An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys

MORRIS H. HANSEN, WILLIAM G. MADOW, and BENJAMIN J. TEPPING*

In this paper we are concerned with inferences from a sample survey to a finite population. We contrast inferences that are dependent on an assumed model with inferences based on the randomization induced by the sample selection plan. Randomization consistency for finite population estimators is defined and adopted as a requirement of probability sampling. A numerical example is examined to illustrate the dangers in the use of model-dependent estimators even when the model is apparently consonant with the sample data. The paper concludes with a summary of principles that we believe should guide the practitioner of sample surveys of finite populations.

KEY WORDS: Randomization inference; Finite population sampling; Consistency; Best estimator; Robustness; Bias.

1. INTRODUCTION

Probability-sampling designs and randomization inferences are widely accepted as the standard approach in sample surveys. However, in recent years there has been a substantial amount of literature challenging the use of probability-sampling designs and inferences based on the induced randomization and suggesting an alternative that conforms more nearly to the model approach often used in traditional statistical inference. These challenges have arisen in what has come to be referred to as discussions of the foundations of survey sampling.

In this article we discuss the problem of inferences concerning characteristics of a finite population. Except as otherwise indicated, we do not here consider either nonresponse or errors of measurement. We assume that with each unit of the population there is associated a well-defined characteristic or vector of characteristics, and that such characteristics are observed without error for each unit in the sample. The mean, sum, ratio, or other summary measures of such characteristics for the population or parts of the population (domains) are to be estimated from characteristics of units in the sample.

Some analysts prefer to base inferences about finite populations on assumed superpopulation models. Among these analysts, some argue that randomization in the selection of a sample is unnecessary and may be undesirable except, perhaps, to avoid accusations of personal bias in sample selection. Other analysts, while accepting a probability-sampling selection plan, prefer inferences that are model-dependent. The assumption of a model may uniquely determine both the optimum sampling plan and the best estimator for that model. However, in model-dependent inferences, no matter how the sampling plan and estimator are obtained, inference is made at least partially on the basis of the model, and the randomization supplied by nature is regarded as sufficient.

In probability-sampling, on the other hand, randomization is introduced in the sample-selection plan and provides the basis for inference. Random selection of the sample is introduced to avoid unnecessary assumptions about the population and the sample. If samples are drawn with design features such as varying sampling fractions or varying probabilities of selection, the probabilities of selection are reflected in the estimators. However, this does not necessarily lead to a unique choice of estimator, especially if auxiliary information is available; alternative estimators may be examined and evaluated. Some advocates of model-dependent inferences have asserted that once the sample is drawn the probabilities of selection are irrelevant, and they regard the assumption of a model and the use of a best estimator under the model as essential. However, while a best estimator may exist for an assumed model, other models consistent with the observed sample may lead to different best estimators, so that a unique best estimator is still not available.

The present authors believe that modeling is an important tool for use in designing probability samples but that, with large samples, models can and should be used within the framework of probability-sampling inference. Thus, design decisions may be guided and evaluated by models, but inferences concerning population characteristics should be made on the basis of the induced randomization, at least when samples are reasonably large.

In this paper we discuss these issues and also give some attention to a few other related issues appearing in the literature. We define some terms before proceeding with the discussion.

* Morris H. Hansen is Senior Vice President, Westat, Inc., 1650 Research Boulevard, Rockville, MD 20850. William G. Madow is a private consultant, 2293 Greer Road, Palo Alto, CA 94303. Benjamin J. Tepping is a private consultant, 401 Apple Grove Road, Silver Spring, MD 20904. This article is dedicated to the memory of William N. Hurwitz. It was originally presented at the 1978 Annual Meeting of the American Statistical Association, at a session dedicated to his memory. The authors thank George T. Duncan, Invited Papers Associate Editor, and the referees for their helpful comments and suggestions.

© Journal of the American Statistical Association
December 1983, Volume 78, Number 384
Invited Papers Section

1.1 Randomization Consistency

Let there be given (a) a sequence of finite populations, U_N, and a sequence of functions, F_N, of the characteristics of the units of U_N, (b) a sequence of probability selection plans, S_{nN}, for selecting $n < N$ units from U_N, and (c) a sequence of estimators, f_{nN}, that are functions of the characteristics of the n units in the sample selected from U_N. We define f_{nN} to be a consistent estimator of F_N if $f_{nN} - F_N$ converges to zero in probability as $n \rightarrow \infty$, $N \rightarrow \infty$.

To apply this definition to particular surveys, it is clearly necessary to define the functions f_{nN} and F_N. For example, f_{nN} may be the sample mean, the ratio of the sample means of two characteristics, or the sample median, and F_N the corresponding function of the units of the population U_N. It is essential to define the sequences of populations U_N, parameters F_N, selection plans S_{nN}, and estimators f_{nN} in such a way that one knows how the design properties change as n and N vary. For example, are the stratum definitions fixed, with the population and sample sizes in each stratum increasing, or does the number of strata increase? In what ways do clusters of elementary units change when the population changes? These questions highlight the difference between this definition of consistency and the definition found in the classical literature on statistical inference. The classical definition assumes that random samples are drawn from a probability distribution, so that it is relatively clear in what manner the functions F_N and f_{nN} change as n and N increase. Those changes must be clearly specified in the context of sampling from a finite population.

A more specific formulation that is fairly general and also reflects these concerns is that of Brewer (1979) for asymptotic unbiasedness. We accept this as a sufficient definition of randomization consistency, provided care is taken in its application, as discussed shortly.

1. The original population of N units is exactly reproduced $K - 1$ times, yielding K populations of N units each.
2. A sample of size n is (hypothetically) selected independently from each of the K populations, using the same sample-selection procedure for each one.
3. The K populations are aggregated to a population of size KN units, and the population parameter is defined to be a specified function F_K of the KN units.
4. The K samples are aggregated to a sample of Kn units, and the estimator of F_K is defined to be a specified function f_K of the Kn units.

By the definition just given, the estimator f_K is said to be consistent if the difference $f_K - F_K$ converges to zero in probability as $K \rightarrow \infty$.

Care must be taken in the application of this definition by appropriate definition of the sampling plan and the functions F_N and f_{nN}. To illustrate, suppose that we wish to estimate the population total Y of some characteristic of the finite population. If the population has been stratified into H strata, the population parameter to be estimated may be written in the form $F = \sum_{h=1}^{H} X_h R_h$, where $X_h = \sum_{i=1}^{N_h} X_{hi}$ is the sum of an auxiliary variable known for every element of the population and $R_h = \sum_{i=1}^{N_h} Y_{hi}/X_h$. The sample selection specifies the selection of a simple random sample of size n_h from stratum h, and the estimator is defined by

$$f = \sum_{h=1}^{H} X_h r_h, \tag{1}$$

where

$$r_h = \sum_{i}^{n_h} y_{hi} \Big/ \sum_{i}^{n_h} x_{hi}. \tag{2}$$

We must now define F_N and f_{nN} and the sampling plan as n and N increase. There are at least two alternatives, as follows: (a) We define

$$F_K = \frac{1}{K} \sum_{k=1}^{K} \sum_{h=1}^{H} X_h R_h, \tag{3}$$

$$f_k = \frac{1}{K} \sum_{k=1}^{K} \sum_{h=1}^{H} X_h r_{hk}', \tag{4}$$

where

$$r_{hk}' = \sum_{i=1}^{n_h} y_{hki} \Big/ \sum_{i=1}^{n_h} x_{hki} \tag{5}$$

is calculated from the kth sample of stratum h. (b) We define

$$F_K = \frac{1}{K} \sum_{k=1}^{K} \sum_{h=1}^{H} X_h R_h, \tag{6}$$

$$f_k = \sum_{h=1}^{H} X_h r_h'', \tag{7}$$

where

$$r_h'' = \sum_{k=1}^{K} \sum_{i=1}^{n_h} y_{hki} \Big/ \sum_{k=1}^{K} \sum_{i=1}^{n_h} x_{hki}. \tag{8}$$

It will be noted that if alternative (a) is adopted, f is not a consistent estimator of F, whereas if alternative (b) is adopted, f is a consistent estimator of F.

Which of the two alternatives is appropriate must depend on a more specific definition of the sample-selection plan. Thus if the sample-selection plan specifies that stratification is to be carried to the point where two units are to be included in the sample from each stratum, then alternative (a) is the appropriate one, and the estimator is not consistent except under special conditions. On the other hand, if the sample-selection plan specifies a fixed set of strata, and that the sample size to be drawn from a stratum increases as n increases, then alternative (b) is appropriate and the estimator is consistent.

One might question the value of the concept of consistency as defined here when the same sample design can be interpreted as either randomization-consistent or randomization-inconsistent. The answer is that consistency is a property of a sequence of estimators, and a single estimator may be a member of many different sequences. Thus consistency does not ensure that biases will be small in relation to the sampling error unless the sample is sufficiently large in those aspects of the sample design that are bias determining. In this particular illustration, which involves a stratum-by-stratum ratio estimator, the bias-determining feature of the design is the size of sample in the individual strata (as discussed in Section 1.4), and the concept of consistency warns that in this case individual-stratum sample sizes must be sufficiently large if negligible biases are to be ensured.

The preceding discussion is illustrative. With other designs similar questions may arise that need to be considered.

1.2 Probability-Sampling Designs

We define a probability-sampling design to consist of (a) a sampling plan such that each member of the population has a known probability greater than zero of inclusion in the sample, and (b) procedures for inference such that for reasonably large samples the correctness of the inferences does not depend on an assumed model. We interpret (b) as requiring estimators that are randomization-consistent. The inferences are then based on the limiting distributions resulting from the induced randomization. For probability-sampling designs the computed confidence intervals, for samples large enough, are valid in the sense that the randomization probability that the confidence intervals contain the value being estimated is equal to or greater than the nominal confidence coefficient, independent of the distribution of the characteristics among the elements of the population from which the sample is drawn. We emphasize that throughout this paper, except as otherwise specified, probability is interpreted in the randomization sense.

We note that the preceding definition of a probability-sampling design is a departure from some earlier definitions. For example, some authors have used the term "sample design" as we use "sampling plan" and "strategy" for the pair "sampling plan and estimators." However, Cochran (1977, p. 9) and Hansen, Hurwitz, and Madow (1953, Vol. II, p. 7) define a sample design to consist of both a sampling plan and an estimator. We include the character of the estimator as part of the definition of a probability-sampling design because the properties that we ascribe to probability sampling depend on both.

