ON INFERENCE AND ESTIMATION FROM SAMPLE SURVEYS [1/]

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

Probability sampling has come to be 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 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 foundations of survey sampling.

A probability-sampling design consists of a sampling plan, or procedure for selecting the sample, and estimators such that (a) each member of the population has a known probability, greater than zero, of inclusion in the sample, and (b) the estimators from the survey are consistent in the sense that the estimator converges in probability to the population characteristic being estimated as both the sample and population sizes increase. For probability-sample designs, confidence intervals can be computed which, for samples large enough, will be valid in the sense that the probability that the confidence intervals contain the value being estimated is equal to or greater than the nominal confidence coefficients, independent of the distribution of the population from which the sample is drawn.

A model-dependent design consists of a sampling plan and estimators for which either the plan or the estimators, or both, are chosen because they have desirable properties under an assumed model, and for which the validity of inferences about the population depends on the degree to which the population conforms to the assumed model. Applications of model-based approaches to survey sampling are often in the form of assumed superpopulation models in which the finite population under study is assumed to be a random realization of the assumed superpopulation. When such a model is assumed it may lead to consequences that substantially alter the approaches to design and inferences from sample surveys, as compared to probability sampling. When an approach is used that is sensitive to the validity of an assumed model, we refer to it as model-dependent. Models are involved, also, explicitly or implicitly, in applications of probability sampling. However, since for large samples inferences do not depend on the validity of an assumed model, we distinguish probability sampling from model-dependent sampling.

We compare the two approaches for a very simple illustration. 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 there are no changes in stores during the year, and that we have information on the approximate size of each store as measured by number of employees in a recent payroll period. Such distributions tend to be highly skewed, with many small establishments, and relatively fewer as size becomes larger, but with the large stores accounting for a high proportion of the total sales.

A simple probability-sampling procedure to estimate total sales might then be to (a) divide the establishments into strata based on the prior approximate information on employment size (and perhaps other information); (b) draw a sample from each stratum, taking account of the principles of optimum allocation; (c) obtain the information on sales from the sampled establishments; and (d) prepare estimates from the sample. In practice such a procedure will yield a higher 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 such a sample is drawn, and that the desired data are collected from the sampled establishments. Then, with probability sampling, the establishments will be weighted by the reciprocals of the probabilities of selection so that the estimator of total sales for all establishments might take the form $\hat{Y}_p = (\bar{y}_w/\bar{x}_w)X$ where $\bar{y}_w = (\Sigma N_h \bar{y}_h)/(\Sigma N_h)$, $\bar{x}_w = (\Sigma N_h \bar{x}_h)/(\Sigma N_h)$, X is the known total employment for all listed establishments, $\bar{y}_h$ 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 are over M strata. The $\bar{y}_w$ and $\bar{x}_w$ are thus weighted means of the $\bar{y}_h$ and $\bar{x}_h$.

The assumption of a superpopulation model may lead to a different approach to estimation, given the observed sample. We might conclude from prior experience, or observe from a scatter chart of the individual sample returns, 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, its sales might be regarded as a random variable whose expected value falls on a regression line which passes through the origin, and with variance around that line proportionate to the expected value, and thus also proportionate to the employment size. These relationships imply, for establishment i, that

$$Y_i = \beta x_i + \epsilon_i; \mathcal{E}\epsilon_i = \emptyset; \sigma_i^2 = \mathcal{E}\epsilon_i^2 = \sigma^2 x_i$$

These equations then represent the superpopulation model. The actual sales, y_i, observed for establishment i is assumed to be a realization of a random variable, Y_i, subject to variance σ_i^2.

If this model holds, 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 least squares estimate of β is simply the ratio of the unweighted sample means, i.e., $\hat{\beta} = \bar{y}_u/\bar{x}_u$, where $\bar{y}_u = \Sigma y_i/n$ and $\bar{x}_u = \Sigma x_i/n$. Then the estimator of total sales for the listed population is $\hat{Y}_M = \hat{\beta}X = (\bar{y}_u/\bar{x}_u)X$. Note that this estimate is similar to the one from the probability-sampling approach except that it is based on a ratio of unweighted sample means.

