

DISCUSSION OF PAPERS BY RAO/SKINNER AND SINGH/WU

M.D. Bankier¹

I am pleased to discuss these two excellent papers on multiple frame estimation. They take quite different approaches to this problem. To place these papers in context, the multiple frame literature will be briefly reviewed.

Hartley (1962) was the first to systematically examine multiple frame estimation. In the decade which followed, various authors (most significantly Fuller and Burmeister, 1972) improved on Hartley's basic approach. In all these papers, estimates were generated separately for each frame and then combined such that the variance was minimized. All these techniques became complicated if more than two frames were used. Many required separate weights for each sample characteristic. Also, most of these techniques were restricted to unstratified simple random sampling from each frame.

Bankier (1986) suggested that a dual frame stratified sample design could be viewed as two samples selected independently from the same frame. He assumed that it was possible to identify and unduplicate those sampled units that fell in both samples. This then allowed the two samples to be treated as one. A Horvitz-Thompson estimator was determined based on the probability of a unit falling into at least one of the two samples. The Horvitz-Thompson weights were adjusted through raking ratio estimation (generalized regression estimation could be used as well) so that sample estimates agreed with known stratum population counts. This significantly reduced the variance. The single set of adjusted weights could be used to produce estimates for all the sample characteristics in a survey. This methodology was used in a dual frame stratified sample of business tax forms.

This method can be easily extended to multiple frame sample designs where the π_i and π_{ij} are known for each frame's sample design for the units selected in at least one frame's sample. This latter condition will not be satisfied, however, for certain complex sample designs such as multi-stage sampling with clusters.

This method, however, has the advantage of allowing well known single frame estimation techniques to be used for many multiple frame sample designs.

Skinner (1991) studied the asymptotic efficiency of the raking ratio estimator proposed by Bankier compared to the estimator proposed by Fuller and Burmeister (see equation (2) of Skinner and Rao, 1996). Dual frames with an unstratified simple random sample selected from each frame were assumed. In addition, the sampling fractions were taken to be negligibly small in the asymptotic arguments. Skinner demonstrated, under certain assumptions, that the Fuller and Burmeister estimator was a MLE of Y . He then showed that the raking ratio estimator was uniformly less efficient than the estimator proposed by Fuller and Burmeister. He pointed out, however, that with several frames or more complex sample designs, the derivation of the Fuller/Burmeister estimator became very complex or might even break down. He then stated that the raking ratio estimator extended naturally to these more realistic situations.

Skinner, Holmes and Holt (1994) discussed multiple frame sampling for multivariate stratification. They used the equation (11) estimator of the Skinner and Rao paper (1996) which is approximately equal to the Horvitz-Thompson estimator when the sampling fractions are small. It does not require the checking for duplicates between the samples and its variance is simpler than the Horvitz-Thompson estimator. Raking ratio estimation was used to incorporate auxiliary information about the strata sizes into the weights. A simulation study was performed which showed that the raking ratio estimator achieved a dramatic reduction in the variance over the Horvitz-Thompson estimator. This methodology was then used to do a survey of farms to measure the readership of various agricultural publications.

Rao and Skinner, in their paper, were able to use the pseudo-maximum likelihood approach to generalize the Fuller/Burmeister estimator. The Rao and Skinner pseudo-maximum likelihood estimator has great appeal

¹ Michael D. Bankier, Social Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6.

in that it can be used under complex designs. It is design consistent, has a simple form and will probably be efficient in practice. It is also intuitively appealing in its relatively simple form and how that can be easily related to the original Fuller/Burmeister estimator. It does not require that duplicates in the sample be identified or that the selection probabilities of the sampled units be known over all frames. It is also very attractive in the way that the same survey weights are used for all variables. It would be difficult, however, to extend their technique to multiple frame sample designs. Rao and Skinner also note that their simulation study is somewhat limited. For example, the single frame estimator would have been more efficient (while still being less efficient than other estimators in some situations) if raking or regression estimation had been used to incorporate information about the number of units in Frame A and Frame B.

Singh and Wu, in their paper, are able to extend the well accepted generalized regression estimation technique to the multiple frame estimation problem. Their technique is applicable to several frames and complex sample designs. Their technique, like the Rao and Skinner estimator, uses a single set of weights for all variables, does not require duplicates to be identified and does not require the selection probabilities over all frames to be known. Thus, their approach has great promise. They use a modified regression estimation technique which allows the use of more general predictors such as the difference of two estimates. Predictors, however, have to be chosen care-

fully since the general or modified regression estimation technique can break down or produce negative weights if too many predictors are used. Also, after the most important predictors are incorporated, the reduction in the mean square error can be slight. For these reasons, it may sometimes be sufficient to use the known auxiliary information as predictors without using the difference of two estimates for one or more sample characteristics. In other cases, however, the difference of two estimates may be one of the better predictors. The comparison, in the Singh and Wu simulation study, of the generalized regression and the Kalton/Anderson estimators to the other estimators would have, again, been more relevant if additional auxiliary information had been incorporated into the former estimators. The authors have stated that this will be done in the final version of their article.

ADDITIONAL REFERENCES

- Skinner, C.J., Holmes, D.J., and Holt, D. (1994). "Multiple Frame Sampling for Multivariate Stratification", *International Statistical Review*, 62, 3, 333-347.
- Skinner, C.J., and Rao, J.N.K. (1996), "Estimation in Dual Frame Surveys with Complex Designs", *Journal of the American Statistical Association*, 91, 349-356.