RESPONSE TO DISCUSSANTS

Morris H. Hansen, William G. Madow, and Benjamin J. Tepping

We are grateful to the discussants for their comments and to the editors of the Proceedings for this opportunity to continue the discussion. We only wish it were practicable for the discussion to continue further so that the points at issue could be further clarified, and, hopefully, disappear.

Royall says that we "repeatedly fail to distinguish between the process of finding optimal strategies under simple models and the entirely different act of advocating the use of these strategies in complex real-world problems." Perhaps we have done so because of the rather forceful way in which these "optimal strategies under simple models" have been presented by Royall, Godambe, and others and (it seems to us) the unnecessarily vociferous attacks in those papers on the use of probability-sampling designs for dealing with real-world problems. We strongly agree with Royall's quotation from his response to comments on his 1971 paper. It is relevant to note that the quotation comes from his rejoinder to those who discussed his paper rather than from the paper itself. It is also noteworthy that at about that time Royall and his colleagues advocated and actually implemented a purposive model-dependent design in conformance with prediction theory for a large-scale sample survey in Afghanistan. For us much of the issue centers on what risks are acceptable and whether and how the risks may be reduced.

Thinking about Royall's, Godambe's, and Eberhardt's comments has led us to feel even more strongly than we have expressed in our paper that those who propose the use of robust procedures with superpopulation model designs are coming close to adopting the probability-sampling point of view.

All the balanced samples (Royall 1973) yield the same conditional variance of the optimal estimator, given the sample. But Royall does not suggest that the balanced samples are therefore equivalent in practice even if equivalent under the model. Rather, he suggests that either further stratification or systematic sampling should be used along with balance within strata. Royall is using the model supplemented by judgment because he knows as well as we that for the real world all balanced samples are not equivalent. Thus, the end product would be that, just as we, he would want to arrive at a set of possible samples such that he would be indifferent about which is selected. Clearly, so also would Godambe. The indifference as to which of the set of possible samples is selected is important from a probability-sample design point of view. Apparently Royall and Godambe have a similar concern.

We might add to or subtract from a Royall or Godambe set of possible samples on the basis of knowledge of the population sampled. We would not use a single model (polynomial in Royall's case) to obtain balanced samples when very likely the relationship is other than polynomial. We might prefer to avoid the compromises made in practice where departures from balance are accepted. But whatever the outcome of such discussions it seems clear that nowadays Royall, Godambe and we all arrive at sets of possible samples believed to be more or less equivalent and that the set of possible samples chosen is based on the desire for robustness, a desire that implies uncertainty about models.

Royall would probably not agree with Godambe that the model-dependent estimator should be unbiased, for each possible realization, in the probability-sample sense. Probably he would agree with us that Godambe is too restrictive in insisting on p-unbiasedness. We insist only that each element of the population have positive probability of being selected, and that the estimator be consistent. Among such possible designs the one chosen should be judged to be reasonably efficient taking costs and time considerations into account. Thus, p-unbiasedness becomes an option rather than a limitation for the survey designer.

If we correctly interpret Royall's comment on advocacy, then Royall feels, as we do, that optimality of an estimator under a model might be desirable but is not by itself an adequate basis for advocating the use of the estimator and sampling plan in complex real-world situations. It would seem to follow that he would not, in practice, object to adjoining other estimators that had good properties from the probability-sampling design (or other model) points of view to the set of estimators to be evaluated. While we are not sure that Godambe would agree, we believe that he would. Nothing in his approach excludes the consideration of different classes of superpopulations, yielding different types of estimators and which lead to no unique strategy. As in statistics in general this is the general case; e.g., uniformly most powerful tests are rarely found in practical situations. Then here, too, at the design stage we would, in the complex real world, end up with a set of candidate consistent estimators none of which is "uniformly" better than the others.

