coefficients (γ_2) in excess of 10 and the skewness coefficients range up to 3.1. However, the relatively large sample size has apparently helped bring the coverage proportions very close to the nominal value.

We also note that the differences among the different variance estimators are, for the most part, neglegible. The estimator v_B gives slightly less coverage than the rest, which is to be expected, since it is a shrinkage estimator.

In the comparison of the average widths of the true 95% confidence intervals, the only population for which substantial differences are observed, is population 1. For this population the design-based Yates-Grundy and Fuller estimators give considerably wider intervals than the rest, while the unweighted ratio estimator gives the narrowest. This seems to be in line with the empirical results in Biyani (1980b) comparing the mean squared errors. It was observed there that the so called model-independent estimators performed worse than the model-based ones when the model was seriously violated.

Although we are not sure that these results can be generalized to most types of populations, the following conclusions are suggested by this study.

1. Student's t distribution may be used to construct reasonable confidence intervals when the distribution of the ratios y_i/x_i departs only moderately from normal.
2. When confidence intervals are based on t distribution, the choice of variance estimator is not critical in determining the actual coverage probability.
3. If an approximation to the true distribution of the Studentized estimators can be obtained for a nonnormal population, then the average width of the confidence interval can be affected by the choice of variance estimator, and model-based estimators are likely to give narrower confidence intervals.

FOOTNOTES

[1] Under a Random Permutation Model on y_i/x_i, with kurtosis approximately that of normal distribution, and within the class of all quadratic invariant estimators.

[2] From the list of 500 largest corporations, excluding those with assets over $10 billion.

REFERENCES

1. Biyani, S. H. (1980a). "On inadmissibility of the Yates-Grundy Estimator in Unequal Probability Sampling", Journal of the American Statistical Association, 75, 709-712.
2. Biyani, S. H. (1980b). "On Variance Estimation in Unequal Probability Sampling". American Statistical Association 1980 Proceedings of the Section on Survey Research Methods, 634-637.
3. Cumberland, W. G. and R. M. Royall (1980). "Prediction Theory in Finite Population Sampling: The Horvitz-Thompson Estimator and Estimates of its Variance", American Statistical Association 1980 Proceedings of the Section on Survey Research Methods, 325-330.
4. Fuller, W. A. (1970). "Sampling with Random Stratum Boundaries," Journal of the Royal Statistical Society, Ser. B, 32, 209-226.
5. Horvitz, D. G. and Thompson, D. J. (1952). "A Generalization of Sampling without Replacement from a Finite Universe," Journal of the American Statistical Association, 47, 663-685.
6. Sampford, M. R. (1967). "On Sampling without Replacement with Unequal Probabilities of Selection," Biometrika, 54, 499-513.
7. Yates, F. and Grundy, P. M. (1953). "Selection without Replacement from within Strata with Probability Proportional to Size," Journal of the Royal Statistical Society, Ser. B, 15, 235-261.