In probability-sampling inference, we emphasize the fact that for large enough samples the validity of randomization inferences does not depend on assumptions concerning the distribution of characteristics in the finite population from which the sample is drawn. Nevertheless, in the principal applications of the methodology the general character of a population to be sampled is intensively studied in advance of designing the sample. Thus, either known or approximate information about the target population or other similar populations is used in an effort to increase the efficiency of the sample, that is, to reduce the length of the confidence intervals per unit of cost. Many special features are introduced in the sample-selection and estimation procedures to achieve this goal. An exceedingly common situation is that one has no direct listing of the elementary units of the population, but finds a way to draw a sample of them with specified probabilities by associating those units with some other population's units for which a list *is* available or can be prepared in some form. This may take the form of area sampling, perhaps with multistage cluster sampling. Often extensive use is made of supplementary information for stratification, possibly with approximately optimum probabilities or optimum allocation, and for regression or ratio estimation. Numerous other design features are also used.

1.3 Model-Dependent Designs

A model-dependent design consists of a sampling plan and estimators such that either the sampling plan or the estimators are chosen because they have desirable properties under an assumed model. We define a model-dependent sampling design to consist of (a) a sampling plan that may or may not require randomization in sample selection, (b) estimators that are model-unbiased or model-consistent, and (c) procedures for inference such that the correctness of the inference depends on an assumed model. Model-dependent approaches to survey sampling often take the form of assuming a superpopulation, with the characteristics of the units of the finite population under study being realizations of the assumed superpopulation. The sampling plan need not be a probability-sampling plan and the estimators need not be randomization-consistent. A model-dependent approach may lead to designs and inferences substantially different from those used in probability sampling.

Model-dependent design and inference may have substantial advantages if the model is appropriate. For example, if the assumed model accurately represents the state of nature, useful inferences can be based on quite small samples, at least for certain models. Also, it may then be possible unequivocally to adopt a best sampling plan or a best estimator. This may result in estimates with sampling errors that are smaller than for alternative possible designs or estimators (for the same level of effort). However, if the assumed model does not accurately represent the state of nature, estimates of population parameters may be substantially biased, and statements about the sampling errors of those estimates may be very misleading. In attempts to avoid this problem, one may possibly relax the model, for example by including additional model parameters. However, even the relaxed model still may not represent the state of nature well enough to prevent misleading inferences.

As noted before, models may also be used to produce model-based designs that are not model-dependent. For example, models of the population may suggest useful procedures for selecting the sample or the estimators. This is often done in probability sampling to great advantage. Thus model-based designs do not need to be model-dependent.

The objective of many surveys is to draw inferences about a particular large finite population. If we use methods that would give us good results on the average over the realizations of a model, then under general conditions these methods will give us good results in the randomization sense in large enough samples for every realization except for a set of realizations of low probability. It follows that if the model holds, a best model-dependent estimator will have a mean squared error over replications of the sample equal to or less than the mean squared error of alternative estimators. (See Note 1 in Appendix.)

Similarly, a characteristic observed for a large finite population considered as a realization of a superpopulation provides a basis for inferring the parameters or other properties of the superpopulation and of other realizations from it. Thus we define an estimator y as better than an estimator y' for a large finite population if the former has a smaller mean squared error over all possible samples of a specified design from the given population. Under general conditions, this estimator y will be better than estimator y' for all except a set of realizations of low probability, for sufficiently large samples. This conclusion is the basis for generalizing from the illustration discussed in the following section.

1.4 Role of Consistency in Probability Sampling

A necessary condition for a probability-sampling design, by the definition we have used, is that it include randomization-consistent estimators. The purpose is to help ensure that biases of estimation for samples of finite size are small relative to the standard error. We emphasize, however, that consistency is not a sufficient condition for the bias of f to be small relative to its standard error for a given sample size n. In the illustration used in defining randomization consistency it is seen that a given sample and estimator can be interpreted either as consistent or as inconsistent depending on how the estimator and sampling plan are assumed to change as sample and population sizes increase. But if the available sample in each stratum is small, the bias of the estimator illustrated may be large relative to its standard error, and knowing what would happen with large sample sizes is not helpful. On the other hand, if the available sample is sufficiently large in each stratum, any bias in f will be small relative to its standard error. However, the requirement of a large sample in each stratum restricts the depth of stratification.

The problem with the illustrated estimator arises because small biases stratum-by-stratum may be more or less systematic and thus the bias of the estimator may not decrease sufficiently simply because the total sample is large. One way to avoid the problem is to use unbiased estimators that are based on weighting by the reciprocals of the probabilities of selection. But in a particular problem this may lead to unnecessarily large variances. On the other hand, an alternative ratio estimator commonly used has the form $f = (\sum\sum y_{hi}/\sum\sum x_{hi})\sum X_h$. This estimator is randomization-consistent but not randomization-unbiased. However, its bias does not depend on the sample sizes stratum-by-stratum but only on the size of the total sample over all strata combined.

The preceding discussion illustrates that one may need to go beyond simply achieving randomization consistency. If biased but consistent estimators are used it is important to have large enough samples at any level of sampling that may affect the bias of the estimator used.

Cochran (1977), and Hansen, Hurwitz, and Madow (1953, Vol. I), provide additional discussion of such alternative estimators.

2. AN ILLUSTRATION COMPARING PROBABILITY- AND MODEL-DEPENDENT SAMPLING

We compare probability- and model-dependent sampling approaches for a very simple hypothetical but highly relevant illustration. There is no implication from these results that the particular probability-sampling designs or model-dependent designs compared here are optimal. Such designs can always be improved. What we claim is that a not unreasonable model-dependent design may lead to unsatisfactory inferences that can be avoided by the use of probability-sampling designs.

2.1 A Simple Example

Suppose we wish to survey a sample of a particular type of retail store at the end of a year to estimate total retail sales for the year. Suppose, for simplicity, that a list is available of the stores in the population under consideration, that the same stores exist during the whole year, and that we have information on the approximate size of each store as measured by the number of employees in a recent payroll period. Such distributions are usually highly skewed, with many establishments of small size and relatively fewer establishments as size becomes larger, but with the large stores accounting for a high proportion of the total sales.

A simple probability-sampling design to estimate total sales might then be to (a) classify the establishments into strata based on the prior approximate information on employment size (and perhaps other information); (b) allocate a sample size to each stratum taking account of the principles of optimum allocation; (c) draw a simple random sample of the specified size from each stratum; (d) obtain the information on sales from the sampled establishments; and (e) prepare estimates from the sample. In practice such a procedure yields a larger fraction of the establishments in the sample for the larger employment size-classes, with decreasing sampling fractions for successively smaller size-classes of establishments. Suppose that such a sample is drawn, and that the desired data are collected from the sampled establishments. The sample observations are weighted by the reciprocals of the probabilities of selection so that an estimator of total sales for all establishments might take the form $\hat{Y}_p = (\bar{y}_w/\bar{x}_w)X$, where X is the known total employment for all listed establishments, $\bar{y}_w = (\sum N_h\bar{y}_h)/(\sum N_h)$, $\bar{x}_w = (\sum N_h\bar{x}_h)/(\sum N_h)$, $\bar{y}_h$ is the average sales for the sampled establishments in size class h, $\bar{x}_h$ is the corresponding average employment figure from the sample, N_h is the number of establishments on the list in size class h, and the sums extend over the strata. The $\bar{y}_w$ and $\bar{x}_w$ are thus weighted means of the $\bar{y}_h$ and $\bar{x}_h$.

The analyst who assumes a superpopulation model may be led to a different approach, given the observed sample. He might adopt a simple and commonly used model (e.g., see Royall 1970; Cochran 1977, p. 158) and conclude from prior experience, or from examining a scatter chart of the individual sample observations, or both, that the relationship between sales and employment could be represented approximately by a straight line through the origin, and that the variability of sales around the regression line increases as employment size increases. More specifically, given the employment size of an establishment, he might regard its sales as a random variable whose expected value falls on a regression line that passes through the origin, and whose variance around that line is proportionate to the employment size. These relationships imply, for establishment i, that

$$y_i = \beta x_i + \epsilon_i, \quad \mathscr{E}\epsilon_i = 0,$$
$$\mathscr{E}\epsilon_i^2 = x_i\sigma^2, \quad \text{and} \quad \mathscr{E}\epsilon_i\epsilon_j = 0, \quad \text{for} \quad \imath \neq j. \qquad (9)$$

Note that $\mathscr{E}$ denotes an expectation over realizations for the superpopulation. These equations then constitute the superpopulation model. Thus, the actual sales, y_i, observed for establishment i are assumed to be a realization of a random variable, Y_i, subject to variance $x_i\sigma^2$.

A careful analyst using the model-dependent approach may make a statistical test of the hypothesis that the model describes the population, and proceed with the model if the test does not reject the hypothesis. In our particular example, as discussed later in this section, such tests for a sample of 400 or less have a high probability of failing to reject the hypothesis.

If the model holds, that is, if it really describes the process that created the finite population from which the sample was drawn, the variance of the sample estimate is reduced by disregarding the procedure by which the sample was selected. We need only estimate the regression coefficient, β, from the sample. The best linear unbiased estimate of β, assuming the model, is simply the ratio of the *unweighted* sample means; that is, $\hat{\beta} = \bar{y}_u/\bar{x}_u$, where $\bar{y}_u = \sum y_i/n$ and $\bar{x}_u = \sum x_i/n$. The estimator of total sales for the finite population is $\hat{Y}_M = \hat{\beta}X = (\bar{y}_u/\bar{x}_u)X$. Note that this estimator is similar to the one given by the probability-sampling approach except that it is based on a ratio of unweighted rather than weighted sample means.

Another difference between the two approaches is in the choice of the variances that are used to measure the precision of the estimators. In the case of probability sampling the variance is defined as the squared deviation of the estimate from its expected value, averaged over all possible samples that would be obtained from the finite population under a specified sample design. For the superpopulation approach, the variance is defined over the possible realizations under the model of the population, conditional on the sample units observed.

If the model is acceptable, or sufficiently so, the superpopulation approach may result in substantial simplifications and other advantages, but substantial disadvantages if it is not. We examine such questions for the example described previously.

Suppose the scatter chart for a sample of 200 observations drawn from 10 strata by an approximately optimum probability-sampling plan from a population of 14,000 establishments looks like that shown in Figure 1. Certain characteristics of the finite population from which the sample was drawn and of the observed sample are also summarized in Figure 1. The generation of the population is described later.