Another difference between the two approaches is in the choice of the variances that are used to measure the precision of the two estimates. In the case of probability sampling the variance is defined as the squared error of the estimate 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 as conditional on the sample actually observed, and the variance is over the possible realizations for the observed sample. If the model holds, the variance of the model-dependent estimator over possible realizations for the fixed sample will be smaller than that of the probability-sampling estimator.

The use of the variances over realizations for a fixed sample results in substantial simplifications and other advantages, if the model is acceptable, or sufficiently so, and substantial disadvantages if it is not. We examine these implications for the illustrative example described above.

Suppose the scatter chart for a sample of 200 observations drawn as developed above looks like that shown in Figure 1. Certain characteristics of the population and of the observed sample are also summarized in Figure 1.

Notice 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 indicated earlier whereby the sample was drawn to achieve approximately optimum allocation for the probability-sampling approach, with considerably higher sampling fractions for the 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 will be exactly on the regression line no matter what sample is drawn. The variance of $\hat{Y}_M$, conditional on the observed sample, would be used under the model-dependent approach to characterize the variability of the estimated average. If the model is valid it is an appropriate measure of the variance of the estimate.

The risk in taking the model-dependent approach is that the model may not hold. Suppose the (unknown) regression line for the population was in fact as illustrated in Figure 1. Both the line through the origin (1) and the assumed true regression line (2) appear reasonably consistent with the observed sample data. However, if the unknown true regression is as shown, then the sales estimated from the model-dependent approach will tend to underestimate

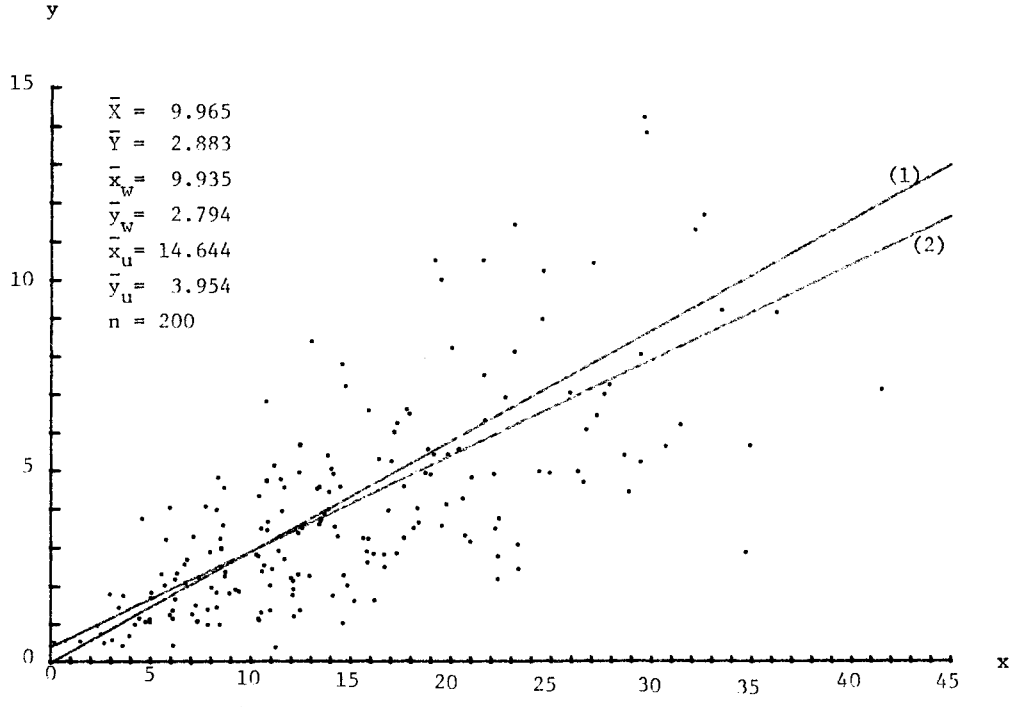

Fig. 1. Scatter chart for a sample of 200, drawn with approximately optimum allocation to ten size strata based on x. The lines shown are (1) the line through the origin and the population means and (2) the population regression line.

