DISCUSSION

Keith R. Eberhardt, Ohio State University

1. Introduction. The tremendous success that survey sampling methods have enjoyed for many years at the U.S. Census Bureau and elsewhere evidences the fact that survey sampling practice is a very highly developed field. However, in a discussion on the foundations of the theory of statistical inference in finite populations, it is relevant to ask if that success derives from the application of classical survey sampling theory, or instead if it is largely due to a phenomenon similar to that studied by Relles and Rogers (1977) in their article, "Statisticians are Fairly Robust Estimators of Location." Many writers have questioned whether the formal probability sampling theory of finite population inference adequately explains the wealth of good practice, or whether an "art" of survey sampling practice is required to bridge the gap between that theory and good practice.

A central issue discussed at some length by the authors is the proper role of superpopulation probability models in finite population inference. They plainly reject the notion that models have no relevance to the theory. They accept the idea that it is often useful to represent the variate values in a finite population as realized values of random variables having some joint probability law. This model is said to be useful for evaluating the relative efficiences of various estimator and probability sampling plan combinations. However, they express strong reservations about the explicit use of the model in making the final inference (e.g. confidence interval) from the sample. Rather than basing the inference on the model, they prefer to make inferences which refer only to the probability sampling distribution actually used to select the sample units.

According to my understanding, the main reasons for the authors' reluctance to use inference statements which refer to a model are the following: (a) The model may be wrong. This raises the worries that we all share about what happens when a model is naively "assumed" in order to endow an estimation procedure with good properties (such as very narrow confidence intervals). (b) Mathematically "valid" inferences can be made without reference to any model. Thus, although an estimation procedure might be chosen because it has desirable properties under what is believed to be a realistic model, the final inference statement is supposed to be safe, even if the model is hopelessly wrong, because it makes no reference to the model. Instead, the "validity" of the inference rests only on the probability sampling distribution, which is completely under the control of the sampler.

In short, Hansen, Madow and Tepping maintain that "models have their place--and that place is in the closet, out of sight."

My aim is to argue that it is time to bring models out of the closet. They should be used explicitly and consistently, but of course robustly. The particular model-based approach with which I am most familiar is the prediction approach in which the problem of making inferences about a population total is reduced to that of predicting the total for nonsample units.

Regarding point (a) above, I wish to maintain that reasonably careful applications of the (model-based) prediction approach will automatically avoid use of procedures which are sensitive to failure of the adopted working model. This idea is employed in section 2 where the Illustrative Example is duscussed.

Point (b) has been strongly challenged by many writers during recent years. Examples which come readily to mind include Basu's (1971) Circus Example, Royall's (1975) example of the population consisting of "an ass, an axe, and a box of old horseshoes," and Lahiri's (1968) examples illustrating the logical tangles one can encounter in trying to describe uncertainty in the unique sample obtained by properties of the probability sample distribution. This challenge begins with the observation that the logical mechanism which makes possible inferences from a given sample to the population depends ultimately on the existence of some connection between the sample and non-sample units. It follows that when this connection is lacking (as in Royall's example) or when the estimator used abuses this connection (as in Basu's example) the resulting inference is unsound--even though it is formally "valid" in terms of the probability sampling theory. In section 3, the "surprise outlier" problem mentioned by Hansen, Madow and Tepping is discussed in this light.

2. Comments on the Simulation Study. The illustrative Example illustrates the fact that estimation procedures which have good properties under one model can perform poorly when a different model holds. An important case in point is the fact that, for unbalanced samples, the unweighted ratio estimator ("model 1") can be seriously biased if the regression of y on x is not a straight line through the origin. Royall and Herson's (1973) result that this bias can be eliminated by the use of balanced samples is also confirmed in the Example. (Balanced sampling, which is not relevant to the "model 2" estimator, does not help it.)

The fact that the "model 1" and "model 2" estimators do not perform as well as the other three estimators studied is supposed to provide evidence of the weaknesses of a model based theory of survey sampling. As a first comment, I believe that one gains more insight into the problem by analyzing the five estimators with respect to the model actually used to generate the study population than one does by examining whether or not weights equal to the reciprocals of inclusion probabilities were used. Secondly, it should be emphasized that the "model 1" and "model 2" estimators studied are far from being

the only choices that the prediction approach has to offer for this problem.