Note that $\bar{x}_u$, the unweighted mean of the sampled x's, is considerably higher for the sample than $\bar{X}$, the mean for the population. This results from the sample-selection procedure whereby the sample was drawn to achieve approximately optimum allocation for the probability-sampling approach, with considerably higher sampling fractions for the strata of larger establishments than for the smaller. If the model holds, it follows that the point $(\bar{x}_u, \bar{y}_u)$ will be approximately on the regression line (within the range of sampling variability) and its expected value conditional on the sample will be exactly on the regression line no matter what sample is drawn. The variance

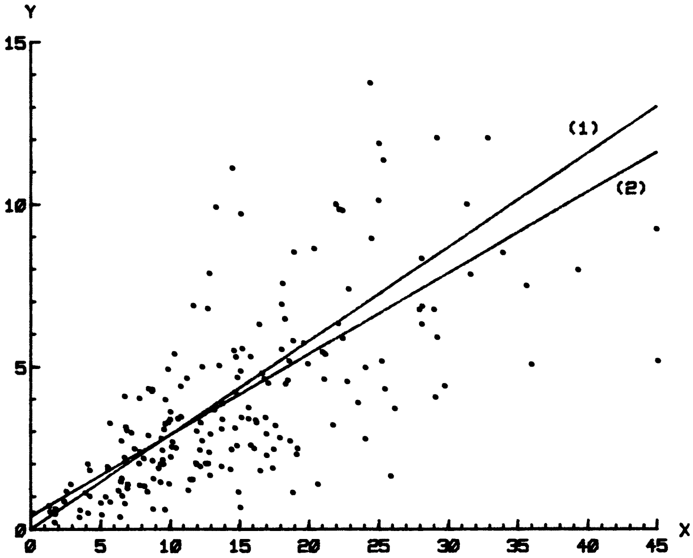

Figure 1. Scatter chart for a sample of 200, drawn from 10 size strata with approximately optimum allocation. The lines shown are (1) the line through the origin and the means of the finite population and (2) the regression line computed for the finite population. $\bar{X}$ = 9.965; $\bar{Y}$ = 2.883; $\bar{x}_w$ = 9.935; $\bar{y}_w$ = 2.794; $\bar{x}_u$ = 14. 644; $\bar{y}_u$ = 3.954; and n = 200.

of $\hat{Y}_M$, over realizations of the population, conditional on the observed sample, would be used under the model-dependent approach to characterize the variability of the estimated total. If the model holds, this variance provides an appropriate measure (a consistent estimator) of the precision of the model-dependent estimator for probability sampling as well as under the model.

The risk in taking the model-dependent approach is that the model may not hold. Suppose the (unknown) regression line for the finite population was in fact as illustrated in Figure 1. Both the line through the origin (1) and the population regression line (2) appear reasonably consistent with the observed sample data. The analyst may then well be willing to assume that the line passing through the origin provides an acceptable model. We show that for this example such an assumption leads to misleading inferences. To the objection that the analyst might introduce an additional parameter or assume some other model that leads to a more robust estimator, thus vitiating our conclusions, we respond that other similar although more complex, examples could be adduced that would lead to the same conclusions that are observed for this simple example. The analysts who use such models emphasize that the model is not assumed to be exactly correct. Our example shows that even moderate departures from an exact representation can lead to large bias even with samples of moderate size.

If the unknown finite population regression line is as shown in Figure 1, then the sales estimated by the model-dependent approach described earlier tend to underestimate sales and to overstate the precision. Thus it provides a confidence interval that includes the value being estimated with a probability that is smaller than the nominal confidence coefficient.

If we know enough about a population, a superpopulation model can give increased precision. But, as we illustrate, the assumption of a model that is not known to hold is at the risk of a possibly large bias arising because the assumed model does not in fact hold. With a probability-sampling approach we can take advantage of the information that led to the assumed model. At the same time we can use estimators whose bias, if any, decreases faster than the sampling error as sample size increases, and which for sufficiently large samples make only a trivial contribution to the mean squared error of the estimate. The probability-sampling estimators may have larger variances, but they avoid the risk of improper inferences because of a bias in the estimator not reflected in the estimated variance.

2.2 The Hypothetical Population

To illustrate some of these points, we have defined a bivariate superpopulation and from it have generated a realized population that we regard as a realistic approximate representation of some populations used as examples in some papers advocating the use of model-dependent designs in drawing inferences about a finite population (see, e.g., Royall and Cumberland 1977).

The realized population was generated as a random sample of 14,000 elements from a bivariate superpopulation in which the variable x has a gamma distribution with density function $f(x) = .04\,x \exp(-x/5)$ and the variable y, conditional on x, has a gamma distribution with density function $g(y;\,x) = (1/b^c\Gamma\,(c))\,y^{c-1}\exp(-y/b)$, where $b = 1.25\,x^{3/2}\,(8 + 5x)^{-1}$ and $c = .04x^{-3/2}\,(8 + 5x)^2$. Hence $\mathscr{E}\,(y \mid x) = .4 + .25x$ and $\text{var}(y \mid x) = .0625\,x^{3/2}$. The variable x is assumed to be known for each element of the realized finite population. The variable y is assumed to be known only for those elements of the realized population that are included in the observed sample.

2.3 A Sampling Plan and the Mean Squared Errors of Five Estimators

The finite population was divided into 10 strata defined by intervals of the variable x, such that the aggregate values of the x variable were approximately the same for each stratum. Then samples of equal size were drawn from each stratum. Thus variable sampling fractions were used, approximately proportionate to the $\bar{x}_h$, following rules of thumb that are sometimes adopted to obtain a rough approximation to optimum allocation of the sample to strata for such populations. Figure 1 displays such a sample, with 20 elements drawn from each stratum.

This sample-selection procedure was independently repeated 1,000 times for each of four sample sizes (2 per stratum, 4 per stratum, 10 per stratum, and 20 per stratum to yield samples of 20, 40, 100, and 200, respectively). For each sample, five estimates of the mean were calculated, along with estimates of their variances. These were

1. the simple unbiased estimator

$$\bar{y}_{(1)} = \frac{1}{N}\sum_{h=1}^{10}\frac{N_h}{n_h}\sum_{i=1}^{n_h} y_{hi} = \bar{y}_w \tag{10}$$

as defined earlier;

2. the regression estimator

$$\bar{y}_{(2)} = \bar{y}_w + b(\bar{X} - \bar{x}_w), \tag{11}$$

where $\bar{x}_w$ is defined analogously to $\bar{y}_w$, $\bar{X}$ is the known finite population mean of x, and b is defined by

$$b = \frac{\sum_{h=1}^{10} N_h{}^2 \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_w)\,(y_{hi} - \bar{y}_w)}{\sum_{h=1}^{10} N_h{}^2 \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_w)^2}\,; \tag{12}$$

3. the ratio estimator

$$\bar{y}_{(3)} = (\bar{y}_w/\bar{x}_w)\bar{X}; \tag{13}$$

4. the best linear unbiased (BLU) estimator under the model

$$y_i = \beta x_i + \epsilon_i;\quad \mathscr{E}\,\epsilon_i = 0;$$

$$\text{var}\,\epsilon_i = \sigma^2 x_i;\quad \text{cov}\,(\epsilon_i,\,\epsilon_j) = 0,\ i \neq j \tag{14}$$

which has the form

$$\bar{y}_{(4)} = \left(\frac{\sum\sum y_{hi}}{n} \Big/ \frac{\sum\sum x_{hi}}{n}\right)\bar{X} = (\bar{y}_u/\bar{x}_u)\bar{X}. \quad (15)$$

(See Note 2 in the Appendix.)

5. the BLU estimator under the model

$$y_i = \beta x_i + \epsilon_i; \quad \mathscr{E}\,\epsilon_i = 0;$$

$$\operatorname{var} \epsilon_i = \sigma^2 x_i^{3/2}; \quad \operatorname{cov}(\epsilon_i, \epsilon_j) = 0,\ i \neq j, \quad (16)$$

which has the form

$$\bar{y}_{(5)} = [(\textstyle\sum\sum y_{hi}/x_{hi}^{1/2})/\sum\sum x_{hi}^{1/2}]\,\bar{X}. \quad (17)$$

For each of the first three estimators, the conventional estimate of the variance was calculated from each sample. In addition, the adjusted estimator of variance (V_H) suggested by Royall and Cumberland (1977) was calculated for the ratio estimator, $\bar{y}_{(3)}$, but this adjusted variance differed trivially from the conventional variance estimate (less than $\frac{1}{2}$ percent for samples of 20, and much smaller differences for larger samples). It has therefore not been included in the summary. For the model-dependent estimators, the variance estimators suggested by Royall and Cumberland (1977) were calculated. (See Note 3 in the Appendix.)

Averages of these variance estimates are shown in Table 1. In addition, for each of the five estimators, the variance, bias, and mean squared error were estimated from the 1,000 replications, and the results also displayed in Table 1.

Figure 1 shows the results of this exercise for a stratified sample of 200 from the realized population, 20 elements being drawn at random from each stratum. We suggested above that it might appear reasonable to consider a line through the origin ás an acceptable approximation for a model. The model-dependent estimators $\bar{y}_{(4)}$ and $\bar{y}_{(5)}$ were chosen on this assumption. They implicitly ignore the sample-selection plan. These two estimators are based on models that assume different conditional variances. These conditional variances are among those that might be regarded as consistent with the observed sample. Actually, the $\bar{y}_{(5)}$ estimator assumes the conditional variances used to generate the population (although not the modest departure of the line from the origin).

If a test of significance were made to decide whether it is reasonable to use a line through the origin, the result would of course depend on the sample size available and the particular sample observed. It would also depend on (but not be sensitive to) the conditional variances assumed in the test. We made such computations (not shown) using the conventional model-based theory. From these computations we conclude that a line through the origin is more likely than not to be accepted as a plausible model for samples of less than about 400, if one made a test before adopting the model, and is more likely to be rejected for larger samples. With a sample of 400 the chances are about even that the estimator $\bar{y}_{(4)}$ would be adopted on the basis of a test of the hypothesis that the intercept of a straight line is zero at the significance level of .05. However, for a sample of 400 the square of the bias of the estimated mean is about four times the variance of the estimated mean.