When so phrased it seems clear to us that, whenever the models considered are more than trivially different, Royall and Godambe as well as we end up by evaluating candidate estimators. Thus, we feel that they, not we, are placing undesirable limits on the designs from which a choice is made, because they insist on the use of estimators implied by a fairly narrow class of population models.

The remaining questions are how a choice is to be made among possible designs and how estimates are to be evaluated at the analysis stage when the survey and its processing are done.

We feel it unnecessarily restrictive and potentially leading to error in "complex real-world problems" to limit evaluation of survey results of the models (ξ's) used in deriving the sampling plan and estimators, even the "least favorable" among them (the model yielding the largest mean square error). To make the sample large enough so that the mean square error is acceptable even under the least favorable model may still lead to substantial downward biases in the estimates of the mean square errors. Consequently, we think it essential that probability-sampling evaluations should be made.

Again, we feel that Godambe would agree, at least in cases where he wasn't secure about the model. We are not sure what Royall would do, but we find it most unlikely that his conclusion would be that even if he were insecure about the model or models he would nonetheless ignore the probability-sampling design evaluation. We have indicated that for small samples biases due to incorrect models may be small relative to the probability-sampling design variance. We believe that Royall and Godambe would also feel that for large samples it would be poor statistical judgment to accept the risks and consequences of choosing an estimator that may be optimal for a model that is in error. Thus we believe that for large samples their choices would necessarily depend on considerations that include probability-sampling design evaluations. To omit probability-sampling design evaluations from consideration seems irresponsible to us, unless one had a level of knowledge rarely if ever available.

Finally we come to the analysis. Here it is the "unique sample" that has provided the data. Why should not all the conclusions be based on the unique sample and a model connecting it to the balance of the population? At this point, a wide range of models might be used whether or not they had been used earlier in the development of candidate designs and the choice among the candidate designs. Given the original uncertainty about models how shall one select which to use in the analysis and how shall one discuss the outcomes? Should one use a Bayesian approach or a minimax approach or some other analysis? Under these conditions the probability-sampling design measures have a uniqueness comforting not only to us but also to users of the outcomes who have not been involved in the choice of a model. For more complex analysis, various tradeoffs need to be considered. Just as we feel it important not to be a prisoner of a model so also at the stage of analysis of relationships we recognize that insights arising from the data must be used if a probability-sampling design analysis of these relationships is not sufficient.

We now respond to some additional specific points made by the discussants.

1. We agree with Cochran's statements on why "sample surveys are different" and for the reasons given above we feel they will continue to be different. The difference arises not because probability-sampling design excludes the use models but because in inference to a finite population the final estimates ordinarily need not and should not depend on the validity of assumed models.

2. We do not agree with Kempthorne that the use of confidence intervals automatically excludes Bayesian thinking -- there are Bayesian confidence intervals. As usually applied, Bayesian approaches yield a model for $Y_1. ..., Y_N$, the N random variables defined for N population elements, rather than for the quantities about which we have better prior information, namely those we wish to estimate. When the prior is for the latter variables we do have a limiting likelihood depending on the sampling design and the estimators used but not on the sufficient statistic. Also, when a Bayesian approach is used on a model the result is another model and our previous comments hold. Our comments in the paper that for certain parts of design, e.g., faulty or incomplete data, models should be considered, apply equally to Bayesian approaches.

3. Several comments have been made on our use of "consistency." First, let us note that our definition is <u>not</u> that the estimate gives the "true" value when the sample includes all elements of the population. It is a limiting definition, as mentioned in our paper, and in more detail in Hansen, Hurwitz and Madow (1953), Vol. II, pp. 72, 74. Godambe is concerned with practical implication of attempts to use large sample theory whether for consistency or limiting distributions. Since Godambe knows as well as we that limiting results are often used as approximations to finite results his reasons must be deeper, but with respect to consistency we put minimal conditions on the sequence of populations and the sample design for the validity of mean square convergence in probability, namely that, as n_s and N approach infinity,

$$\lim \ E(y_s' - y_N)^2 = \lim \sigma^2_{y'_s} + E(y_s' - Y_N)^2$$

for each realization, where y_s' is the estimator based on the sample s, n_s is some measure of size of sample, and y_N is the population characteristic estimated. (Given the population and the nature of the design, we can define a sequence of possible sample designs and populations that begin with small sizes of each and as they pass through n and N are expected to be large enough for the limiting results to be a good approximation.)