sales, and to overstate the precision (i.e., underestimate the width of the confidence interval of the estimate). It will provide a confidence interval that will include the value being estimated with a probability that is smaller than the nominal confidence coefficient. On the other hand, if the true regression is indeed a line through the origin, and the conditional variances around the regression line are as assumed, the model-dependent approach provides an appropriate estimated variance and computed confidence interval.

The probability-sampling estimator will have a higher variance, and a longer confidence interval, than the model-dependent estimator. However, probability sampling protects against systematic errors due to model failure because the variance of an estimator is computed over all samples possible according to the design. With sufficiently large samples probability sampling will provide a confidence interval that will include the value being estimated with a probability equal to or greater than the confidence coefficient, independent of the population distribution or model assumptions.

2. An Illustrative Example

To illustrate some of these points, we have assumed a bivariate superpopulation and have generated a realized population from it that we regard as a realistic approximate representation of some populations that have served as examples in some papers that advocate the use of model-dependent designs in drawing inferences about a finite population (see, for example, Royall and Cumberland [1977]).

The realized population was generated as a simple random sample of 14,000 elements from the superpopulation in which the variable x has a Gamma distribution. The x-variable is assumed known and available for the realized population.

The variable y was also generated for the 14,000 members of the realized x-population, from a superpopulation with a conditional distribution which is also approximately a Gamma distribution such that

$$\mathcal{E}(Y|x) = .4 + .25\,x, \quad \mathrm{Var}\,(Y|x) = \frac{1}{16}x^{3/2}.$$

The y-values are presumed unknown to the analyst except as observed for a sample.

We chose to generate a hypothetical population for an illustration, instead of using an available actual population or sample, because it would then be feasible to know the superpopulation that generated the realization, and the characteristics of a realized population, as well as of repeated samples of specified design from that population.

The population was divided into ten 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.

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

(1) the simple unbiased estimator

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

(2) the regression estimator

$$\bar{y}' = \bar{y}_w + b(\bar{X} - \bar{x}_w)$$

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

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

(3) the ratio estimator

$$\bar{y}'' = (\bar{y}_w/\bar{x}_w)\,\bar{X}$$

(4) the best linear unbiased estimator under the model

$$\mathcal{E}Y_i = \beta x_i, \quad \mathrm{Var}\,Y_i = \sigma^2 x_i, \quad \mathrm{Cov}\,(Y_i, Y_j) = \emptyset, \; i \neq j$$

which has the form

$$y''' = \left(\frac{\Sigma\Sigma y_{hi}}{n}\Big/\frac{\Sigma\Sigma y_{hi}}{m}\right)\bar{X} = (\bar{y}_u/\bar{x}_u)\,\bar{X}.$$

We shall refer to this as model 1.

(5) the best linear unbiased estimator under the model

$$\mathcal{E}Y_i = \beta x_i, \quad \mathrm{Var}\,Y_i = \sigma^2 x_i^{3/2}, \quad \mathrm{Cov}\,(Y_i, Y_j) = \emptyset, \; i \neq j$$

which has the form 2/

$$\bar{y}^{iv} = [(\Sigma\Sigma y_{hi}/x_{hi}^{1/2})/\Sigma\Sigma x_{hi}^{1/2}]\bar{X}.$$

We shall refer to this as model 2.

For each of the first three estimators, the conventional estimator 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, but differed trivially (less than 1/2 percent for samples of 20, and much smaller differences for larger samples) and has not been included in the summary. For the model-based estimators, the variance estimators suggested by Royall and Cumberland (1977) were calculated.