Table 1. Some Results for the Example, Based on 1,000 Replications

Estimates	Sample Size	$\bar{y}_{(1)}$	$\bar{y}_{(2)}$	$\bar{y}_{(3)}$	$\bar{y}_{(4)}$	$\bar{y}_{(5)}$
Variance[a]	20	.0978	.131	.0920	.0836	.0808
	40	.0599	.0640	.0580	.0510	.0491
	100	.0190	.0184	.0181	.0169	.0157
	200	.0103	.00991	.00993	.00906	.00875
Bias[b]	20	−.002	−.000	−.001	−.122	−.063
	40	−.001	.000	−.002	.131	−.070
	100	−.001	.000	−.001	−.135	−.072
	200	.003	.003	.002	−.130	−.068
MSE[c]	20	.0979	.131	.0921	.0986	.0849
	40	.0599	.0640	.0580	.0682	.0540
	100	.0190	.0184	.0182	.0351	.0209
	200	.0103	.00993	.00995	.0260	.0133
Average of variance estimates[d]	20	.101	.0812	.0952	.0789	.0929
	40	.0515	.0454	.0479	.0394	.0457
	100	.0208	.0190	.0193	.0158	.0185
	200	.0103	.00955	.00965	.00792	.00917
Theoretical variance[e]	20	.1029	.0957	.0958	.0792	.0919
	40	.0514	.0478	.0479	.0396	.0460
	100	.0206	.0191	.0192	.0158	.0184
	200	.0103	.00957	.00958	.00792	.00919

[a] The variance among the 1,000 replicate estimates.

[b] The mean of the 1,000 replicate estimates minus the finite population mean.

[c] The mean square of the 1,000 differences between the replicate estimate and the finite population mean.

[d] Average over 1,000 replicates of variance estimates for each estimator. (See Note 3 in Appendix for variance estimate formulas.)

[e] See Note 3 in Appendix for definitions of theoretical variances.

The first deck of numbers in Table 1 shows the variances of the 1,000 replications for each of the estimators. (See Note 4 in the Appendix.) For $\bar{y}_{(4)}$ the estimated variance is approximately 9 percent smaller than the variance of the ratio estimator $\bar{y}_{(3)}$, and about 12 or 13 percent smaller for $\bar{y}_{(5)}$. If $\sigma_{y\mid x}{}^2$ had been proportionate to x, instead of to $x^{3/2}$, and if the regression line were through the origin, the ratio estimator would have been the optimum for simple random samples, contrary to the results that follow from the use of the Taylor approximations with the terms usually retained. For large samples, the variance of the regression estimator is equal to or less than that of the ratio estimator, and they are about equal in the illustration for samples of 100 or more. Also, the Taylor approximation to the variance for the regression estimator appears to be unsatisfactory for small samples (20 and 40) from the illustrative population.)

Also, the estimated biases of $\bar{y}_{(1)}$, $\bar{y}_{(2)}$, and $\bar{y}_{(3)}$ are trivial for all sample sizes illustrated (the expected bias of $\bar{y}_{(1)}$ is zero). The biases of the model-dependent estimators, estimated from the 1,000 replications, are negative and approximately constant for all the sample sizes, being about 4 percent for $\bar{y}_{(4)}$ and about 2 percent for $\bar{y}_{(5)}$. Thus, even though the model-dependent estimators have moderately smaller variances than the conventional estimators for all the sample sizes, their mean squared errors are greater for the higher sample sizes, and would be much greater for still larger samples. The break-even point appears to occur at a sample size of about 20 for $\bar{y}_{(4)}$, and between 40 and 100 for $\bar{y}_{(5)}$ (which assumes the true conditional variance of y in the superpopulation). The bias of $\bar{y}_{(5)}$ is somewhat smaller than the bias of $\bar{y}_{(4)}$, since the estimator $\bar{y}_{(5)}$ assigns relatively smaller weights to observations in the higher strata than does $\bar{y}_{(4)}$, but the bias is still substantial.

In the case of the model-dependent estimators, the observed bias is clearly the result of the fact that the estimators take no account of the sample design. Obviously, the bias of the model-dependent estimators would vanish or could be made trivial, in this particular illustration, by adopting a model in which the regression line is not required to go through the origin. However, this alternative would not provide a general solution. For example, a model that would be appropriate for a particular population might be nonlinear. In general, more robust estimators and designs could be used in an effort to resolve such problems of model-dependent approaches. However, as discussed later (Sec. 4.4), the problems of model failure will remain unless the designs are so robust as to be nearly model-independent, in which event they are essentially equivalent to probability-sampling designs.

Royall and Cumberland emphasize the bias of the conventional ratio estimator $\bar{y}_{(3)}$ and illustrate it with simple random samples. We see in this illustration that stratification has satisfactorily controlled the biases of the regression and ratio estimators; that is, the estimated biases are negligible even for samples as small as 20.

2.4 The Effect of Proportionate Sampling

Royall and Herson (1973) and Royall and Cumberland (1977) suggest the use of balanced samples, with various means of balancing, to achieve robustness by reducing the bias of the estimator $\bar{y}_{(4)}$. An overall balanced sample (as suggested by Royall and Herson) is very closely accomplished by proportionate stratified sampling with sufficient stratification. In practice stratification is often carried to the point of selecting two units per stratum. For the illustrative population described earlier, only 10 strata are sufficient to achieve reasonably good balance in a proportionate sample of 100. For such a sample the coefficient of variation of the sample mean of x is less than 1.4 percent. Table 2 presents the results of such samples, replicated 1,000 times. Note that the mean squared errors of the three randomization estimators and the model-dependent estimator $\bar{y}_{(4)}$ are all approximately equal, and that the squares of their estimated biases make negligible contributions to their mean squared errors. On the other hand, the square of the estimated bias of the model-dependent estimator $\bar{y}_{(5)}$ constitutes more than 40 percent of the mean squared error. Thus it is seen that the use of $\bar{y}_{(5)}$, even though the model assumed the correct conditional variance and the sample is well balanced with respect to x, failed to attain a reasonable degree of robustness.

One sees by comparing Tables 1 and 2 that for this illustrative population a balanced sample results in a substantially *increased* bias for the estimator $\bar{y}_{(5)}$. Moreover, a requirement that the sample be balanced with respect to the mean restricts the sample design, and does not allow the sometimes substantial gains provided by stratified sampling with approximately optimum allocation. Also, balancing may not be achieved for various domains of interest or for estimating various characteristics of the population. The price paid for overall balancing is an increase in variance as compared to approximately opti-

Table 2. Results of 1,000 Replications of an Approximately Proportionate Stratified Sample[a] of Size 100

Average of Estimates from 1,000 Replications	$\bar{y}_{(1)}$	$\bar{y}_{(2)}$	$\bar{y}_{(3)}$	$\bar{y}_{(4)}$	$\bar{y}_{(5)}$
Variance	.0215	.0215	.0211	.0213	.0199
Bias	−.011	−.009	−.012	−.015	.119
MSE	.0217	.0216	.0213	.0216	.0341
Theoretical variance[b]	.0240	.0229	.0229	.0203	.0260
Average of variance estimates[b] from 1,000 replications	.0236	.0217	.0225	.0199	.0259

[a] The design is referred to as approximately proportionate stratified because of the trivial variation in the selection probabilities resulting from the requirement that the n_h be integral. In the results given here, $\bar{y}_w \cong \bar{y}_u \cong \bar{y}_{(4)}$ and $\bar{x}_w \cong \bar{x}_u$. These approximate equalities would be exact equalities if the samples were exactly proportionate. For proportionate stratified samples of 100 (either exact or approximate) the coefficient of variation of $\bar{x}_u$ is only 1.4 percent and the coefficient of variation of $\bar{x}_w$ is only 1.5 percent.

[b] See Note 3 in Appendix.

mum allocation, as can be seen in comparing the variance for samples of 100 in Table 1 with those in Table 2. The price paid would be far more for many commonly encountered populations that are much more skewed than that used in our illustration. (See Hansen, Hurwitz, and Madow 1953, pp. 139–145 for some examples.)

We recognize that in subsequent papers Royall and others have adopted stratification with differential sampling fractions, and with balancing only within strata, and thereby have moved very close to probability-sampling procedures that do take account of the sample-selection plan in estimation. We have not adopted such procedures for the model-dependent estimators in this illustration, to demonstrate the importance of using such procedures.

We note that with exactly proportionate stratified sampling, the estimators $\bar{y}_{(3)}$ and $\bar{y}_{(4)}$ become identical, although the variance estimators are still different. The two estimated means in Table 2 are not identical only because exact proportionality could not be achieved since the sample sizes n_h are necessarily integers. However, the estimators yield results very close to what they would be if the sample were an exactly proportionate stratified sample.

2.5 Coverage of the Confidence Intervals

As suggested by one of the referees, we examine the coverage properties of the confidence intervals corresponding to the various estimators considered previously.

For each of the 1,000 samples on which Table 1 is based, we summarize in Table 3 the Studentized deviation from the population value $\bar{Y}$, namely $t = (\bar{y} - \bar{Y})/s_{\bar{y}}$, where $\bar{y}$ denotes one of the five estimators and $s_{\bar{y}}$ denotes its estimated standard error. For each of the three probability-sampling estimators $s_{\bar{y}}^2$ is the usual probability-sampling estimator of variance; for $\bar{y}_{(2)}$ and $\bar{y}_{(3)}$ it is based on a Taylor approximation. In the case of $\bar{y}_{(3)}$, it was denoted as V_C by Royall and Cumberland (1977). For $\bar{y}_{(4)}$, we used the error variance (denoted as V_L) as defined in Note 3 in the Appendix, and also estimators V_C and V_H in the style suggested by Royall and Cumberland (1977). For $\bar{y}_{(5)}$, we used estimators analogous to V_L and V_C. (See Note 5 in the Appendix for definitions.) To obtain a 95 percent confidence interval, the practitioner assuming a normal distribution would construct the interval $\bar{y} \pm 1.96\, s_{\bar{y}}$. That is, he or she would expect that the computed quantity t would be less than -1.96 about 2.5 percent of the time and greater than 1.96 about 2.5 percent of the time. Table 3 shows the actual results for the 1,000 samples, for each sample size. For sample sizes greater than 40, the coverage is close to the desired 95 percent for each of the three probability-sampling estimators. For the model-dependent estimators, the coverage is substantially less than 95 percent and becomes worse as the sample size increases, although it is fairly close to 95 percent for $\bar{y}_{(4)}$ using v_C for sample sizes less than 200. Also, the distribution of t is highly skewed for the model-dependent estimators.

