CHAPTER 10

Design and Model

10.1. USES OF DESIGN AND MODEL IN SAMPLING

In the design-based approach to survey sampling, the values of a variable of interest (y-values) of the population are viewed as fixed quantities and the selection probabilities introduced with the design are used in determining the expectations, variances, biases, and other properties of estimators. In the model-based approach, on the other hand, the values of the variables of interest in the population are viewed as random variables, and the properties of estimators depend on the joint distribution of these random variables.

One reason for the historical reliance on design-based methods in sampling, in addition to the elimination of personal biases in selecting the sample, is that in many cases—and especially with natural populations—very little may be known about the population. Most researchers find it reassuring in such a situation to know that the estimation method used is unbiased no matter what the nature of the population itself. Such a method is called *design-unbiased:* The expected value of the estimator, taken over all samples which might be selected, is the correct population value. Design-unbiased estimators of the variance, used for constructing confidence intervals, are also available for most such designs.

One area of sampling in which the model-based approach has received considerable attention is in connection with ratio and regression estimation. In many sampling situations involving auxiliary variables, it seems natural to researchers to postulate a theoretical model for the relationship between the auxiliary variables and the variable of interest. A model can, of course, also be assumed for populations without auxiliary variables. For example, if the N variables $Y_1, \ldots, Y_N$ can be assumed to be independent and identically distributed, many standard statistical results apply without reference to how the sample is selected. However, it is difficult to cite examples of survey situations in which a model of independent, identically distributed y-values can be assumed with confidence. In fact, a pervasive problem with the model approach to sampling is that for many real populations, attempts to specify models have been far from adequate. Typically, the

Sampling, Third Edition. Steven K. Thompson.

models become mathematically complex while still not being suitably realistic. In particular, any model assuming that the y-values are independent (or have an exchangeable distribution) ignores the tendency in many populations for nearby or related units to be correlated.

Moreover, many survey programs need to produce results that will be used by people of widely different viewpoints and often conflicting preferences regarding whether an estimate should be higher or lower. For example, a demographic survey may be used to allocate governmental resources from one district to another; a fishery survey may be used to determine the amount of commercial catch allowed. It would be hard in such a situation to propose a model that would seem acceptable or realistic to all interested parties. In such a situation, the elimination of ordinary human selection biases through some sort of random selection procedure can be a powerful pragmatic argument in favor of an approach that is at least partially design-based.

With some populations, however, experience may have established convincingly that certain types of patterns are typical of the y-values of that type of population. For example, in spatially distributed geological and ecological populations, the y-values of nearby units may be positively correlated, with the strength of the relationship decreasing with distance. If such tendencies are known to exist, they can be used in obtaining efficient predictors of unknown values and in devising efficient sampling procedures. This model-based approach has been prevalent in sampling for mining and geological studies, in which the cost of sampling is particularly high and the economic incentive is strong for obtaining the most precise possible estimates for a given amount of sampling effort.

Sources of nonsampling error must be modeled if they are to be taken into account. Problems of differential response, missing data, measurement errors, and detectability must be modeled in some way in order to adjust for biases and to assess the uncertainty of estimates.

10.2. CONNECTIONS BETWEEN THE DESIGN AND MODEL APPROACHES

Let $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ denote the vector of y-values associated with the N units of the population. From the model viewpoint, these y-values are random variables with some joint distribution F. Let P(s) denote the probability under the design of selecting sample **s**, where **s** is a sequence or subset of the units in the population.

From the sample of n units, one wishes to estimate or predict the value of some quantity y_0, where y_0 may, for example, be the population mean, the population total, or the y-value at a unit not in the sample. The predictor or estimator $\hat{y}_0$ is a function of the y-values of the sample.

An estimator or predictor $\hat{y}_0$ is said to be *design-unbiased* for y_0 if its conditional expectation, given the realization of the N population y-values, is the realized value of y_0, that is, if

$$\mathrm{E}(\hat{y}_0|\mathrm{y}) = y_0$$

Notice that, although y_0 may be viewed as a random variable, with a distribution determined by F, the design-unbiased estimator $\hat{y}_0$ is unbiased under the design for the realized value of y_0—the actual value that y_0 has taken on at the time of the survey. The distribution F, which produced the population y-values, is thus irrelevant to this unbiasedness.

An estimator or predictor $\hat{y}_0$ is said to be *model-unbiased* for y_0 if, given any sample **s**, the conditional expectation of $\hat{y}_0$ equals the expectation of y_0, that is, if

$$\mathrm{E}(\hat{y}_0|\mathrm{s}) = \mathrm{E}(y_0|\mathrm{s})$$

No matter what sampling design gave rise to the sample **s**, the model-unbiased predictor $\hat{y}_0$ is unbiased under the population distribution F for y_0 given the sample **s** obtained. The design that produced the sample **s** is thus irrelevant to this unbiasedness.

An estimator or predictor $\hat{y}_0$ is *unbiased* (i.e., unconditionally unbiased) for y_0 if the expectation of $\hat{y}_0$ equals the expectation of y_0, that is, if

$$\mathrm{E}(\hat{y}_0) = \mathrm{E}(y_0)$$

Any estimator that is *either* design-unbiased or model-unbiased for y_0 will be (unconditionally) unbiased for y_0, by a well-known property of expectation.

Thus, if the desired end is simply unbiasedness, it can be achieved through either the design or the model approach. However, some authors philosophically demand one or the other types of unbiasedness—design unbiasedness, so that assumptions about the population are not relied upon, or model unbiasedness, so that the particular sample obtained is taken into account.