A simple alternative is the separate ratio estimator,

$$\bar{y}_{SR} = \Sigma(N_h/N)\bar{X}_h\bar{y}_h/\bar{x}_h,$$

where $\bar{y}_h$ and $\bar{x}_h$ are sample means, and $\bar{X}_h$ is the the population mean for stratum h. This estimator has been studied by Royall and Herson (1973-part II) under various polynomial regression models in which the conditional y-variance is taken to be proportional to x. One property they obtain is that, if well-balanced samples are chosen in each stratum, $\bar{y}_{SR}$ is somewhat more robust than the unweighted ratio estimator ($\bar{y}'''$) with simple (unstratified) balanced sampling when failure of the straight-line-through-the-origin model is contemplated. In addition, they show that $\bar{y}_{SR}$ with stratified balanced samples can be more efficient than $\bar{y}''$ with simple balanced samples. Although the robustness of this relative efficiency result has not been carefully studied, the separate ratio estimator has enough appeal, from a prediction theory point of view, to merit consideration along with the five estimators considered by the authors.

When (as in the Example) the variables y and x are related by a model with $\mathcal{E}(Y) = \alpha + \beta x$ and $\mathrm{Var}(Y) = \sigma^2 x^{3/2}$, it is easy to show that, for stratified balanced samples, the mean squared error of $\bar{y}_{SR}$ is approximately

$$\mathcal{E}(\bar{Y}_{SR} - \bar{Y})^2 = \sigma^2\Sigma(N_h/N)^2(\overline{x^{3/2}})_h/n_h ,$$

where $(\overline{x^{3/2}})_h$ is the sample mean of $x^{3/2}$ in stratum h, and the approximation is due to ignoring finite population correction factors. A similar calculation for the weighted ratio estimator yields, for a given sample,

$$\mathcal{E}(\bar{Y}'' - \bar{Y})^2 = \alpha^2(\bar{X} - \bar{x}_w)^2/\bar{x}_w^2 + \sigma^2(\bar{X}/\bar{x}_w)^2\Sigma(N_h/N)^2(\overline{x^{3/2}})_h/n_h .$$

Since the average value of $(\bar{X}/\bar{x}_w)^2$ in stratified random sampling exceeds unity, these calculations indicate that $\bar{y}_{SR}$ would have compared favorably with the other estimators if it had been included.

The simulation study also confirms that the model-based weighted least squares variance estimators exhibit persistent negative biases under the conditions of the study. This is consistent with the results for the Model 1 estimator reported in Royall and Eberhardt (1975) and Royall and Cumberland (1977).

The variance estimator v_H is mentioned in passing by the authors in connection with the weighted ratio estimator. This variance estimator was originally suggested from prediction theory, and studied empirically, for use with the simple ratio estimator. The available theoretical and empirical results indicate that v_H would produce variance estimates consistently larger than those produced by the Model 1 variance estimator used in the study.

3. Surprise Outliers. The authors present an interesting example in section 3.2 which I believe supports the view that basing inferences on the probability sampling distribution is not sufficient to guarantee sound inferences. For ease of discussion, I will rephrase the example slightly as follows.

A large sample of blocks is selected with probabilities proportional to an available measure of size, z_i (e.g., number of housing units on block i). After selection of the sample, it is discovered during the field work that the size measure used for block #1, z_1, was grossly in error. In fact, the correct size measure, x_1, is about 500 times as large as z_1. The usual estimator of the y-mean is

$$\bar{y}_z = (\bar{Z}/n)\Sigma_s y_i/z_i ,$$

where y_i denotes a total over block i, s denotes the set sample blocks, and $\bar{Z} = \Sigma_1^N z_i/N$. When block #1 belongs to the sample, s, the value of $\bar{y}_z$ will be very large, apparently producing a serious overestimate of $\bar{Y}$. Faced with a sample containing block #1, the sampler may therefore decide to reduce the weight applied to y_1, perhaps by using

$$\bar{y}_x = (\bar{X}/n)\Sigma_s y_i/x_i ,$$

where $x_i = z_i$ for $i \neq 1$ and x_1 is the true size for block #1, (or some other value which the sampler finds appealing). I assume that the original measures of size were obtained and checked with sufficient care so that the sampler is confident that the difference between $\bar{Z}$ and the mean of the true size measures is small.