Table 3. Coverage of Alternative Confidence Intervals

Estimator and Variance	Sample Size	Proportion of 1000 Replicates		
		$t < -1.96$	$t > 1.96$	$-1.96 < t < 1.96$
$\bar{y}_{(1)}$ v_C	20	.053	.030	.917
	40	.045	.015	.940
	100	.046	.024	.930
	200	.034	.018	.948
$\bar{y}_{(2)}$ v_C	20	.076	.074	.850
	40	.051	.034	.915
	100	.039	.025	.936
	200	.032	.025	.943
$\bar{y}_{(3)}$ v_C	20	.055	.038	.907
	40	.042	.021	.937
	100	.033	.022	.945
	200	.031	.025	.944
$\bar{y}_{(4)}$ v_L	20	.115	.012	.873
	40	.141	.004	.855
	100	.220	.004	.776
	200	.338	.000	.662
$\bar{y}_{(4)}$ v_C	20	.025	.001	.974
	40	.030	.000	.970
	100	.050	.000	.950
	200	.107	.000	.893
$\bar{y}_{(4)}$ v_H	20	.109	.010	.881
	40	.131	.003	.866
	100	.195	.004	.801
	200	.300	.000	.700
$\bar{y}_{(5)}$ v_L	20	.059	.012	.929
	40	.071	.005	.924
	100	.094	.004	.902
	200	.122	.002	.876
$\bar{y}_{(5)}$ v_C	20	.066	.013	.921
	40	.078	.077	.915
	100	.108	.004	.888
	200	.129	.002	.869

For the probability-sampling estimators, the coverage of the confidence intervals is improved at the smaller sample sizes if one assumes a Student t distribution instead of the normal distribution. For samples of 20 (two per stratum) the suggested number of degrees of freedom is 10, and for samples of 40 (four per stratum) it is 30. The coverage of the resulting confidence intervals for $\bar{y}_{(1)}$, $\bar{y}_{(2)}$, and $\bar{y}_{(3)}$ are, respectively, 94.9, 88.4, and 94.3 percent for samples of 20, and 94.9, 92.6, and 94.6 percent for samples of 40.

3. DIFFERING SETTINGS OR GOALS OF SAMPLE SURVEYS, AND CHOICE OF APPROACH

The setting in which a sample survey is taken, and the related goals it is to serve, have important impacts on the choice of design. Surveys come in an enormous variety of circumstances and settings, and with widely differing goals. We now illustrate a few.

Some surveys are one-time and others are repetitive. Some have only a few dozen cases in the sample, or perhaps a few hundred, and others involve thousands or tens of thousands of cases. In some surveys a single characteristic is to be estimated from the survey. In others several estimates are to be made, and in others many hundreds or even many thousands of different estimates will be made from the survey. With hundreds or thousands of statistics to be estimated, often on a tight time schedule, choosing a possibly different estimator for each statistic may be impractical.

From a somewhat different perspective, some surveys are taken to guide specific decisions or actions, may be used by one or a few closely related decision makers (guided in a substantial or a minor way by the survey results), and are not in the public domain or for public use. The decisions may be made by the individual or individuals conducting the survey, or they may be made by others. However, many surveys are taken for widespread use, by government, labor, and industry, for action programs, for legislative guidance or to meet legislative requirements, for research, and other uses.

Sometimes the inferences from a survey relate only to the specific finite population from which the sample was drawn. In other instances the principal goal of the survey is to study causal systems, as a means of understanding and predicting future developments and relationships. Occasionally, but rarely, decisions are based solely on the survey results.

There is an endless variety of circumstances and goals to be served, with costs of a survey ranging from a few hundred dollars, or even a few person-hours, to millions of dollars.

It seems reasonably obvious that the desirable approach will depend on such settings, circumstances, and goals. Surveys such as the Current Population Survey (CPS) of the Bureau of the Census or the Consumer Price Index (CPI) of the Bureau of Labor Statistics cost millions of dollars directly each year, and major public and private decisions and programs are guided by them that may involve hundreds of billions of dollars. In such surveys it seems desirable, to the extent feasible, to avoid estimates or inferences that need to be defended as judgments of the analysts conducting the survey. To the extent that model-dependent methods are used, if the model is not necessarily a good description of reality the analyst may have introduced judgments that are subject to charges, even if ordinarily unfounded, that the analyst has advertently or inadvertently altered or manipulated data to obtain a desired or expected result. Such problems are reduced by staying, to the fullest extent feasible, within the framework of objective instead of judgmental methods. On the other hand, the analyst using, for example, CPS or CPI data, in an attempt to develop a fuller understanding of socioeconomic forces and relationships will often formulate the area of study as some stated model, and will be concerned with estimating the parameters or other characteristics of that model. The analyst then has no choice but to use model-dependent approaches, either explicitly or implicitly, in the effort to understand the information provided by the data that are presented.

4. SOME SURVEY DESIGN IMPLICATIONS

In probability-sampling survey theory and applications, extensive attention has been given to ways in which the various aspects of survey design interact, but this has received relatively little attention in the literature on the foundations of survey sampling in which model-dependent approaches are advocated. Often the goals of a survey design include not only making estimates of various population aggregates and relationships, but also obtaining the results within a limited time schedule and under other administrative restrictions. Other factors that have substantial impact on survey design include the available relevant resources (such as lists or the lack of them covering all or part of the population to be sampled, relevant statistical information, data-processing resources, theory, etc.), alternative methods of data collection, procedures for recruiting, training, and supervision of survey personnel, quality control, and numerous other factors. A general goal is to maximize the relevant information obtained per unit of cost, possibly subject to certain time and other constraints.

4.1 Efficient Survey Design

In the simple case where the goal of a survey is to estimate a single characteristic, survey efficiency may be defined in terms of the mean squared error of the estimator for a fixed total cost of the survey. This definition of survey efficiency assumes similar time and other administrative restraints among surveys whose efficiencies are compared. Commonly, however, a survey is designed to yield several or many statistics, and then efficiency is not so easily defined. It may be defined in terms of some function of individual variances or MSE's per unit of cost. (See, e.g., Cochran 1977, pp. 119–123, for such definitions.) This approach has found relatively little applicability in practice because of the requirements for detailed advance judgments on the uses of the survey results, as well as on unit variances. In such circumstances the relative efficiency of alternatives often becomes a judgment, guided by evaluations for several important statistics.

4.2 Models in Survey Design

In selecting a particular probability-sampling design for a sample survey, the use of models plays an important role. In such uses the models are of various types. Often a model specifies the total cost of the survey as a function of the design, and also the relationships among unit costs and the relevant components of the variance of some statistics, or even assigns values to those components. In other cases, the model specifies, in more or less complete detail, the form of a superpopulation of which the target population is assumed to be a random realization. However, with probability-sampling, the chosen design provides consistent estimators and valid confidence intervals for sufficiently large sample sizes regardless of whether the models used are good descriptions of the real world.

The value of the model in these uses is that it guides the choice of a design, which will be more or less efficient to the degree that the model is a good description. The model, together with probability-sampling theory and data from previous surveys, may guide the definition of strata, the choice of the types and sizes of clusters of the elementary units that are to be the sampling units, the allocation of the sample to strata at the various stages of selection in multistage samples, the forms of the estimators, and other aspects of the design. Ordinarily it is necessary to tailor design decisions to the specific types of situations and populations that are confronted. Rarely is it possible to identify procedures that are uniformly best for all surveys or all populations.

4.3 Role of "Best" Estimators

Godambe (1955) showed that there is no uniformly best linear unbiased estimator for estimating a population total; Godambe defined linearity so that the coefficients of the estimator could depend on the obtained sample and thus included many classes of estimators. Survey statisticians early recognized that there was no best estimator, and in any event had not confined themselves to unbiased or linear estimators (whatever the definition of linearity). Thus, Godambe's results did not create any practical difficulties for the practitioner. The standard practice has been to determine for which types of finite populations a particular design has a smaller mean squared error than an alternative design. In many cases, it was possible and helpful to determine an approximate optimum *within* a specific class of design.

It should be noted that Godambe's proof does not contradict the existence of best linear unbiased estimators as discussed by Neyman (1934) (and others) since Godambe considered a broader class of estimators. It should be noted, also, that insistence on unbiasedness of estimators often results in much larger mean squared errors than necessary. (See, e.g., Hansen and Hurwitz 1943.) Instead, consistency of estimators has been insisted on, with minimization of the mean squared error being the general criterion for choice among designs considered. It seems obvious that for the general class of consistent estimators there are no uniformly best estimators in this sense.

Among advocates of model-dependent approaches, much interest is still expressed in best estimators. The need for assuming models in order to have best estimators is expressed by a number of authors (Godambe 1978, Basu 1971, Royall 1970, Särndal 1978, Cassel, Särndal, and Wretman 1977, and others) and is summarized by Smith (1976) in discussing Neyman's 1934 paper. He states:

> . . .Although a best estimator may be found in each class [of linear estimators] this does not imply that any one of the estimators is best for all classes. This limits the value of Neyman's concept of efficiency. (p. 186).
>
> . . .One consequence of this [Godambe's] nonexistence theorem is that no empirical comparison can ever be conclusive, for in any particular case somebody may be able to construct a better estimator. . .
>
> The problem of a lack of best estimators arises because of the generality of Neyman's formulation of the solution to the inference problem. Inferences are made with respect to the p-distribution [the p distribution refers to the distribution over all possible samples from a probability-sampling design] for *any population Y*, regardless of its structure. But this is too much freedom for a satisfactory theory of inference and no optimum properties can be found for all populations. (p. 187)

Neyman's comment in his fundamental 1934 paper is relevant and expresses our point of view. He says:

> . . .The problem of the choice of estimates has—as far as I can see—mainly a practical importance. If this is not properly solved (granting that the problem of confidence intervals has been solved correctly) the resulting confidence intervals will be unnecessarily broad, but our statements about the values of estimated collective characteristics will still remain correct. Thus I think that the problems of the choice of estimates are rather the technical problems, which, of course, are extremely important from the point of view of practical work. . . (Footnote p. 191)

Added costs of an estimation procedure (in dollars or time) may exceed the gains from reduced variance. In many sample surveys a great many statistics are involved. Timeliness of results is often an important consideration. One may then find it advantageous to adopt estimators based on uniform procedures but with larger variances than those of available alternative estimators tailored to some or all of the statistics. This is a common situation, illustrated by the Current Population Survey of the Bureau of the Census, a complex repetitive survey serving many different purposes. One of the most important purposes is to produce labor-force statistics each month. Thousands of different estimates are published each month, and the estimates are prepared and published within about three weeks after the close of the eight-day period during which the data are collected. Uniform estimation procedures are necessarily applied without, for example, evaluating some available alternatives for each estimate and attempting to choose from among them the consistent estimator with the lowest variance, or worse, a model-dependent estimator. For most of the estimates that are published, improved estimators are known or could be found, often with trivial or minor, but sometimes substantial, improvements. However, the variances computed for the more important statistics provide valid confidence intervals.