4. The criticisms by Särndal, Royall, and Eberhardt, and implicitly by Godambe, that in the worked-out example we could have used the weighted ratio estimator, seem to us to miss the point of why we gave this example. A weighted estimator based on stratification with the same or different models used in each stratum will also be poor if the models are not sufficiently valid. Indeed, the mean square error of the estimator will then include contributions resulting from the biases within strata due to incorrectness of the models. It doesn't really matter that, given an

example, one can construct a model. Our point is that assumed models will ordinarily be false and that for large samples the penalty is unnecessary.

5. Royall indicates that we have misrepresented the recommendations of prediction theory, in that "prediction theory explicitly warns against" the use of the model 1 estimator when the zero-intercept model fails. The illustration we have given is one in which the designer would likely conclude that the model has not failed. Should he nevertheless reject the use of the model 1 estimator which he would reasonably conclude is optimal or near optimal for the illustrative example represented? If one must be certain that a model holds for a given sample, there appears to be little utility to prediction theory in choosing an estimator.

6. Royall also complains that we "failed to report what prediction models have to say in support of the weighted ratio estimator." Eberhardt makes a similar comment. The weighted estimator they refer to and that is discussed by Royall and Herson in 1973 is a consistent estimator, not model-dependent, and perfectly acceptable in probability sampling. It is presented and evaluated, as such, in Hansen, Hurwitz and Madow (1953). The only difference is that they propose balanced samples within strata. The prediction models that lead to the use of that estimator do not ignore sample design, as recommended in earlier discussions of prediction theory. The more recent concern of Royall and others with stratification to achieve robustness is discussed in our paper, where we say that the difference between probability-sampling and model-dependent approaches has substantially disappeared (totally, if probability sampling with approximately optimum allocation to strata is employed). We would be much surprised if any estimator used in a probability-sampling design were not optimal under some model-dependent designs.

Unfortunately, we cannot agree with Godambe that his approach enables one "to choose from all estimators and all sampling designs" or that Godambe's optimality criterion "provides logically equal status to both the model and the design." As we see it Godambe's assertions do not hold for a fixed model since estimators are required to be p-unbiased. And the importance of the model itself seems to us to be far greater than might be expected from Godambe's statements. In most discussions of robustness with respect to alternative models, the models are close to the kernel model; e.g., the kernel model may specify a linear regression and the others polynomial regressions. Godambe's alternatives in this (1978) paper are similarly close; i.e., the kernel model really dominates the choices.

8. Godambe indicates that mention of a "best linear estimate" in Hansen and Hurwitz (1943) "was unfortunate and particularly puzzling in view of the authors' statement in the present paper that survey statisticians already recognized that there was no best estimator. . ." He adds: "If this recognition had received clear expression earlier indeed much of the ensuing confusion could have been avoided." We believe the point was clearly made in the 1943 paper, as follows:

> "Under these circumstances the 'best linear unbiased estimate' of X from a sample of m clusters turns out to be $M/m \sum_{i}^{m} X_i/N$. However, a smaller mean square error is often obtained by the use of a ratio estimate from the sample such as $\sum_{i}^{m} X_i/\sum_{i}^{m} N_i$. This estimate is excluded by the 'best linear unbiased' criterion because it is nonlinear and biased, although the bias is usually negligible and the estimate is consistent...
>
> "A recent paper by Cochran [1942] gives a number of consistent though biased estimates of $\bar{X}$ . . .
>
> "Both types of biased estimates mentioned above are consistent, and usually have a smaller mean square error than the best linear unbiased estimate for sampling systems in which the sampling units vary in size. Thus, improved sample estimates will be obtained by modifying the 'best linear unbiased estimate' criterion to include estimates that are nonlinear, consistent, but have a smaller mean square error than the best linear unbiased estimate."