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

Table 1. Results of 1,000 replications for sample sizes 20, 40, 100 and 200, for the illustrative example.

	$\bar{y}_w$	$\bar{y}'$	$\bar{y}''$	$\bar{y}'''$	$\bar{y}^{iv}$
Estimates from 1,000 replications					
Variance					
Samples of 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					
Samples of 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					
Samples of 20	.0979	.131	.0921	.0986	.0849
40	.0599	.0640	.0580	.0682	.0540
100	.0190	.0184	.0182	.0351	.0209
200	.0103	.00903	.00995	.0260	.0133
Theoretical variance*					
Samples of 20	.1037	.0986	.0966	.0792	.0919
40	.0518	.0493	.0483	.0396	.0460
100	.0207	.0197	.0193	.0158	.0184
200	.0104	.00986	.00966	.0072	.00919
Average of variance estimates from 1,000 replications**					
Samples of 20	.101	.0812	.0952	.0789	.0929
40	.0515	.0454	.0479	.0394	.0457
100	.0208	.0190	.0193	.0158	.0185
200	.0103	.00955	.00965	.00792	.00917

* For the variances of $\bar{y}'$ and $\bar{y}''$, the numbers shown are the conventional Taylor approximations to the variance.

** For the weighted mean and the regression and ratio estimators, the variance estimators are the conventional ones. For Models 1 and 2, the variance estimators are the so-called "error-variances," namely

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

and

$$1/(n-1)\ \{\Sigma\Sigma\ [y_{hi} - x_{hi}(\Sigma\Sigma y_{hi}/\sqrt{x_{hi}})/\Sigma\Sigma\sqrt{x_{hi}}]^2/x_{hi}^{3/2}\}\ \bar{x}^2/\Sigma\Sigma\sqrt{x_{hi}}$$

respectively, where $\bar{x}_s = (1/n)\Sigma\Sigma x_{hi}$, $\bar{y}_s = (1/n)\Sigma\Sigma y_{hi}$ and n is the total sample size. For Model 2, the computed error variance is the mean square difference between the estimate and $\beta\bar{X}$ rather than the difference between the estimate and the random variable $\bar{Y}$. There is some question as to which is appropriate. The difference between them is small.

Figure 1 shows the results of this exercise for a stratified sample of 200 from the realized population obtained by the procedures by which the 1,000 replicates were obtained. We suggested in Section 1 that it might appear reasonable to consider a line through the origin as an acceptable approximation for a model. The model-dependent estimators 1 and 2 were chosen on this assumption, along with ignoring the sample-selection plan.

If a test of significance were made to evaluate whether or not it was reasonable to use a line through the origin, the test 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 doing the test. From such computations (not shown) 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 a model 1 estimator would be adopted on the basis of a test of the hypothesis that the intercept of a straight line is zero with a significance level of .05. However, for a sample of 400 the bias squared is about four times the variance.

Royall and Eberhardt (1975) emphasize the bias of the conventional estimator and illustrate it with simple random samples. It is seen in this illustration that stratification can control the bias satisfactorily.

The variances estimated from the 1,000 replications (the first deck of numbers in Table 1) have approximately the expected relationships. 3/ For model 1 the estimated variances average approximately 9 percent smaller than the variances of the ratio estimator, and about 12 or 13 percent smaller for model 2.

Also, the estimated biases of $\bar{y}_w$, $\bar{y}'$, and $\bar{y}''$ are trivial for all sample sizes illustrated (the expected bias for $\bar{y}_w$ is zero). The biases of the model 1 and 2 estimators, estimated from the 1,000 replications, are negative and approximately constant for all the sample sizes, being about 4 percent for model 1 and about 2 percent for model 2. Thus, even though the model-dependent estimators have moderately smaller variances than the conventional estimators for all the sample sizes, their mean square 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 model 1, and between 40 and 100 for the model 2 estimator (which assumes the correct conditional variance of y).