4.4 Robustness in Surveys

Robustness is usually understood to mean that inferences made from a sample are insensitive to violations of the assumptions that have been made. In principle, and ordinarily in fact, robustness is achieved in probability-sampling surveys by the use of sampling with known probabilities (i.e., randomization) and consistent estimators, and using a large enough sample that the central limit theorem applies, so that the estimates can be regarded as approximately normally distributed. Also, some estimators are asymptotic approximations that are satisfactory for large enough samples. Much of the work on sample design consists of the study of the population to be surveyed. This is done (a) to adapt the sample design to various special characteristics of the population to improve efficiency, and (b) to identify a sample size such that, at least for the principal items estimated, the estimates have acceptably small variances, and the distributions of the statistics will be approximately normal. Asymptotic approximations are then acceptable.

Occasionally such efforts are not fully successful and a problem occurs when an outlier is observed in a sample, however outliers may be defined. Note that, with probability sampling, the outlier is known to be a member of the population unless, of course, it is simply the result of a mistake (such as a recording error). Sometimes there is the related, but often much less serious, problem that there are cases not in the sample but that would be regarded as outliers if they happened to be selected for the sample. What constitutes an outlier in a sample is not easy to define, but an illustration is a case where a single sampling unit in a sample of, say, 1,000 such units, influences an estimate by, say, 10 percent or more, and also has a substantial impact on the variance estimate. What may be considered an outlier depends, of course, on the size of the sample. An outlier in a sample of 100 might not be considered to be an outlier if the size of the sample were increased to 1,000. Some unknown potential outliers in the population (and not in the sample) are of no concern if they are insufficient to influence the characteristics of the population to be estimated in an important way. If they are large enough to have such important impact, they can be dealt with adequately only through an appropriate choice of the sample design.

When an outlier occurs in a sample, the use of normal approximations may not be acceptable. As indicated earlier, the problem is avoided if the initial efforts at design are successful. If they are not and one or a few outliers occur, several kinds of efforts are ordinarily taken in probability sampling. One is to exclude the outlier from the sample. Another is to reduce its weight in the sample estimate. A preferred procedure is to investigate why the outlier occurred, and to take steps to remove or reduce such problems in the entire population from which the sample was drawn, or in a stratum, or in a larger sample, or in rotating samples, and thus avoid or reduce the impact of an outlier.

An illustration may help. A sample of city blocks may be drawn as first-stage units, a listing of housing units (*hu*'s) in the sampled blocks prepared, and a subsample of *hu*'s drawn from the listed *hu*'s. Varying probabilities of selection of the blocks may be used in an effort to control variation in size, based on prior information on approximate block sizes. This information may be supplemented by special work with building permits so that the presence of new construction is reflected in the measures of size. Alternatively, a separate sample may be drawn from new construction occurring since the date of the prior information. Such an approach ordinarily is quite effective. However, it may still be found that the sample includes a block that is an outlier. This might arise, for example, because the block identification was incorrect. In any event, one sample block expected to have only a few or no housing units may be found to contain a large development, with perhaps 100 housing units in the sample after subsampling from the listing. Weighting the sample observations by the reciprocals of the probabilities of selection might lead to that block accounting for, say, 20 percent of an estimate. In such a case it may be feasible to take corrective action, for example, by cruising past all or a large sample of blocks with small measures of size and increasing the sample size for blocks that are found to be large (or using a post-stratification estimator) so that there will no longer be such outliers. If this or an equivalent procedure is not feasible, it may be desirable to reduce the weight of the sampled outlier. In such an event, the computed sampling error is not readily interpreted, and it is important to qualify carefully any statements of precision of the sample estimate by indicating the potential effects of such action.

If one is applying a model-dependent instead of a probability-sampling approach, the kind of outlier problem just illustrated may or may not be a serious problem. In a model-dependent design an outlier is an observation that deviates considerably from the model. If the outlier in the probability sample comes about as the result of multiplying by the reciprocal of a small probability of selection, then the model-dependent approach, if it ignores the probabilities of selection, may avoid the problem in that the estimate is not influenced importantly by this observation. However, the consequences in terms of potential bias in the model-dependent estimate may not be removed. An outlier in a probability-sampling approach may or may not be an outlier with a model-dependent estimator given the same sample, and vice versa.

The issue of robustness arises especially when model-dependent methods are used for sample selection as well as estimation. Royall (1970) suggested cutoff samples in some situations where a size measure (an x value) is available for each unit in the population. For a sample of n, this calls for selecting the n cases with the largest x values. This approach was suggested for populations having characteristics similar to the example in Section 3, in which the line through the origin seems to be an acceptable fit to the data, and with increasing conditional variances of y as x increases, but at a rate such that the conditional variance of y given x increases less rapidly than x^2. For the particular illustrative population, such a design results in a large bias for quite small samples, and with moderately large samples the bias squared greatly outstrips the variance as sample size increases.

Royall has discussed various procedures to reduce the risk of bias with model-dependent designs, and seems, by striving for robustness, to have moved successively closer to methods used in probability sampling, to the point where relatively little difference exists between some of his more recently recommended approaches and probability sampling. Thus, Royall (1970, 1971) considered the use of cutoff methods in situations where some commonly encountered models seemed to hold. He also recommends (1970) disregarding the sample-selection procedure in estimation from the sample. However, Royall and Herson (1973) gave special attention to robustness and recommended balanced samples, apparently because the lack of robustness of the methods recommended earlier became increasingly obvious. Further, they recommended disproportionate sampling with optimum allocation of samples to strata (taking account of costs), and with balanced sampling within strata. They use model-based approaches to optimize the definition of the size-strata, as would a probability sampler. In these papers they also recommend taking account of the design in estimation, using separate ratio estimates to individual stratum aggregates of an independent variable. Thus the variable selection probabilities are appropriately reflected in the estimator. At this point the difference between probability-sampling and the recommended model-dependent approach has substantially disappeared (totally, if random selections are made within strata).

Having gone this far, final selection of the sample by a probability process within the approximately optimized strata would increase the variance trivially, if at all, as compared with purposive balanced sampling within strata, even if the assumed model holds. It seems highly desirable at this stage to avoid the risk of bias and the necessity of defending an assumed model, and concern if it does not hold. Indeed, if after the steps described, the model assumptions make more than trivial reduction of the length of confidence intervals, there is still the risk of seriously misleading estimates and confidence intervals. Such risk is especially serious when large sample sizes are used to ensure relatively precise and accurate results.

4.5 Role of Varying Probabilities in Sample Selection

There is a substantial literature, which we do not attempt to discuss or summarize here, on the use of varying probabilities of selection of sampling units in the design of a sample.

Varying probabilities of selection may be introduced through stratification with varying sampling fractions in different strata, or through the selection of individual units with varying probabilities. In certain circumstances there is relatively little difference in the results obtained by the two procedures, but in others there are important differences. In the example in Section 2 it would have made relatively little difference whether the units were selected with individually varying probabilities (approximately equivalent to the probabilities resulting from the allocation to the strata), or through the stratification procedure that was followed.

Here we wish to emphasize some of the issues that have been raised in discussions of the foundations of survey sampling by some who may have misunderstood the principal role and uses of varying probabilities.

Several purposes may be served by using varying probabilities in sample selection, in addition to reducing variances of estimators in certain circumstances. A particularly important use is illustrated by the Current Population Survey (CPS) of the Bureau of the Census, and many related types of surveys (often with smaller sample sizes). It was in the design of the CPS that varying probabilities of selection with supporting theory were first introduced by Hansen and Hurwitz (1943).

The primary sampling units (PSU's) in the CPS are individual counties or small groups of adjacent counties within which subsampling is carried out. A resident interviewer can reasonably travel over a primary sampling unit without staying away from home overnight, and typically, in such applications, a workload for one (or two) interviewers is provided in each PSU. We assume, for purposes of simplifying this discussion, that the goal is a general-purpose sample of the population with equal overall probability of including each household in the population. Commonly, in such a design, the primary sampling units are first stratified into very broad size classes, and by type, with stratification of the first-stage units to the point that two units (or sometimes only one) are included in the sample from each stratum. Ordinarily, to achieve near-optimum stratification, the strata are defined so as to be approximately equal in total population. In a particular stratum, suppose there are 10 such PSU's and that the largest one has a measure of size of 50,000 (i.e., it contained approximately 50,000 households at the last Census of Population), the smallest has a measure of size of 10,000, and the total measure of size for all 10 PSU's in the stratum is 250,000. The measures of size are approximately proportionate to the current numbers of households or population in each PSU. Now consider two alternatives for selecting the sample in order to obtain a sample of two PSU's from the stratum, and an overall sampling fraction of .0004, that is, an expected sample from the stratum of approximately 100 households.

Procedure (a). Draw two PSU's from the stratum, such that every PSU in the stratum has the same probability of selection. Then subsample households with the same subsampling fraction (.002) from each selected PSU. With this procedure the number of households in the sampled PSU's in the stratum will be somewhere between about 20 and 100, and on the average will be about 50 per PSU.

Procedure (b). Draw two PSU's into the sample without replacement and with probabilities proportionate to their measures of size, and then subsample at a rate such that the overall probability of selection is the same for each household in the stratum. If P_j is the probability of selecting PSU j, and f_j is the subsampling rate within that PSU, then $f_j = .0004/P_j$ so that $P_j f_j = .0004$, the desired uniform overall probability of selection. The number of households in the sample will be approximately 50 in each of the selected PSU's, assuming only that the measures of size are reasonably good, which in this kind of situation is usually true.

On the average, the total sample size will be the same for both procedures. However, the sample size will vary widely among individual PSU's for procedure (a), but not widely for procedure (b).

A sample size of approximately 50 for each PSU is desirable because (we assume here for illustration) it is the workload that a resident interviewer can conveniently handle in the allotted interviewing period. Under procedure (a) some PSU's have an inadequate workload to keep a resident interviewer interested and trained, and in some others it takes two interviewers to meet the required time schedule, each with an inadequate workload. In procedure (b) the workload in each PSU can be handled by a single interviewer; there is no wasted training of people with low workloads as in procedure (a). The total number of resident interviewers required is smaller for procedure (b).

There is an additional dividend for procedure (b). The between-PSU contribution to the variance is ordinarily smaller for procedure (b) than for procedure (a). (See Hansen and Hurwitz 1943.) In some situations a greater variance reduction might be achieved by using probabilities proportionate to some other quantity—for example, the square root of size, but the considerations illustrated before may be of greater importance and lead to selection with probability proportionate to size. It is this kind of logic that has led to the use of varying probabilities of selection in many surveys in which the situation is approximately that just described.