9. Godambe proposes that estimators be p-unbiased. The unnecessary stress on using only p-unbiased estimators will often lead to estimators that have poor characteristics in practice. Godambe attempts to avoid this whole question in his 1978 paper by referring to ratio estimators as "nearly unbiased." We prefer the usual procedure of recognizing them as one of the class of biased but consistent estimators.

10. Kempthorne's comment on pivotal estimates is related to Godambe's comment on consistency. As in our remark on consistency it seems to us that the question is whether one can imbed the current population and design in a sequence of populations and designs such that for that sequence the limiting properties hold and the sample size and population size are large enough for the distribution and other properties of the estimator to be approximated by those resulting from the limiting distribution.

11. Godambe refers to a statement of ours concerning optimum probabilities that we had to delete from the full paper to meet the space requirements of the Proceedings. We like much of Godambe's approach in the 1978 paper that he cites, and as reviewed in his discussion, in which he introduces a model-based selection plan and estimator. We only comment here that Godambe's estimator e* is not model-dependent when, as we have both assumed, the units have been selected with probabilities proportionate to $\sqrt{x_i}$. It is a consistent probability-sampling estimator. He refers to it as having the property of "near optimality" under a very general model.

We emphasize, however, e* does not in general meet the conditions of either an unbiased or a "best" estimator and, in fact, is the kind of ad hoc estimator that Godambe objects to in his discussion. Moreover, e* is a special case of the ratio estimator discussed by Hansen and Hurwitz (1943, 1949), and is the ratio of what are generally referred to as Horvitz-Thompson estimators. Such an estimator results from a simple extension of the philosophy of the estimation procedures Neyman used with disproportionate stratified sampling, in which observations are weighted by the reciprocals of the probabilities of selection.

12. Eberhardt challenges the ability to draw inferences to a population through probabilty sampling without assuming a model and refers to Royall's example (1975) of an ass, an axe, and a box of old horseshoes as an illustration. We discussed this illustration in an earlier paper published in the proceedings of a 1977 conference at Chapel Hill (Namboodiri 1978) as follows:

> "For example, if one is trying to estimate the total weight of the elements of a finite population with probability sampling then estimators exist; they do not depend on an assumed model; and they have known properties for large samples. These properties depend on both the population distribution and the survey design, and are valid even if the population consists of axes, asses, and boxes of old horseshoes.
>
> "The sample data may or may not tell much about the weight associated to any individual element not in the sample -- it might tell a great deal if the units were classified by type in the analysis. In any event, it enables confidence intervals to be computed for the total weight and the total weight associated to elements not in the sample, and those confidence intervals can be made as small as desired by appropriate design and size of sample."

13. With respect to outliers we agree with Eberhardt's discussion that an outlier situation can cause problems in probability sampling, as elsewhere. That was what our discussion of this topic indicated. We also indicated that often the problem can be handled without appeal to assumptions, but that in some situations it may be necessary to compromise with probability-sampling principles, and introduce assumptions or apply model-dependent procedures essentially as described by Eberhardt. However, we stress the need to question the validity of and confidence in the inferences when such compromises are made, which Eberhardt seems not to be greatly concerned about. The outlier may represent a group of such units in the population that have (as is often the case in practice) quite different average characteristics from the balance of the population. In such a situation the procedure Eberhardt describes will result in underrepresenting them in the sample estimate, with potentially serious consequent bias. The problem can ordinarily be avoided by appropriate design.

