Statistical Science
2011, Vol. 26, No. 2, 260–261
DOI: 10.1214/11-STS346B
Main article DOI: 10.1214/10-STS346

Discussion of "Impact of Frequentist and Bayesian Methods on Survey Sampling Practice: A Selective Appraisal" by J. N. K. Rao

J. Sedransk

Abstract. This comment emphasizes the importance of model checking and model fitting when making inferences about finite population quantities. It also suggests the value of using unit level models when making inferences for small subpopulations, that is, "small area" analyses.

Key words and phrases: Diagnostics, hierarchical structure, model checking, model fitting, small area statistics, unit level models.

Professor Rao has written an excellent review of the alternative methods of making inference for finite population quantities. This is an underserved field of research and, hopefully, this paper will encourage some readers to make contributions to this important, practical area.

Rather than commenting on detailed aspects of the paper, I will discuss two broad areas. Both are treated briefly in this article, but have not been considered in the survey sampling literature as fully as I think they should be. The first is the fitting of models to complex survey data, and the second is model checking.

Except for the design-based approach, all of the inferential methods described in this paper rely significantly on models. And, over the past thirty years great strides have been made to develop models that are consistent with observed data. My impression, though, is that survey statisticians have been slow to adopt these methodological advances. In Section 1 Rao writes, referring to Hansen, Madow and Tepping (1983), "Unfortunately, for large samples [model dependent approaches] may perform very poorly under model misspecifications; even small model deviations can cause serious problems." This example (in Hansen, Madow and Tepping, 1983) was analyzed almost thirty years ago and *only by the authors*. One would hope that current methodology and skills in data analysis would provide an improvement over the Hansen, Madow and Tepping (1983) "straw man," the usual ratio estimator. As noted by Hansen, Madow and Tepping (1983), one should use robust methods. But, there have been other advances in diagnostic techniques and inferential methods (e.g., model averaging). Moreover, this is a *single* example and, before drawing general conclusions, it would be preferable to consider this example again and analyze other examples typical of sample survey data. Finally, though, it is important to note that there are challenging problems in modeling data from complex sample surveys because there may be several stages of cluster sampling, small sample sizes (typically in inconvenient places), possible selection biases, nonresponse and measurement errors.

When the objective is inference for "small area" quantities there are special issues with modeling. In my experience almost all of the *applications* use an area-level model; see, for example, Section 5 of this paper and Rao (2003). (Moreover, there are many applications that are not reported in the refereed literature, and I do not know of any that use a unit-level model.) In a small area analysis one is concerned about the quality of the direct estimator, $\hat{\theta}_i$, and, thus, uses a model that adds information about other small areas to improve inference about θ_i. Clearly, then, the quality of the estimated variance of $\hat{\theta}_i$, $v(\hat{\theta}_i)$, is even more questionable. (Rao notes this in Section 5, i.e., "the second

J. Sedransk is Professor (Emeritus), Department of Statistics, Case Western Reserve University, 10600 Crossing Creek Road, Potomac, Maryland 20854, USA (e-mail: jxs123@cwru.edu).

assumption of known sampling variances is more problematic.") Moreover, is it reasonable to assume that $(\hat{\theta}_i - \theta_i)/\sqrt{v(\hat{\theta}_i)}$ is satisfactorily approximated by a standard normal distribution? A transformation of $\hat{\theta}_i$ may be helpful. But, choosing the transformation and verifying that the associated standardized quantity is approximately distributed as $N(0, 1)$ is a challenging exercise. There is a better way, though, and that is to model the *unit* level data as, for example, in Battese, Harter and Fuller (1988), Malec, Sedransk, Moriarity and LeClere (1997) and Malec (2005). Doing so has a second benefit. In such circumstances one can investigate alternative ways to make inference about the θ_i from an *area-level* model (because the microdata are now available and one can investigate sampling distributions of the transformed $\hat{\theta}_i$'s).

Model checking is an essential part of the modeling process. In Section 5, Rao writes that "some of the default HB model-checking measures that are widely used may not be necessarily good for detecting model deviations. For example, the commonly used posterior predictive p-value (PPP) for checking goodness-of-fit may not be powerful enough to detect non-normality of random effects… because this measure makes 'double use' of the data…." There are methods that take care of this problem, for example, the partial PPP and conditional PPP (Bayarri and Berger, 2000), and the newer CPPP (Hjort, Dahl and Steinbakk, 2006). While these are computationally intensive, this should not be a major limitation in the current era. (See Ma, Sun and Sedransk, 2010, for a recent implementation of CPPP.) I think, though, that there are other considerations that are probably even more important. First, choosing the appropriate test quantities to assess the fit of the currently entertained model is essential. And, this is difficult because an appropriate selection depends on guessing the nature of the aberration of the currently entertained model from one that is closer to the one that generated the observed data. See, for example, Yan and Sedransk (2006, 2007, 2010) who investigated in detail the problem of detecting unknown hierarchical structure (e.g., fitting a model with a single stage when, in actuality, there are two stages). Moreover, is it important to detect relatively small discrepancies from the model currently being entertained? One may be requiring more "power" than is warranted by the intended use of the data. Additionally, tests of goodness-of-fit are problematic, especially in the frequentist paradigm since such tests are constructed to *reject* null hypotheses whereas one would like to accept a postulated model if the data are concordant with it.

Finally, in Sections 4 and 5, Rao has discussed some applications of Bayesian methods to sample survey data. Sedransk (2008), referenced in Rao's paper, describes other areas where the use of Bayesian techniques should be useful, and also points out some limitations.

REFERENCES

BATTESE, G., HARTER, R. and FULLER, W. (1988). Empirical Bayes estimation of finite population means from complex surveys. *J. Amer. Statist. Assoc.* **83** 28–36.

BAYARRI, M. J. and BERGER, J. O. (2000). p values for composite null models (with discussion). *J. Amer. Statist. Assoc.* **95** 1127–1142. MR1804239

HANSEN, M., MADOW, W. and TEPPING, B. (1983). An evaluation of model-dependent and probability sampling inferences in sample surveys. *J. Amer. Statist. Assoc.* **78** 776–793.

HJORT, N. L., DAHL, F. A. and STEINBAKK, G. H. (2006). Post-processing posterior predictive p-values. *J. Amer. Statist. Assoc.* **101** 1157–1174. MR2324154

MA, J., SUN, J. and SEDRANSK, J. (2010). Sample size determination for unordered categorical data. Technical report, Dept. Statistics, Case Western Reserve Univ., Cleveland, OH.

MALEC, D. (2005). Small area estimation from the American Community Survey using a hierarchical logistic model of persons and housing units. *Journal of Official Statistics* **21** 411–432.

MALEC, D., SEDRANSK, J., MORIARITY, C. and LECLERE, F. (1997). Small area inference for binary variables in the National Health Interview Survey. *J. Amer. Statist. Assoc.* **92** 815–826.

RAO, J. N. K. (2003). *Small Area Estimation*. Wiley, Hoboken, NJ. MR1953089

SEDRANSK, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics* **24** 495–506.

YAN, G. and SEDRANSK, J. (2006). Exploring the use of subpopulation membership in Bayesian hierarchical model assessment. *J. Data Sci.* **4** 413–424.

YAN, G. and SEDRANSK, J. (2007). Bayesian diagnostic techniques for detecting hierarchical structure. *Bayesian Anal.* **2** 735–760. MR2361973

YAN, G. and SEDRANSK, J. (2010). A note on Bayesian residuals as a hierarchical model diagnostic technique. *Statist. Papers* **51** 1–10.