DISCUSSION

Richard Royall, The Johns Hopkins University

Both of these papers are good examples of the process of developing estimators using conventional finite population sampling theory. We can pick out three important stages in this process:

1. Assuming observations on certain variables are available, scratch your head and write down an estimate which has some intuitive appeal.

2. Try to get a handle on bias and variance. (Having done this a few times and produced a few estimates, compare their mse's. Find one estimate is better than another under certain assumptions about population parameters.)

3. Get a real population and try out the estimates to see which works better under various realistic conditions.

After, or along with, these three basic steps comes the secondary problem of measuring the uncertainty in an estimate. This usually boils down to finding a nearly-unbiased estimate of an approximation to the variance or mse. Unfortunately, these variance estimates rarely have the "face validity" or obvious reasonableness of the original statistic. For example, the synthetic estimates are in a gross sense reasonable. They obviously won't give really precise estimates, but they will be, if not in the right ballpark, at least in the right city. The variance estimate, on the other hand, might not even be on the right planet -- a negative variance estimate might be reasonably described as "lost in space".

I would like to see a different approach used, and I think the problem at hand, estimation for small areas, is one in which this approach would yield different and better results than the conventional one, particularly with regard to providing estimates of mean square errors to use as measures of uncertainty. This approach would begin not with an estimate, but with an attempt to express the basic relationships among the relevant variables through a probabilistic model. The model would then be used to generate estimates, provide a framework for comparing estimates, and to provide estimates of standard errors. Often the conventional intuitive estimates are optimal or nearly so under a simple probability model, but sometimes the model suggests practical improvements, especially in the conventional measures of uncertainty. Varying the model can give valuable insight into the robustness of estimators. This general approach has been called "the prediction approach" because, when viewed in the context of (super-population) probability models, many finite population inference problems are mathematically equivalent to classical prediction problems. "The prediction approach" actually has many facets -- simple linear least-squares [4,5], esoteric fiducial [2], and full-blown Bayesian [1] prediction techniques are only some of those available.

What would be the results of applying the prediction approach (least-squares variety) to the present problem? Two important general results I would expect are:

1. New estimators and new variance estimators for the old ones.

2. New insight into relationships among estimates already proposed, and increased understanding of their strengths and weaknesses.

Specifically ... I don't know what results would be obtained. The work has not, to my knowledge, been done. But some relevant comments can be made.

The "ratio-correlation" method and the "regression-sample data" method aren't so much two different methods as two different estimates, each more or less appropriate under its own prediction model. Although the two models do employ slightly different functions of births, etc. as regressors, the most important differences between these two estimates come not from different assumptions concerning the relationships among the relevant variables, but from different assumptions about available data. The ratio-correlation method is not allowed to use the sample data, while the regression method employs only data from the sample and the most recent census, ignoring the previous census. In both models the total for a local area at one time is represented as a multiple of the total at an earlier time plus an error whose variance is proportional to the square of the earlier total. (We might ask whether a different error-variance might be more appropriate. If it is, this would suggest different estimates.) The multiplicative factor for a given area is a function of various bits of data concerning births, deaths, number of school children, etc. in that area. In this factor are certain coefficients which change over time. The "ratio-correlation method" uses estimates of out-of-date coefficients, while the "regression-sample-data" method uses less precise estimates of more timely coefficients.

When the "ratio-correlation" estimate is used as a "symptomatic indicator" in the "regression-sample data" estimate, we are, in effect, using a particular linear combination of estimates of the "old" coefficients and the "new". I think a formal model, in which coefficients for one time interval are stochastically related to those for an earlier interval, would be quite useful in evaluating this and other estimates based on all the data, from both censuses as well as the sample.

In much the same way, the choice between direct estimation and imputation in the synthetic estimation paper is really the choice between a high-variance estimate of a directly relevant parameter and a low-variance estimate of a different quantity. The choice need not be made -- surely a combination of the two is better than either taken alone. A probability model can express the

relationships whose existence makes the whole notion of "imputation" reasonable. Such a model would generate (via standard linear prediction techniques) statistics which would give proper weight to both direct and imputed estimates.

I think, however, that one of the possibilities suggested by Gonzales and Waksberg in their Vienna paper [3] is more promising -- before really good local area estimates are produced, the synthetic estimation approach must move towards Ericksen's in making greater use of available local area variables.

REFERENCES

[1] Ericson, W.A., "Subjective Bayesian Models in Sampling Finite Populations," Journal of the Royal Statistical Society, Ser. B, 31, No. 2 (1969), 195-224.

[2] Kalbfleisch, J. and Sprott, D.A., "Applications of Likelihood and Fiducial Probability to Sampling Finite Populations," in Johnson, N.L., and Smith, H., Jr., eds. New Developments in Survey Sampling, New York: John Wiley and Sons, Inc., 1969.

[3] Gonzalez, M.E., and Waksberg, J., "Estimation of the Error of Synthetic Estimates." Paper presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, August 18-25, 1973.

[4] Royall, R.M. and Herson, J., "Robust Estimation in Finite Populations I," Journal of the American Statistical Association, 68, No. 344 (1973), 880-9.

[5] Royall, R.M., "Linear Regression Models in Finite Population Sampling Theory," in Godambe, V.P., and Sprott, D.A., eds., Foundations of Statistical Inference, Toronto: Holt, Rinehart and Winston of Canada, Ltd., 1971.