We have belabored this illustration because of the apparent misunderstanding of some (but by no means all) of those who have advocated model-dependent procedures. Thus, Basu (1971) expresses the opinion that it was because of efforts to make the ratio estimator "look good" (by eliminating the bias of ratio estimators of the form $c\sum y_i/x_i$ where sampling is with probability proportionate to x_i) that "surveyors got mixed up with the idea of unequal probability sampling."

Actually, a principal motivation for introducing sampling with probability proportionate to size of individual units is in multistage sampling, as illustrated, or with additional intermediate stages of sampling, and not to avoid a bias in the ratio estimator. In fact, it does not have this effect, except in cases that rarely occur in practice. Even the selection of samples with probability proportionate to the size of the sample has little or no practical use for this purpose, because, again, the measure of size ordinarily is not the desired denominator of a ratio to be estimated from the sample. With several or a large number of ratios to be estimated from the sample, as is the common case in practice, many or most of the ratios have different denominators that are random variables. At most, ratios with a denominator in common could be made unbiased by sampling with probability proportionate to that measure. This would be unusual in a real survey. Instead of such a simplistic reason for using varying probabilities, there are indeed some good ones, as we have indicated, that have to do with administrative convenience, cost reduction, improved interviewer performance, and variance reduction.

Another illustration of the use of varying probabilities of selection is to increase the effectiveness of stratification. For example, sometimes varying probabilities may be used to replace some stratification by size; this may still get the principal gains that would have been obtained by such stratification. Such an approach makes it feasible to stratify on other variables that are more effective than size in achieving increased homogeneity within strata.

Varying probabilities can be misused, as can other available design features. Unless reasonably good measures are available to determine the varying probabilities, substantial variance increases rather than decreases may result from their use, especially in situations in which a small measure of size has been assigned to a unit for which an extremely large measure should have been used. Numerous analysts have encountered such problems to their regret. However, substantial benefits can often be achieved when varying probabilities are properly used.

Additional consideration of the potential of varying probabilities is beyond the scope of this article, except to refer to some discussion by Godambe.

Godambe (1978) discusses the use of varying probabilities in sample selection along with choice of estimators and alternative superpopulation models. His results are consistent with those on the choice of optimum probabilities as discussed by Hansen and Hurwitz (1949) and Hansen, Hurwitz, and Madow (1953). However, the approach of the latter authors also takes approximate account of costs as well as variance components.

5. SOME ADDITIONAL REMARKS AND COMMENTS

5.1 Analysis of Survey Results

Survey results often are used simply to describe characteristics of a finite population—for example, the number of unemployed at a point or interval of time. Survey results are also used for analyses that relate to the specific finite population, as in testing a hypothesis concerning a difference between ratios for two groups (e.g., the unemployment rates of males and females).

Very often, on the other hand, the analysis is concerned with inferences about a causal system. In this situation the finite population is treated as a realization of that causal system. What a probability-sampling approach can do is provide appropriate information about the available realization, or successive or different realizations, of the causal system, and for this the preceding discussion is directly relevant. However, inferences about the causal system or prediction of future developments may be expressed in terms of estimation of the parameters of a superpopulation or a stochastic process. Then only model-dependent approaches are relevant. Great caution in interpretation is needed, however, as witnessed by many experiences with failures of inference and prediction.

A major issue in inferences to causal systems on the basis of surveys of finite populations has involved two differing points of view on the use of survey results. One view often expressed is that the inferences to a causal system should not depend on the survey design, and that the design of the sample in such instances should be ignored. The analysis is done as if the only source of variation were random sampling from a hypothetical superpopulation. Our view is that the design is relevant, including especially the effects of intraclass correlations from cluster sampling, and perhaps also variable sampling fractions and other aspects of design. Failure to recognize such effects may lead to serious understatement of confidence intervals and overstatements of precision in inferences to the causal system. We believe that misinterpretations are especially likely when design effects due to cluster sampling are not included in the models used for inferences. However, discussion of this topic, beyond a simple mention, is beyond the scope of this paper. Kish and Frankel (1974) pioneered some of the work in this area. More recent work is reported by Fellegi (1980), Holt, Smith, and Winter (1980), Nathan and Holt (1980), and Rao and Scott (1981).

The topic deserves additional work and communication, including the attention and contributions of those concerned especially with the foundations of survey sampling.

5.2 Inferences From Prediction Theory and From Probability Sampling

A criticism of probability sampling by some who advocate model-dependent approaches is that probability-sampling survey theory ignores the fact that to estimate, say, a population total Y is equivalent to predicting the total of Y for the part of the population not in the selected sample. Hence, it is asserted, assumptions must be made that relate those elements in the sample to those not in the sample so that any inference about those not in the sample will be meaningful. If such relations exist, and are known, probability selection is unnecessary except to ensure that no inadvertently biased selection is made. This criticism is related to another that asserts that when the sampling is done all one has is the unique sample, and that the selection process should be ignored. In this view, if there is no model that provides a relationship between the sample and the balance of the population, how that sample was selected cannot create the relationship. However, this argument totally ignores the fact (easily proved mathematically and demonstrated many times empirically) that probability-sampling methods provide a confidence interval for the population characteristic being estimated, and that for large enough samples the confidence interval is valid and short enough to provide as precise statements as desired about the value being estimated. The appropriate consideration concerns the gains and losses from the assumption of a particular superpopulation, with increased risk of misleading results from dependence on the model as sample size increases.

Again, if we know enough about a relationship in the population we should use that knowledge. We can use it in ways such that the validity of the inferences does not depend on the validity of any assumptions. But to make apparently reasonable assumptions just to provide a relationship and perhaps to increase mathematical tractability, and to ignore the type of inference that can be made without such assumptions, takes unnecessary risks and may result in misleading inferences. We prefer to have both approaches available. As few assumptions as possible should be made when the size and nature of the sample permit. As many assumptions as are needed to make sense of the data should be made, but one should not claim too much for the results unless any assumptions made can be supported firmly.

5.3 Conditioning on the Observed Sample

It is an attractive idea to make inferences that are conditional on the observed sample. This is sometimes legitimate in the framework of probability sampling. For example, the estimators $\bar{y}_{(2)}$ and $\bar{y}_{(3)}$ take account of the values of x observed in the sample. So also do the conventional estimators of their variances. However, the latter variances refer to the variability among all samples selected in accordance with the specified sample-selection plan. In some circumstances, the variance can refer to a well-defined subset of all possible samples, which includes the observed sample. In his comment on the paper by Royall and Cumberland (1981), Fuller (1981) has stated it well:

> The use of the randomization population of samples as a basis for inference is *not* restricted to the use of the population of *all* possible samples. Post-stratification is a case in point. The inference can be based on the population of samples that have the same number of observations in the strata that occur in the realized sample. The estimation of proportions [or means] in subdivisions of the population is another situation . . . where proponents of the use of the randomization population of samples recognize the possibility of considering a subset of the population of samples. . . . In the case of simple random sampling and ratio estimation, the population of all samples has traditionally remained the inference population because there is no identifiable subset for which probabilities can be calculated when only the mean of the auxiliary variable is available. If all x values are known at the estimation stage, post-stratification with ratio estimation can be used. That is, when the population of x values is known, it is possible to identify relevant subsets of the population of samples. This emphasizes the importance of clearly stating the nature of the available information when comparing alternative approaches.

5.4 Some Concluding Remarks

When selecting samples and making estimates for a finite population it is always possible to assume a model of the population that implies shorter confidence intervals per unit of cost than can be obtained by probability-sampling methods without those assumptions. However, the probability that the computed interval covers the population value may be substantially less than the cited confidence coefficient. Probability-sampling methods ordinarily can be applied with little or no increase in costs for achieving a confidence interval equal to the computed interval for reasonably carefully applied model-dependent methods. One reason for this is that the randomization mean squared error can be evaluated for various models and compared to the model-dependent mean squared error. One can then make changes in the probability-sampling design to approximate more closely the mean squared error of the model-dependent design while still using probability sampling. If sufficiently important public policy or other issues are involved—as is often the case—it may be well worth paying even a substantial increase in costs to obtain the assurance provided by the probability methods.

On the other hand, in most practical problems the application of probability-sampling theory is essentially assumption-free only if the sample is acceptably large. When surveys use relatively small samples, the samples may be too small for the application of the theory to be essentially assumption-free. Under such conditions, model-dependent inferences may be preferable. Much research needs to be undertaken on the applicability of asymptotic theory to relatively small samples, as for some of the small samples in the illustration of Section 2.

Major advantages stem from the acceptability and face validity of results that can be supported without having to defend assumptions. This may not be important for personal uses of data, but is often vital when sample estimates are for finite populations and results are to be used for important public-policy actions or by opposing factions with different interests when the stakes are high.

Often methods that depend on assumptions based on prior experience work well under stable situations. However, it is especially desirable to use assumption-free methods (or methods involving only mild assumptions) during times of change, especially when one is estimating important measures. For example, in the early part of World War II the Current Population Survey (at that time called the Labor Force Survey) conducted by the U. S. Bureau of the Census yielded estimates of agricultural and nonagricultural employment that were used in guiding wartime manpower policies. Before 1943 the sample was not a probability sample at the final stage of selection and involved an assumed model that, it seemed, would hold reasonably well. However, in 1943 a probability-sampling selection and estimation system was introduced that showed that those assumptions and the estimates based on them were seriously in error for some important estimates. Important changes were made in manpower policy when the new results became available. Presumably the assumptions on which the initial sample was based would have continued to be reasonably satisfactory if there had not been the upheavals in population distribution associated with the war. The new probability-sampling methods were introduced at a time when valid estimates were urgently needed. Their robustness had major advantages.

A similar situation occurred at the end of World War II. One of the authors was in Japan immediately after the war and a friend who had just arrived there indicated that the Labor Force Survey had proved to be unsatisfactory and would soon be abandoned because it failed to show the very substantial increase in unemployment that everyone claimed was occurring with demobilization. He was assured that even if the survey showed surprising and unexpected results, users could believe those results and abandon their preconceptions about what was happening, provided that the sampling errors were small enough, as was the case. It turned out that the principal policy makers did base important policy decisions on the survey results. The validity of those results could be demonstrated by the randomization confidence intervals, and was soon confirmed, without the necessity of defending assumptions. Other similar cases could be cited. Robust methods are equally important for crises such as the energy crisis in 1973, a sharp recession, or other situations in which substantial changes occur in economic or social conditions.

