Journal of Official Statistics, Vol. 28, No. 3, 2012, pp. 349–351

Discussion

Alan H. Dorfman[1]

Rod Little is to be congratulated for giving us an article which is likely to galvanize a renewed discussion on the very basics of survey sampling. There is much to agree with, in particular the point captured in his catch-phrase "inferential schizophrenia" to characterize two opposed approaches to handling survey data that get embraced, not only by one and the same statistical agency, but by individual statisticians. The small area versus large sample divide has an element of irony: precisely where there is least data on which to assess our models where we need to, models are leaned on; where there is ample data with which to assess and possibly improve the model, the modeling (predictive) approach is commonly eschewed.

Little is Bayesian but very sympathetic to design-based protocol – at bottom what he mainly means by "calibrated" is to use a "robust model" which he takes to mean a model which automatically reflects the (probabilistic) sample design. In the currently dominant model-assisted approach (Särndal et al. 1992), an estimate based on a model is modified by a sum of sample-weighted residuals that is intended to correct for model failure. In the case of some models, the adjusting sum is algebraically zero; these are the robust models that Little prefers. Thus calibrated Bayes sample estimation can be viewed, in large samples, as a particular version of model-assisted sample estimation. (It may be noted that the basic device, of guarding against model failure by adding weighted sums of residuals at the price of some possible loss in efficiency, also has its place in pure prediction-based sampling (Chambers et al. 1993; Dorfman 1993)).

Outside the sampling context, the calibrated Bayes protocols of Box (1980), Rubin (1984) and others have a much stronger diagnostic edge than we see in this article. In that very interesting "hybrid" tradition, inference is Bayesian, based on models for prior and data, but these models can be called into question by the data at hand through tests in the frequentist tradition. Little's frequentist "formulating and assessing the model" means primarily making sure it fits the sample protocol, quite apart from bringing to bear residual analysis, the predictive distribution, or other post-collection data scrutiny. The models he prefers could be chosen *before* any data is seen.

In general terms, Little refers to diagnostics favorably. In particular, he notes the idea of comparing sample-weighted and non-sample-weighted estimates under a given model, as a test of the validity of the model, and cites Dumouchel and Duncan (1983). The place of this sort of diagnostic in Little's calibrated Bayes is a little puzzling: the approach

[1]U.S. Bureau of Labor Statistics, 2 Massachusetts Ave., NE, Washington DC 20212-0001, U.S.A. Email: dorfman.alan@bls.gov

advocated would do estimation using a "robust model" that embodies the sampling structure. Suppose a test or inspection shows marked discrepancy of the estimates between this robust version and an alternative version. What does one then do? In particular, does one simply fall back on the weighted "robust" version of the model? Or must one seek another "more robust" model, and if so how?

But more worrisomely puzzling is the fact that Little cites favorably the famous paper by Hansen et al. (1983) (HMT), on the *in*effectiveness of diagnostics in the sampling context (Section 4.2). Compare also Little (2004, Example 8), where he simply dismisses the counter-study in Valliant et al. (2000, Section 3.7), that refutes the HMT contention. The HMT paper has been a clarion call for design-based samplers for thirty years, but it is best regarded as a naïve, misleading paper. What HMT actually shows is that inference does not stand well in the presence of *half-baked* diagnostics. The sampling world could stand much more use of the tools of regression diagnostics, especially those that can be automated.

Little recognizes the existence of a model-based theory that ignores prior distributions (option (b) in Section 2), often referred to as the prediction-based approach, as in Valliant et al. (2000). One of the important implications of the prediction-based theory of sampling is that sample designs that eschew randomization not only *can* be acceptable, but are under some circumstances *preferable.* Robustness is often to be sought in various forms of balanced design, chosen purposively. From this point of view, focus on choosing priors and incorporating the random design into the chosen model seems like a major *distraction.*

Little talks about the advantages of combining standard survey data with administrative data, etc., but it is not obvious how his version of robust models, which accommodate the sample weights, enfold this non-random-design data. This same worry applies to small area estimation.

Small area estimation is a central consideration in this article. A special virtue of calibrated Bayes is said to be that it reaches down to where the sample is small, in contrast to both the design-based and standard prediction-based approaches. Credible intervals based on hierarchical Bayes will tend to be larger and therefore *better* than other approaches to small area estimation, including the direct approach. (In effect, Little is saying that "calibrated Bayes" boils down to "Bayes" in the small area context.)

"Better" however is not equivalent to "good". Although the field of small area estimation seems well established and has drawn on the ingenuity of many very talented statisticians and economists, must it not in general be regarded as one of the more tenuous of statistical activities? It is typically carried out under a great deal of budgetary and, often, political pressure. Here is a quote from a report on SAIPE:

"Although the Census Bureau's 1995 estimates of poor school-age children have potentially large errors for many school districts, the [National Academy of Sciences] panel nonetheless concludes that they are not inappropriate or unreliable to use for direct Title I allocations to districts as intended by the 1994 legislation. In reaching this conclusion, the panel interprets 'inappropriate and unreliable' in a relative sense. Some set of estimates must be used to allocate funds . . . . [these] estimates are generally as good as–and, in some instances, better than– estimates that are currently being used." (The National Academies 1999).

But are they good enough? I venture that, despite a lot of theoretical work, it is very difficult in small area estimation to verify the validity of our models. Few small area projects are able, as in Beresovsky et al. (2011), to verify the coverage of their prediction or credible intervals, if they produce them at all. Without such verification, must not our inferences remain dubious? Do we not need some good criteria by which to recognize occasions when the data are simply insufficient for sound inference? I am therefore in doubt that wrapping small area estimation in the protective mantle of an overall "Bayesian philosophy" would indeed be a positive thing.

References

Beresovsky, V., Burt, C.W., Parsons, V., Schenker, N., and Mutter, R. (2011). Application of Hierarchical Bayesian Models with Poststratification for Small Area Estimation from Complex Survey Data. 2011 Proceedings of the American Statistical Association. Alexandria: Survey Research Methods Section.

Box, G.E.P. (1980). Sampling and Bayes Inference in Scientific Modeling and Robustness. Journal of the Royal Statistical Society, Series A, 143, 383–430.

Chambers, R.L., Dorfman, A.H., and Wehrly, T.E. (1993). Bias Robust Estimation in Finite Populations using Nonparametric Regression. Journal of the American Statistical Association, 88, 268–277.

Dorfman, A.H. (1993). Comparison of Design-based and Model-based Estimators of the Finite Population Distribution Function. The Australian Journal of Statistics, 35, 29–41.

Dumouchel, W.H. and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analysis of Stratified Samples. Journal of the American Statistical Association, 78, 535–543.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys (with discussion). Journal of the American Statistical Association, 78, 776–793.

Little, R.J. (2004). To Model or not to Model? Competing Modes of Inference for Finite Population Sampling. Journal of the American Statistical Association, 99, 546–556.

Rubin (1984). Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistican. Annals of Statistics, 12, 1151-1172.

Särndal, C.-E., Swensson, B., and Wretman, J.H. (1992). Model Assisted Survey Sampling. New York: Springer.

The National Academies (1999). Small-Area Estimates of School-Age Children in Poverty: Interim Report 3. C.F. Citro and G. Kalton (eds). Washington: National Academies Press.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: Wiley.

Received June 2012