Consider the properties of $\bar{y}_x$ and $\bar{y}_z$ with respect to the probability sampling distribution actually used, in which blocks are selected with inclusion probabilities proportional to the z_i (ppz). First note that $\bar{y}_z$ is ppz-unbiased, consistent, and its variance can be unbiasedly estimated from the sample. On the other hand, $\bar{y}_x$ is biased, but should have smaller mean squared error than $\bar{y}_z$. The estimator $\bar{y}_x$ may not be consistent in the probability sampling sense, depending on the properties of the infinite sequence of populations to which the present population is considered to belong. For example, if the proportion of units for which $z_i << x_i$ remains bounded away from zero, $\bar{y}_x$ (under ppz sampling) has a bias which does not vanish as $n,N \to \infty$. In terms of formal probability sampling theory, it would appear that mathematically "valid" inferences can be made from $\bar{y}_z$, but inferences based on $\bar{y}_x$ are not theoretically justifiable.

For the unique sample obtained, there are at least two aspects of the situation which are inadequately explained by probability sampling theory. First is the matter of bias. If block #1 is selected into the sample, the estimate

produced by $\bar{y}_z$ is an unacceptable overestimate of $\bar{Y}$ - yet it is "unbiased." If one corrects the estimate to $\bar{y}_x$, he incurs a bias. Furthermore, since the ppz-bias in $\bar{y}_x$ does not depend on the sample obtained, its magnitude (and sign) is the same whether we use it to describe the bias incurred in making a large negative adjustment from $\bar{y}_z$ to $\bar{y}_x$ (if block #1 is in the sample) or a small positive adjustment (if block #1 is not in the sample but the error in z_1 is discovered during field work). Of course the presumed improvement in the ppz mean squared error also has the same magnitude whether the actual difference between $\bar{y}_z$ and $\bar{y}_x$ is small or large for the sample obtained.

As the authors remark, and as intuition suggests, the problem with the estimator $\bar{y}_z$ is ordinarily much less serious if block #1 does not happen to belong to the sample. However, it is hard to see how this idea comes from probability sampling theory. If the probability sampling distribution is relevant to assessing after-sampling uncertainty in the estimate, then $\bar{y}_z$ should have the same logical status whether or not the sample happens to contain block #1.

A second issue is the applicability of the normal approximation. The authors mention the difficulty that "when an outlier occurs in a sample the normal approximation ... may not be acceptable." While this statement almost reads as if the normal approximation should apply to the unique sample obtained, I will discuss it in terms of the probability sampling distribution. Suppose that, owing to a large-enough sample size, etc., the ppz-distribution of $\bar{y}_z$ is well approximated by a normal distribution except for samples containing block #1. If the total probability of all samples containing block #1 is small the disturbance to the sampling distribution, due to samples containing block #1, would not seriously affect the goodness of approximating normal probability calculations. Moreover, as z_1 becomes smaller, the total probability of samples including block #1 diminishes, and the normal approximation improves. Of course, if the sampler has the great misfortune of obtaining a sample containing block #1, the value of $\bar{y}_z$ becomes more extreme as z_1 decreases and, for him, the difficulty apparently becomes worse. Clearly, something more than approximate normality of the probability sampling distribution is needed to make the final confidence statement sound.

Peeking into the closet, I submit that the reason for the sampler's reluctance to use $\bar{y}_z$ when block #1 is in the sample is that the value of the estimate (for this unique sample) is intolerably inconsistent with his understanding of the relationships among the y and x values for sample and nonsample units. If y_i is roughly proportional to x_i throughout the population, then an estimator based on the sample mean of y_i/x_i is appealing. If not (as where x_1 is replaced by z_1) then that estimator loses its appeal. When block #1 is in the sample, the practical value of $\bar{y}_z$ as an estimate of $\bar{Y}$ is not enhanced by the fact that ppz sampling was used to obtain the sample. It is hard to see how one can usefully "avoid ... the necessity of defending an assumed model" by an appeal to the assumption-free probability sampling theory of $\bar{y}_z$.

It seems clear that the kind of reasoning required to produce a useful remedy for the problem that is encountered when block #1 is in the sample uses implicit assumptions about the relationships among the y and x values for different blocks in the population. That is to say, some sort of vague, implicit model is forced out of the closet to rescue the sampler when a problem of this magnitude is encountered.

