Journal of Official Statistics, Vol. 18, No. 2, 2002, pp. 157–161

Discussion
An Outsider's Perspective

William R. Bell[1]

1. Introduction

Graham Kalton gives a nice review of some issues with the design-based and model-based approaches to inference in sample surveys. His perspective benefits from his years of experience with the practice of survey sampling. I label my discussion, ''An Outsider's Perspective,'' because though I have spent over 20 years working for a large survey organization (the U.S. Census Bureau), relatively little of my experience over that time has been directly associated with standard survey practice. For about 15 years my work concentrated on time series analysis and seasonal adjustment, and for the last six or so years I have concentrated on small area estimation. Both are topics that essentially require use of statistical models. I have worked on some projects (e.g., census coverage estimation), however, that focused on construction of design-based survey estimates. From these experiences, and from general exposure to design-based ideas in my 20 plus years at the U.S. Census Bureau, I feel (hope) I have learned something about the design-based approach. In fact, I have often found it interesting to contrast the design-based and model-based approaches with respect to particular issues that have arisen. My model-based perspective has definitely made me something of an outsider in discussions of these issues.

Graham Kalton makes the case for limited use of models in survey practice, citing several areas where use of models seems advisable or even required. I tend to agree with him about the need to use models in these areas, and so my discussion will not attempt to either amplify these points or disagree with some of their specifics. Instead, my discussion will raise some other issues where I see interesting contrasts between the design-based and model-based approaches. With most of these issues the model-based perspective sees something odd if not problematic about the design-based approach. My aim in doing this is to try to raise some interesting questions, rather than to necessarily provide answers, because it has seemed to me that some of these questions arise precisely from looking at things from a model-based perspective, and such questions may not be so obvious to those looking at the world from a design-based perspective.

[1] U.S. Census Bureau, Washington, DC, U.S.A.
Acknowledgment: The author thanks Donald Malec and Tommy Wright for valuable comments on an earlier version of this discussion.

2. Time Series Analysis and Change Estimates

As noted in the Introduction, time series analysis, including seasonal adjustment, essentially requires model-based perspective. In fact, the strict design-based perspective applied to a repeated survey would imply that values over time of the indvidual population units, and hence also of the population characteristic of interest, would be a series of fixed unrelated quantities, in which case time series analysis would be irrelevant. On the other hand, statistical agencies that conduct repeated surveys leave much of the time series analysis up to data users (as they do for other forms of analyses), although seasonal adjustment is a common practice implying some adoption of a model-based perspective. As an aside, to facilitate time series analysis by data users, statistical agencies would do well to make available estimates of sampling error autocorrelations (not just sampling variances) so that users could potentially account for time series properties of the sampling errors in their time series analyses of repeated survey estimates.

A related area that points out a problem with the design-based approach involves tests of statistical significance for change estimates. Smith (1978) pointed out this problem via an example where estimates from a repeated survey marched predominantly downward over time, but where the latest change estimate was not judged statistically significant (from comparison of the latest change estimate to a suitable multiple of its sampling standard error). Conventional design-based survey practice would simply note that the latest estimated change was statistically insignificant, but this hardly seemed a satisfactory answer given the overall downward movement of the series. These problems with the design-based approach are not limited to Smith's interesting example. Consider that for a survey with very little or even no sampling error (foreign trade statistics typically are examples) change estimates would always or nearly always be statistically significant, yet this would not tell us anything very useful. More attention is needed to this issue, and it is worth noting that making vague references to ''trends'' is not enough to address this problem.

3. Outliers

The interesting thing about outliers is that from a strict design-based perspective they do not exist. Under the design-based view the data are observations of fixed unrelated quantities there is no basis for declaring any observation to be an ''outlier.'' (I am taking an ''extreme'' position to make a point here.) In my view, outliers are data points that strongly suggest deviations from *model assumptions*. Hawkins (1985, p. 539) says, ''Note that the key fact that outliers are only in relation to a distributional model. If the model is changed, they may become concordant.'' One could avoid the difficulties with using the term ''outlier'' in a design-based context by referring instead to ''influential observations,'' but then one runs into the problem that observations can certainly be influential without being outliers. Hidiroglou and Srinath (1981) offer a probably better terminology, ''large or extreme observations,'' and discuss how to reduce their effect in a design-based context, though they also refer to such data points as outliers.

The difficulty with talking about outliers in a design-based context does not stop survey statisticians from worrying about them, especially about what to do if outliers occur in the sample (as opposed to merely occurring in the population). A concern I have sometimes heard expressed is that outliers (in the sample) will, ''blow your variances out of the

water.'' (It would be more precise to say they may blow your *variance estimates* out of the water. Whether or not outliers occur in the sample has no effect on the true design-based variances of the estimates.) Hidiroglou and Srinath (1981) express the concern by noting that the presence of extreme observations in the sample will make the estimated population total deviate substantially from the true population total. But these concerns about the properties of any particular samples have always seemed to me to be at odds with a strict design-based approach. Since the concept of outliers seems to me to be inherently model-based, it also seems to me that to deal with outliers there may be advantages to adopting a model-based perspective.