One special caution is needed, to avoid claiming too much even for probability-sampling results. In the preceding discussion we have ignored the existence of measurement or response error. Also we have only briefly mentioned problems of control and treatment of nonresponse. In these areas finite-population concepts do not apply, and we have no choice. Models must be assumed, and to the extent that good models are applied they may aid in improving inferences from a sample survey. However, in our judgment the need for model-dependent methods in some phases of survey work does not justify abandoning the use of probability methods in other important aspects of surveys.

5.5 Guiding Principles

We conclude with a summary of guiding principles that we believe are reasonable inferences from the discussions and illustrations that have been presented:

1. ''Best'' estimators are not possible except by unduly restricting the class of estimators. A model-dependent ''best'' estimator depends on the validity of the model and may yield confidence intervals that are seriously misleading.

2. It is advantageous, as compared with use of ''best'' model-dependent estimators, and sufficient to have a ''good'' estimator based on a reasonably large probability sample that provides a valid confidence interval.

3. Model-dependent designs, including those that use ''robust'' procedures, face the risk of substantially understating the mean squared error, even when the model appears to be satisfactory. Model-dependent approaches in which the model has been tested for the sample data and found not to be inconsistent with those data may still substantially understate the lengths of the confidence intervals.

4. Models are appropriately used to guide and evaluate the design of probability samples, but with large samples the inferences should not depend on the model.

5. Probability-sampling methods, when carefully applied and with reasonably large samples, provide protection against failures of assumed models, and can provide robustness for any sample estimates for which the sample

is large enough, including estimates for domains or subsets that may be identified from the sample and for which the samples are reasonably large.

6. It is our judgment and experience that, with reasonably large samples, sampling plans and estimators based on good probability-sampling methods lose relatively little in efficiency as compared with model-dependent methods even when the models are perfect descriptions of the population.

7. Except in a few special cases such as inferences about a binomial distribution, for which exact results are available, we cannot be assured of valid inferences from small samples, with either probability or model-dependent designs. However, model-dependent methods may have an advantage with quite small samples, for which probability-sampling methods may not be appropriate. Probability samples should be so designed and large enough that statistics of particular interest are approximately normally distributed and such that any needed asymptotic approximations are acceptable. What constitutes a large enough sample depends on effective use (adapted to the population under study) of such devices as stratification, cluster sampling, varying probabilities, use of supplementary information in estimation, and so on. Ordinarily, simple random sampling is *not* appropriate.

8. No general rules can be given for what is a large enough sample. That is a function of the population being sampled and the sample design. Ordinarily, one can reasonably regard samples of less than 25 as small, and of 100 or more (at least with reasonably good probability-sampling designs) as large for populations commonly encountered in survey practice.

These rules relate to the number of first-stage units in a sample or the number of first-stage units contributing nontrivially to an estimate for a domain or a subset of the population.

APPENDIX: NOTES

1. The conclusions stated in this and the following paragraph seem intuitively reasonable. Proofs can be supplied on request.

2. The effect of using this BLU estimator instead of the predictor BLU estimator here is trivial.

3. For $\bar{y}_{(1)}$ the theoretical variances shown are given by the expression

$$s_{\bar{y}}^2 = \frac{1}{N^2} \sum_{h=1}^{10} N_h^2 \left(\frac{N_h - n_h}{N_h - 1}\right) \left(\frac{1}{n_h}\right) \sigma_h^2. \qquad (18)$$

The variance estimate for each sample is the same expression with σ_h^2 replaced by $s_h^2\,(N_h - 1)/N_h$, where

$$s_h^2 = \left(\frac{1}{n_h - 1}\right) \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2. \qquad (19)$$

The theoretical variances shown for $\bar{y}_{(2)}$ and $\bar{y}_{(3)}$ are the conventional Taylor approximations computed for the realized population. The variance estimates are the same approximations computed for each sample. For $\bar{y}_{(4)}$ and $\bar{y}_{(5)}$ the variance estimates used for each sample are the so-called error variances, namely

$$\frac{\bar{X}^2}{n}\left\{\frac{1}{\bar{x}_u\,(n-1)}\,\frac{\sum\sum[y_{hi} - x_{hi}\,\bar{y}_u\,/\,\bar{x}_u]^2}{x_{hi}}\right\} \qquad (20)$$

and

$$\frac{\bar{X}^2}{n}\left\{\frac{1}{\frac{n-1}{n}\sum\sum\sqrt{x_{hi}}} \times \frac{\sum\sum[y_{hi} - x_{hi}\sum\sum y_{hi}\,/\,\sqrt{x_{hi}}]^2}{x_{hi}^{3/2}}\right\} \qquad (21)$$

respectively, where n is the total sample size and $\bar{x}_u$, $\bar{y}_u$ are the sample means defined previously. Note that for $\bar{y}_{(5)}$ the computed error variance is the mean square difference between the estimate and $\beta\bar{X}$ rather than between the estimate and the random variable $\bar{Y}$. There is some question as to which is appropriate, but the difference between them is small. The theoretical variances shown for $\bar{y}_{(4)}$ and $\bar{y}_{(5)}$ used the same formulas but with the sample values in the expressions replaced by their mathematical expectations conditional on the realized population, for a sample of the given design.

4. The samples of size 40 are atypical in that their variances are larger than expected for all estimators. However, the relationships among the estimators are similar to those for the other sample sizes. Also, for this population and sample selection plan, the regression estimator and the ratio estimator show little difference, except that the ratio estimator has a smaller variance than the regression estimator for small samples.

5. We define, for $\bar{y}_{(4)}$,

$$V_C = \frac{1}{n(n-1)}\sum\sum[y_{hi} - x_{hi}\bar{y}_s/\bar{x}_s]^2$$

and

$$V_H = V_C\,\frac{(\bar{X}/\bar{x}_s)^2}{1 - \sum\sum(x_{hi} - \bar{x}_s)^2/n(n-1)\bar{x}_s^{\,2}}\,.$$

Similarly, for $\bar{y}_{(5)}$,

$$V_C =$$

$$\frac{1}{n(n-1)}\sum\sum[y_{hi} - x_{hi}\,(\sum\sum y_{hi}/\sqrt{x_{hi}})/\sum\sum\sqrt{x_{hi}}]^2.$$

[*Received January 1979. Revised January 1983.*]

REFERENCES

BASU, D. (1971), "An Essay on the Logical Foundations of Survey Sampling, Part One," in *Foundations of Statistical Inference*, eds. V.P. Godambe et. al., Toronto: Holt, Rinehart and Winston, 203–233.

BREWER, K.R.W. (1979), "A Class of Robust Sampling Designs for Large-Scale Surveys," *Journal of the American Statistical Association*, 74, 911–915.

CASSEL, C.-M., SARNDAL, C.-E., and WRETMAN, J.H. (1977), *Foundations of Inference in Survey Sampling*, New York, John Wiley.

COCHRAN; W.G. (1977), *Sampling Techniques* (3rd ed.), New York: John Wiley.

FELLEGI, I.P. (1980), "Approximate Tests of Independence and Goodness of Fit Based on Stratified Multistage Samples," *Journal of the American Statistical Association*, 75, 261–268.

FULLER, W.A. (1981), "Comment" on "Ratio Estimator and Estimators of Its Variance," *Journal of the American Statistical Association*, 66, 78–80.

GODAMBE, V.P. (1955), "A Unified Theory of Sampling From Finite Populations," *Journal of the Royal Statistical Society*, Ser. B, 17, 369–278.

——— (1966), "A New Approach to Sampling from Finite Populations—I, II," *Journal of the Royal Statistical Society*, Ser. B, 28, 310–328.

——— (1978), "Estimation in Survey Sampling: Robustness and Optimality," unpublished manuscript.

HANSEN, M.H., and HURWITZ, W.N. (1943), "On the Theory of Sampling From Finite Populations," *Annals of Mathematical Statistics*, 14, 332–362.

——— (1949), "On the Determination of Optimum Probabilities in Sampling," *Annals of Mathematical Statistics*, 20, 426–432.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953), *Sample Survey Methods and Theory*, Vols. I and II, New York: John Wiley.

HOLT, D., SMITH, T.M.F., and WINTER, P.D. (1980), "Regression Analysis of Data from Complex Surveys," *Journal of the Royal Statistical Society*, Ser. A, 143, 474–487.

KISH, L., and FRANKEL, M.R. (1974), "Inference for Complex Samples," *Journal of the Royal Statistical Society*, Ser. B, 36, 1–37.

NATHAN, G., and HOLT, D. (1980), "The Effect of Survey Design on Regression Analysis," *Journal of the Royal Statistical Society*, Ser. B, 42, 377–386.

NEYMAN, J. (1934), "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection," *Journal of the Royal Statistical Society*, 109, 558–606.

RAO, J.N.K., and SCOTT, A.J. (1981), "The Analysis of Categorical Data from Complex Sample Surveys: Chi-squared Tests for Goodness of Fit and Independence in Two-way Tables," *Journal of the American Statistical Association*, 76, 221–230.

ROYALL, R.M. (1970), "On Finite Population Sampling Theory under Certain Linear Regression Models," *Biometrika*, 57, 377–387.

——— (1971), "Linear Regression Models in Finite Population Sampling Theory," in *Foundations of Statistical Inference*, eds. V.P. Godambe et al., Holt, Rinehart & Winston of Canada, Ltd., 259–279.

ROYALL, R.M., and CUMBERLAND, W.G. (1977), "An Empirical Study of Prediction Theory in Finite Population Sampling I: Simple Random Sampling and the Ratio Estimator," presented at the Symposium on Survey Sampling, Chapel Hill, North Carolina, April 1977.

ROYALL, R.M., and CUMBERLAND, W.G. (1981), "An Empirical Study of the Ratio Estimator and Estimators of the Variance," *Journal of the American Statistical Association*, 76, 66–77.

ROYALL, R.M., and EBERHARDT, K.R. (1975), "Variance Estimates for the Ratio Estimator," *Sankhya*, C, 37, 43–52.

ROYALL, R.M., and HERSON, J. (1973), "Robust Estimation in Finite Populations," *Journal of the American Statistical Association*, 68, 880–893.

SÄRNDAL, C.-E. (1978), "Design-Based and Model-Based Inference in Survey Sampling," *Scandinavian Journal of Statistics*, 5, 27–52.

SMITH, T.M.F. (1976), "The Foundations of Survey Sampling: A Review," *Journal of the Royal Statistical Society*, Ser. A, 139, 183–204.