In the case of the model 1 and 2 estimators, the observed bias is clearly the result of the fact that the estimators are inconsistent because they take no account of the sample design. The bias is reduced by model 2, since the estimator assigns relatively smaller weights to observations in the higher strata, but is still substantial.

Royall and Herson (1973) and Royall and Cumberland (1977) suggest use of balanced samples, with various means of balancing, to achieve robustness by reducing the bias of the model 1 estimator. 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. In our illustration only ten strata are sufficient to achieve reasonably good balance (that is, small departure of $\bar{x}$ from $\bar{X}$).

We present in Table 2 the results from proportionate sampling (using the same ten strata). More strata could be introduced, but the variability in the $\bar{x}$ for these strata is small enough that little more could be gained for the added work. The bias is seen to be trivial for the model 1 estimator, as for the first three estimators. However, the robustness of balanced sampling depends on the assumed model.

Table 2. Results of 1,000 replications of an approximately proportionate stratified sample* of size 100.

	$\bar{y}_w$	$\bar{y}'$	$\bar{y}''$	$\bar{y}'''$	$\bar{y}^{iv}$
Average of estimates from 1,000 replications					
Variance	.0215	.0215	.0211	.0213	.0199
Bias	-.011	-.009	-.012	-.015	.119
MSE	.0217	.0216	.0213	.0216	.0341
Theoretical variance**	.0240	.0229	.0229		
Average of variance estimates** from 1,000 replications	.0236	.0217	.0225	.0199	.0259

* The design is referred to as approximately proportionate stratified because of the trivial variation in the weights. Note that $\bar{y}_w \doteq \bar{y}_u \doteq \bar{y}'''$ (and $\bar{x}_w \doteq \bar{x}_u$) in these results. The weighted and unweighted averages would be identical except for trivial variations in the weights because n_h must be integral and cannot be proportionate exactly. The coefficient of variation of either $\bar{x}_u$ or $\bar{x}_w$ for stratified samples of 100 is 1.4 percent.

**See the footnotes in Table 1.

Thus, it is seen by comparing Tables 1 and 2 that for this illustrative population a balanced sample results in a substantially increased bias for an estimator based on model 2. Moreover, an overall balanced-sample approach restricts the sample design, and does not allow the sometimes substantial gains from stratified sampling with approximately optimum allocation. Also, it may not achieve balancing for various subdomains of interest. The price paid for overall balancing is increased variance as compared to approximately optimum allocation, as is seen by comparing Tables 1 and 2. The price paid would be considerably more for many commonly encountered populations that are much more skewed (see Hansen et al. [1953] for examples).

Obviously, the biases of the model 1 and model 2 estimators could be made trivial, in this particular case, by using a line not required to go through the origin. However, this would not be a general solution. For example, a model that would fit a population better 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, the problems of model failure will remain unless the designs are sufficiently robust as to be model-independent, in which event they are or are essentially equivalent to probability-sampling designs.

3.1 Role of "Best" Estimators

Godambe (1955) showed that there is no uniformly best linear unbiased estimator for estimating a population total, where linear was defined in a general manner that 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 linear). Thus, Godambe's results did not create any difficulties. The standard practice had been to attempt to determine for which types of finite populations a design that included an estimator y_1 had a smaller mean square error than an alternative design that includes an estimator y_2. Minimization methods had been used only within specific classes of estimators or other aspects of design.

It should be noted that Godambe's proof did not contradict the existence of best linear unbiased estimators as discussed by Neyman (and others) since Godambe adopted a different definition of linearity. It should be noted, also, that in probability sampling unbiasedness often results in much larger mean square errors than necessary. Instead, consistency of estimators has been the general criterion adopted. It seems obvious that for the general class of consistent estimators there are no uniformly best estimators.

Much interest is still expressed in best unbiased estimators in model-based approaches. The need for assuming models in order to have best estimators is expressed by a number of authors (Godambe [1978]), Basu [1971], Royall [1970], Sarndal [1978], Cassel et al. [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 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 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. 101)

Added costs of an estimation procedure (in dollars or time) may exceed the gains from reduced variance. In many sample surveys multiple statistics are involved, often a great many of them. Also, timeliness of results is often an important consideration. One may then find it advantageous to adopt estimators with larger variances than those of available alternative estimators for many 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. Several thousand different estimates are published each month, and the data are processed 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.