4. The Design-Based Approach Is Less Systematic for Generalizing Methods to Address New Problems

One advantage to the model-based approach is that a model can be modified to address new features discovered about the present problem, or to provide a good starting point for developing models to use in related problems. While design-based estimators can also be generalized and extended, this has always seemed to me to be more difficult than extending models, primarily because design-based estimators do not rest on a foundation of explicit assumptions. If we do not know what assumptions led to the choice of a particular estimator, how do we know how to modify the estimator to produce something suitable for a different but related problem?

5. The Design-Based Approach Hinders Communication Between the Survey World and the Rest of Statistics

Like it or not, the field of statistics, apart from survey sampling, overwhelmingly follows a model-based view. The differences between the respective approaches to inference hinder communication between design-based survey samplers and other statisticians in both directions: other statisticians have difficulty understanding survey problems when they are presented from a design-based perspective, and survey statisticians may have some difficulties understanding and using statistical modeling techniques.

The consequences of this communication problem for survey organizations are not trivial. Research on sample surveys is to some degree limited, I believe, because fully understanding the design-based foundation of sample surveys presents a significant hurdle to academic statisticians who have not specialized in this area. This limits the number of academic statisticians doing research on sample surveys, which in turn limits opportunities for graduate training in sample surveys (fewer sample survey course offerings and fewer opportunities for PhD thesis research on surveys). This situation adversely affects the recruiting and training of statisticians to work for survey organizations. (The Joint Program in Survey Methodology of the University of Maryland, University of Michigan, and Westat seeks to address some of these issues in the U.S.) In the other direction, statisticians who have worked all their careers in survey organizations may have had little experience with statistical modeling since graduate school, and may thus have difficulty using newly developed or even well-established statistical modeling techniques. This is of concern since, as Graham Kalton emphasizes, survey organizations do face some problems that require use of statistical models.

6. Are Data Users Better Served by Design-Based or Model-Based Estimates?

In 1989 in celebration of the sesquicentennial of the American Statistical Association (ASA), the Washington, DC chapter of the ASA sponsored a series of special seminars. I organized one of these seminars. The general topic was economic statistics, and one of my speakers was Charles Schultze of the Brookings Institution, who had been chairman of the U.S. Council of Economic Advisors under President Jimmy Carter. In his talk Dr. Schultze addressed the question, ''What role should policy makers have in setting priorities for economic surveys?'' His short answer to this question was, essentially, ''not very much.'' Instead, to summarize Dr. Schultze, he thought it would be better to design surveys to produce data that would be useful for economic researchers, and then hope that the policymakers would pay attention to the researchers.

One interesting implication of Dr. Schultze's recommendation is that accuracy is not the only consideration appropriate for judging survey estimates! Survey estimates may serve as input data to formal analyses by other researchers. If researchers are to account for the sampling error properties of these estimates (as, ideally, they should), then keeping the statistical properties of these estimates as simple as possible may be more important, for this particular use of the estimates, than making the estimates as accurate as possible. The U.S. state poverty estimates from the SAIPE program provide an example to illustrate this point. These estimates are constructed by applying a Fay-Herriott (1979) type model to direct, design-based state estimates from the Current Population Survey (CPS). The direct state estimates have relatively large sampling variances, but the sampling errors of these estimates can be assumed approximately uncorrelated since the CPS sample is selected independently within states. In contrast, while the model-based SAIPE estimates have improved accuracy, the errors in these estimates are correlated across states in a complicated way as a result of applying the model. This is not of much concern for those who only wish to directly interpret the state poverty estimates. Researchers wishing to model state poverty data, however, may be better off using the direct estimates (in conjunction with other information from the regression variables used in the SAIPE models) rather than attempting to deal with (or ignoring) the more complicated statistical properties of the SAIPE model-based estimates.

What does this say about design-based versus model-based estimates? Perhaps it does not say very much directly. Design-based estimates may tend to be somewhat simpler than model-based estimates, and thus more likely to produce estimates with simple statistical properties that are better suited for use as inputs to further analysis by formal models. It need not always be the case that design-based estimates are simpler, however. But the general points that *(i)* some thought can be given to what type of survey estimate will best meet the needs of particular data users, and *(ii)* accuracy is not the only relevant consideration for evaluating survey estimates, are worth keeping in mind.

7. References

Fay, R. E. and Herriott, R. A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. Journal of the American Statistical Association, 74, 269–277.

Hawkins, D. M. (1985). Outliers. In the Encyclopedia of Statistical Sciences, Vol. 6, eds. Samuel Kotz and Norman L. Johnson. New York: John Wiley and Sons, 539–543.

Hidiroglou, M. A. and Srinath, K. P. (1981). Some Estimators of a Population Total from Simple Random Samples Containing Large Units. Journal of the American Statistical Association, 76, 690–695.

Smith, T.M.F. (1978). Principles and Problems in the Analysis of Repeated Surveys. In Survey Sampling and Measurement, ed. N. K. Namboodiri. New York: Academic Press, 201–216.