14. We agree with many of Särndal's comments and we commend the book by Cassel, Särndal, and Wretman for its exposition of the relationship between probability-sampling and model-dependent designs. We must point out, however, that survey sampling in practice is not confined to the estimation of finite-population means and totals, nor are probability-sampling principles applicable only to such estimates. To cite one example, an analyst studying a hypothetical model of a causal system may well, as a first step, estimate regression coefficients for a realized finite population. Another example is the estimation of intraclass correlations for various types and sizes of units, to provide guidance in the design of future surveys. In both these examples, and many others, probability-sampling principles should govern the way in which the estimates are made. We believe, as does Särndal, that more work is needed in inference to cause systems from complex samples. We believe that disregarding correlations introduced by complex designs as well as those implicit even if simple random samples are drawn from some realized population can be and have been important sources of improper inference.

15. We are puzzled by Särndal's remark that our choice of estimators to compare is somehow unfair to the model-dependent approach. He states that "the probability-sampling estimators (but not the model-dependent ones) benefit from knowing something about the population, namely, its shape." Actually, the three probability-sampling estimators make use of knowledge of the sample design, which was based on an anticipation that the variance within a stratum would be roughly proportional to the mean value of the concomitant variable x in the stratum. Both the model 1 and model 2 estimators were chosen to be near optimum on the basis of what might reasonably be inferred from an observed sample from it, such as the rather large sample (n = 200) represented in Figure 1. Moreover, the model 2 estimator derives from a prediction model which assumes the correct values of the conditional variances in the superpopulation, although it assumes a zero intercept of the regression line. Our point was that, on the basis of what appears to be a proper assessment of the data either the model 1 or model 2 prediction theory estimators might be regarded as appropriate. For these models the sample design (involving a disproportionate allocation of the sample to the strata) is irrelevant, from the model-dependent point of view.

16. Särndal has raised a question (as did T.M.F. Smith in the paper cited) as to why survey samplers should be different by not using model-based approaches. Our answer is in three parts:

(a) There are many parts of statistics in which the statistician has no alternative to dependence on models. (Also the ability to compare with a "complete count" rarely exists in areas other than sampling. It was not accidental that the quality of public opinion polls improved after the 1936 and 1948 presidential elections.)

(b) An alternative does exist in sampling from finite populations. Less risk is taken, with even moderately large samples, by not using model-dependent methods, as discussed and illustrated in our paper.

(c) We have not suggested that models be avoided; they can be highly useful in guiding survey design. We only suggest that serious risk of misleading inferences can be avoided by using models in ways such that the results are not model-dependent, at least with reasonably large samples, that is, by observing probability-sampling principles.

17. Eberhardt's summarization of our position as being that the "place [of models] is in the closet, out of sight" is, of course, incorrect. If the inferences made are dependent on a model, the model should be stated explicitly. This would be the case, for example, when there are certain types of adjustments for nonresponse in a survey. We could cite other cases in which it is necessary to invoke a model in making inferences. However, as we said in our paper, the need to use model-dependent methods in some phases of survey work does not justify the complete abandonment of probability-sampling methods, with the consequent loss of their substantial advantages.

18. As Cochran and Godambe have mentioned, graduate courses on sample surveys have not been popular. We doubt that this has been because of a lack of generalizing principles or intellectual content in survey-sampling theory, as Godambe suggests. One reason may be the mathematical orientation of graduate students of mathematical statistics (an orientation that we believe desirable), compounded by the fact that in relatively few universities are courses in sample surveys offered by competent teachers with experience in the planning and implementation of sample surveys. Graduate students often find the derivation of the mean square errors associated with realistic complex designs boring. They may also have trouble doing them correctly. Moreover, there is a conflict between their need for obtaining basic theoretical results suitable for a Ph.D. thesis and the acquisition of the ability to recognize suitable sample-survey designs and the skill required to derive the associated sampling properties.

REFERENCE

Hansen, M.H. and Madow, W.G. (1978), "Estimation and Inference from Sample Surveys -- Some Comments on Recent Developments," in Namboodiri, N.K., ed., Survey Sampling and Measurement, New York: Academic Press, Inc. (in press).