3.2 Robustness in Surveys

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, variance estimates for

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. This is done in order to adapt the sample design to various special characteristics of the population in order that (a) the possible samples result in acceptably small variances at least for the principal items estimated, (b) for a moderate sample size the distributions of the statistics will be approximately normal and asymptotic approximations will be acceptable.

Occasionally such efforts are not fully successful and a problem occurs when an outlier is observed in a sample. Sometimes there is the related, but often much less serious, problem that there are undetected cases not in the sample but that would be outliers if they had 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. When the outlier is excluded the estimate is 10 percent different than if the outlier is included. What constitutes outliers, in the sample or not in the sample, depends, of course, on the size of the sample. An outlier in a sample of 100 might not be an outlier if the sample were increased to a thousand. Some unknown potential outliers in the population (and not in the sample) are of no concern if they are insufficient to influence importantly aggregate or average characteristics of the population. If they are large enough to have such important impact, they can be adequately dealt with only through appropriate treatment in sample design.

When an outlier occurs in a sample the normal approximations (and any asymptotic 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 two outliers occur, one of two kinds of efforts, or both, are ordinarily taken in probability sampling. The first is to exclude the outlier from the sample, or to arbritrarily reduce its weight in the sample estimate. The second 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 larger sample, 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, with listing of housing units (hu's) in the sampled blocks, and subsampling from the listed hu's. Varying probabilities of selection of the blocks may be used to control variation in size, based on prior information on block sizes. This information may be supplemented by special work with building permits so that new construction is reflected in the measures of size. Alternatively, a new procedure may be used for sampling new construction occurring since the date of the information that provides the measure of size. Such an approach ordinarily is quite effective. However, it may be found that the sample includes one block that is an outlier. This might arise, for example, because the block identification was improper. In any event, one block is found in the sample that was expected to have only a few or no housing units, and in fact contains a large development, with perhaps a hundred housing units in the sample after the subsampling from the listing. The result is substantial so that multiplying by the reciprocal of the probability of selection would lead to that block accounting for, say, 20 percent of a sample-survey estimate. In such a case it may be feasible to take corrective action, e.g., by driving past all or a large sample of blocks with small measures of size, and redetermine the probability of selection of such blocks so that there will no longer be outliers. If such or an equivalent procedure is not feasible for an outlier it may be desirable to reduce the weight of the sampled outlier. In such an event it is important to carefully qualify any statements of precision or MSE 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 is made an outlier by 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 are not 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. This particular problem seems to be more or less a stand-off in the two procedures.

The issue of robustness arises when model-dependent methods are used for sample selection or 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 taking the n cases with the largest x values. This approach was suggested for populations having characteristics similar to the illustrative 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 increased, but at a rate such that

$$\sigma^2_{y|x}/x^2$$

decreases as x increases. For the particular illustrative population, such a design results in a relatively larger bias, even for quite small samples, and the bias will greatly outstrip the variance.

Royall has discussed various procedures to reduce the risk of bias with model-dependent designs, and seeems, 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 approches and probability sampling. Thus, in 1970 and 1971 he recommended cut-off methods in situations where some commonly-encountered models seemed to hold. He also recommended, in 1970, disregarding the sample-selection procedure in estimation from the sample. However, in 1973 Royall and Herson recommended balanced samples, 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), with balanced sampling within strata, and using separate ratio estimates to individual-stratum aggregates of an independent variable. The variable selection probabilities are reflected so that this becomes a consistent estimator. By this time the difference between probability-sampling and model-dependent approaches has substantially disappeared (totally, if random selections are made within strata). They use model-dependent approaches to optimize the definition of the size-strata, as would a probability sampler. Having gone this far, they would, by taking the final selections by a probability process within the approximately optimized strata, increase the variance trivially 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 confidence intervals, there is still the risk of seriously misleading estimates and confidence intervals. Such risks will be especially serious when relatively precise and accurate results are needed and paid for with large sample size.