If we make explicit use of a simple model for the situation, a straight-forward analysis of the problem may begin along the following lines. Consider representing each y_i as the realized value of a random variable Y_i, which has conditional mean βx_i (given the "true" size measure, x_i). Then, with respect to the model, the bias of $\bar{Y}_z$ for a given sample is $\beta[(\bar{Z}/n)\Sigma_s x_i/z_i - \bar{X}]$. In agreement with intuition, the magnitude and sign of this quantity depend strongly on whether $x_i = z_i$ for blocks in the sample. Of course, under the same model, $\bar{Y}_x$ is unbiased for any sample. If another, more complicated, model is thought to be more realistic, a different analysis can incorporate it. The utility of the approach lies in its ability to give formal, explicit expression to whatever insights are available into the structure of the population under study.

4. Summary Remarks. It is agreed that the use of artificial randomization can be a useful tool in designing a sample (e.g. Royall, 1976), but it does not necessarily follow that randomization is sufficient to produce sound inferences--even if the sample size is "large enough," etc. While randomization can be expected to produce samples with good properties, that is, samples for which extraneous factors tend to be balanced out, the soundness of an inference based on a given sample ultimately depends on the characteristics of that sample. Inference from a "bad" sample is not logically strengthened by appealing to the characteristics of the randomized sample selection procedure which produced it.

Similarly, if, for the unique sample obtained, an estimation procedure is grossly inconsistent with the relationship which exists between sample and nonsample units (such as $\bar{y}_z$ if unit #1 is in the sample), any randomization used in the selection of the sample has not helped to make the estimator robust. In this case, a serious failure of the implicit model has been encountered. Keeping the model in the closet is not the solution.

The least squares prediction theory of finite population inference appears to be a promising path to the development of robust estimators and sampling plans which have good properties under a wide range of conditions, and also to the characterization of the range of conditions under which a contemplated estimator/sampling-plan combination can be expected to produce good results.

ADDITIONAL REFERENCES

Lahiri, D. B. (1968), "On the Unique Sample, the Surveyed One," presented at the Symposium on Foundations of Survey Sampling, Chapel Hill, North Carolina, April, 1968.

Relles, Daniel A. and Rogers, William H. (1977), "Statisticians are Fairly Robust Estimators of Location," Journal of the American Statistical Association 72, 107-111.

Royall, Richard M. (1975), "The Likelihood Principle in Finite Population Sampling Theory," 40th Session, International Statistical Institute.

Royall, Richard M. (1976), "Current Advances in Sampling Theory: Implications for Human Observational Studies," American Journal of Epidemiology, 104, 463-473.

DISCUSSION

William G. Cochran

To me, when I was young, the model-based approach to estimation problems in sample surveys was the standard, natural one--that was what I learned at Cambridge. When I got to Rothamsted I used it in field and laboratory research in agriculture, in problems in which there were at most two or three measured response variables and in which it was relatively easy to collect extra data to check that the proposed model seemed to be correct. However, when we began to consider surveys of farm practices that might have over 50 questions, I saw the point of Yates' reliance on randomization and on results calculated over the sample space produced by his randomization method. Construction of over 50 models, some on variables with which I was not at all familiar, did not seem appealing.

When the model is well-behaved, the simplicity of some of its exact small-sample consequences is attractive, and I use them when I feel confident of the model. I have at times wondered, however, if experts in operations with models might not contribute more in the area of observational studies. In this area, regression and ratio adjustments to remove initial biases have been found to be unreliable and vulnerable to attack. Short of abandoning observational studies, about the only positive method of attack on such problems is to try to develop more realistic (and presumably more complex) models and work out their consequences when used in attempts to reduce bias.

I agree with Dr. Godambe that lecture courses on sample surveys fail to attract. This has saddened me. I have been teaching sample surveys ever since I started teaching, and have always had the impression that in a Ph.D. program the sample surveys course was not popular, and somehow stuck out like a sore thumb. At times I have tried in lectures to relate sample survey randomization theory to the techniques taught in the mainstream courses. But if I did too much of this, I felt that I had stopped teaching sample surveys, and was just teaching another course in math. stat. I agree that books like Dr. Sarndal's will help in bridging this gap.