4. Some Additional Remarks and Comments

4.1 Analysis of Survey Results

Survey results often are used to describe characteristics of a finite population -- for example, the number of unemployed at a point or interval of time, or for analysis. Analysis may relate to the specific finite population, as in testing a hypothesis concerning a difference between ratios for two groups, say, the unemployment rates of males and females.

Very often, on the other hand, the analysis is concerned with inferences about a cause system. In this situation the finite population can only be regarded as a realization of that cause system. What a probability-sampling approach can do is provide appropriate information about the available realization, or successive realizations, of the cause system, and for this the above discussion is directly relevant. However, for inferences about the cause system, and the ability to predict future developments, 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 cause systems from surveys of finite populations has involved, again, the differing points of view that have been discussed earlier, on use of survey results. One view often expressed is that the inferences to a cause system do not depend on the survey design, and that the design of the sample in such instances should be ingnored. The analysis is done as if the only source of variation were simple random sampling from a hypothetical superpopulation. Another view is that the design is relevant, including effects of intraclass correlations from cluster sampling, variable sampling fractions, and other aspects of design, and that failure to recognize their effect will lead to understatement of confidence intervals and overstatements of precision in inferences to the cause system, and that such factors should be appropriately reflected in the models, as in drawing inferences about the finite population. We concur strongly with the latter view. However, discussion of this topic, beyond simply mentioning it, is beyond the scope of this paper. Kish, and Kish and Frankel (1974) have pioneered some of the work in this area.

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

4.2 Inferences from Prediction Theory and from Probability Samples

A criticism of probability sampling by some who advocate model-dependent approaches is that probability-sample-survey theory ignores the fact that to estimate, say, the population total is equivalent to predicting the total of Y for the part of the population not in the selected sample. Hence, it is asserted that there must be assumptions made, in the sense of probabilistic dependence on the parameters, that relate those in the sample to those not in the sample, in order that any inference about those not in the sample will be meaningful. If such relations exist, and are known, probability selection is unnecessary. 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 information for the selected sample and the information for the balance of the population, how that sample was selected cannot create the relationship. However, we wish to emphasize that probability-sampling methods provide a confidence interval for the population characteristic being estimated, and for large enough samples the confidence interval will be valid and short enough to provide as precise statements as desired about the value being estimated. The question becomes one of the gains and losses from the assumption of a particular superpopulation, with increased risk of misleading results from dependence on superpopulation models and their consequences as sample size increases.

4.3 Some Concluding Remarks

It is always possible, in selecting samples and making estimates for a finite population, to make assumptions that will yield shorter computed confidence intervals per unit of cost than can be obtained by applying probability-sampling methods. However, the probability that the computed interval covers the population value may be substantially less than the cited confidence coefficient. The essentially assumption-free probability-sampling methods, taking advantage of models only to guide design choices often can be applied with little or no increases in unit costs for achieving a given length of confidence interval as compared with the computed interval for reasonably carefully applied model-dependent methods. 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 unit costs to obtain the additional assurance provided by the probability methods.

On the other hand, when surveys are taken with relatively small samples, even if finite-population estimates are the goal, the samples may be too small for the essentially assumption-free aspects of finite theory to be reasonably applicable. It may then be an advantageous use of resources to use methods that depend on the validity of superpopulation models, perhaps select purposive samples, and make estimates based on the assumed distributions or models. In most practical problems the essentially assumption-free aspects of probability-sampling theory are applicable only with acceptably larger samples.

There are major advantages in the acceptability and face validity of results that can be supported without having to defend model-based assumptions. This may not be so important for personal uses of data, but it 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 stakes are large.

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

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

1. "Best" estimators are not possible except by unduly restricting the class of estimators. A "best" estimator obtained by assuming a model depends on the validity of the model and may yield confidence intervals that are seriously misleading. Moreover, model-dependent approaches in which the model is a considered one are feasible for one or a few statistics but not for many statistics from a survey since multivariate models may be required.

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 and acceptably small confidence interval.

3. Model-dependent designs, even those that use "robust" procedures, face the risk of substantially understating the MSE, even when the model appears satisfactory. Model-dependent approaches in which the model is adjusted to be consistent with survey results may substantially understate the lengths of the confidence intervals, whether the confidence interval is viewed as conditional on a single sample or for all possible samples. This is especially true as sample sizes increase.

4. Probability-sampling methods, with reasonably large samples, provide protection against failures of assumed models, and provide robustness for all estimates, including estimates for any subdomains for which the samples are reasonably large. Models are appropriately used to guide in design of probability samples.

5. Ordinarily, 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 valid.

6. When inferences are made to a cause system, instead of to a finite population, there is no choice. Probability-sampling methods are not available for drawing samples from a causal system, but only from some finite realization of that cause system. One must use model-dependent inferences. Nevertheless, probability samples of finite realizations of the cause system may be highly useful in arriving at inferences in many situations.

7. The proper use of models has much to contribute to survey design. We urge continuing strong efforts, taking the fullest feasible advantage of models, but ordinarily within the framework of probability sampling, i.e.,

using designs and estimators that are not model-dependent.

1/ At the suggestion of V.P. Godambe, this paper is dedicated to the memory of our very dear friend and close collaborator, William N. Hurwitz. We thank Professor Godambe for the suggestion.

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

3/ 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, 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. If the $\sigma^2_{y|x}$ has been proportionate to x, instead of to $x^{3/2}$, 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. 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.

REFERENCES

Basu, D. (1971), "An essay on the logical foundations of survey sampling, Part One," in V. P. Godambe and D. A. Sprott, eds., *Foundations of Statistical Inference*, Toronto: Holt, Rinehart and Winston of Canada, Ltd., 203-233.

Cassel, C. M., Sarndal, C. E. and Wretman, J. H. (1977), *Foundations of Inference in Survey Sampling*, New York: John Wiley & Sons.

Godambe, V. P. (1955), "A unified theory of sampling from finite populations," *Journal of the Royal Statistical Society*, B17, 369-278.

------- (1966), "A new approach to sampling from finite populations - I, II," *Journal of the Royal Statistical Society*, B28, 310-328.

------- (1978), "Estimation in survey sampling: robustness and optimality," unpublished.

Hansen, M. H. and Hurwitz, W. N. (1943), "On the theory of sampling from finite populations," *Annals of Mathematical Statistics*, 14, 332-362

------- and Hurwitz, W. N. (1949), "On the determination of optimum probabilities in sampling," *Annals of Mathematical Statistics*, 20, 426-432.

-------, Hurwitz, W. N., and Madow, W. G. (1953), *Sample Survey Methods and Theory*, New York: John Wiley & Sons, Inc.

Kish, L. and Frankel, M. R. (1974), "Inference for complex samples," *Journal of the Royal Statistical Society*, B36, 1-37.

Royall, R. M. (1970), "On finite population sampling theory under certain linear regression models," *Biometrika*, 57, 377-387.

------- and Herson, J. (1973), "Robust estimation in finite populations," *Journal of the American Statistical Association*, 68, 880-893.

------- and Eberhardt, K. R. (1975), "Variance estimates for the ratio estimator," *Sankhya*, 37, Series C, 43-52.

------- 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.

-------- (1978), "Variance estimation in finite population sampling," *Journal of the American Statistical Association*, 73, 351-358.

Sarndal, C. E. (1978), "Design-based and model-based inference in survey sampling," *Scandinavian Journal of Statistics* (in press).

Smith, T. M. F. (1976), "The foundations of survey sampling: a review," *Journal of the Royal Statistical Society*, A139, 183-204.