The mean square error associated with predicting y_0 with $\hat{y}_0$ is

$$\mathrm{E}(y_0 - \hat{y}_0)^2$$

the expectation being taken with respect to both the distribution of the population values and the design. If $\hat{y}_0$ is unbiased for y_0, the mean square error is the variance of the difference:

$$\mathrm{E}(y_0 - \hat{y}_0)^2 = \mathrm{var}(y_0 - \hat{y}_0)$$

From the model viewpoint, interest focuses on the conditional mean square error, given the sample **s**. If $\hat{y}_0$ is model-unbiased for y_0, this mean square error is a conditional variance:

$$\mathrm{E}[(y_0 - \hat{y}_0)^2|\mathbf{s}] = \mathrm{var}(y_0 - \hat{y}_0|\mathbf{s})$$

From the design viewpoint, the concern is with the conditional mean square error given the realized population y-values. When $\hat{y}_0$ is design-unbiased for y_0, this conditional mean square error is

$$\mathrm{E}[(y_0 - \hat{y}_0)^2|\mathbf{y}] = \mathrm{var}(\hat{y}_0|\mathbf{y})$$

If $\hat{y}_0$ is design-unbiased, the unconditional mean square error may be written

$$\mathrm{E}(y_0 - \hat{y}_0)^2 = \mathrm{E}[\mathrm{var}(\hat{y}_0|\mathbf{y})]$$

If $\hat{y}_0$ is model-unbiased, the unconditional mean square error may be written

$$\mathrm{E}(y_0 - \hat{y}_0)^2 = \mathrm{E}[\mathrm{var}(y_0 - \hat{y}_0|\mathbf{s})]$$

Estimators of variance may in similar fashion be design- or model-unbiased. A variance estimator that is either design-unbiased or model-unbiased will be unconditionally unbiased. Thus, with the design simple random sampling, the usual estimator

$$\widehat{\mathrm{var}}(\hat{y}) = \left(\frac{N-n}{N}\right)\frac{s^2}{n}$$

which is design-unbiased for $\mathrm{var}(\overline{y})$, is unbiased for the true mean square error no matter what distribution may give rise to the population.

10.3. SOME COMMENTS

A main result of the preceding section is that a sampling strategy is unconditionally unbiased if it is either design-unbiased or model-unbiased. Even so, the two approaches may lead to conflicting recommendations. An assumed-ratio model may suggest purposive selection of the units with the highest x-values; such a procedure is certainly not design-unbiased. The sample mean may be design-unbiased under simple random sampling; but under an assumed model the sample mean for the particular sample selected may not be model-unbiased. Some advantages of a design-based approach include obtaining unbiased or approximately unbiased estimators (and estimators of variance) that do not depend on any assumptions about the population—a sort of nonparametric approach—obtaining estimates acceptable (if grudgingly) by users with differing and conflicting interests, avoiding ordinary human biases in selection, obtaining fairly representative or balanced samples with high probability, and avoiding the potentially disastrous effects of important but unknown auxiliary variables. Some benefits of a model-based approach include assessing the efficiency of standard designs and estimators under different assumptions about the population, suggesting good designs to use—or good samples to obtain—for certain populations, deriving estimators that make the most efficient use of the sample data, making good use of auxiliary information, dealing with observational data obtained without any proper sampling design, and dealing with missing data and other nonsampling errors.

For a real population, however, even the best model is something one not so much believes as tentatively entertains. Under the assumption of the model, one can outline an efficient course of action in carrying out a survey. It is also nice to be able to say that if that assumption is wrong, the strategy still has certain

desirable properties—for example, the estimator is still unbiased, if less efficient. One approach combining design and model considerations uses the best available model to suggest an efficient design and form of estimator of the population mean or total while seeking unbiasedness or approximate unbiasedness under the design, and using estimators of variance that are robust against departures from the model. With this approach, one looks for a strategy with low unconditional mean square error, subject to the required (exact or approximate) design unbiasedness. Such an approach has been useful in the development of such survey methods as the generalized ratio and regression estimators under probability designs. "Model-assisted" strategies such as these, using models to suggest good sampling designs and inference procedures but seeking to have good design-based properties that provide robustness against any possible departures from the assumed model, are described in depth in Särndal et al. (1992).

Reviews of the ideas and issues involved in the relationship of design and model in sampling are found in Cassel et al. (1977), Godambe (1982), Hansen et al. (1983), Hedayat and Sinha (1991), Särndal (1978), Smith (1976, 1984), Sugden and Smith (1984), M. E. Thompson (1997), and Thompson and Seber (1996).

10.4. LIKELIHOOD FUNCTION IN SAMPLING

In the design-based, fixed-population approach to sampling, the values $(y_1, \ldots, y_N)$ of the variable of interest are viewed as fixed or given for all units in the population. With this approach, the unknown values y_i of the variable of interest in the population are the unknown parameters. For designs that do not depend on any unobserved y-values the likelihood function is constant, equal to the probability of selecting the sample obtained, for every potential value $\mathbf{y}$ of the population consistent with the sample data (Basu 1969).

In the model-based approach, the population values $\mathbf{y}$ are viewed as realizations from a stochastic distribution. Suppose that there is a population model $f(\mathbf{y}; \theta)$, giving the probability that the y-values in the population take on the specific set of values $\mathbf{y} = (y_1, y_2, \ldots, y_N)$. This probability may depend on an unknown parameter θ as well as on the auxiliary variables. The distribution may also depend on auxiliary variables. However, the dependence of the data, sampling design, and model on auxiliary variables will be left implicit in this section for notational simplicity. Also for ease of notation, assume that the variable of interest is a discrete random variable, so that sums rather than integrals are involved in the likelihood function.

The likelihood function is the probability of obtaining the observed data as a function of the unknown parameters. The data in sampling consist of the units in the sample together with their associated values of the variable of interest and any auxiliary variables recorded. For simplicity, the data can be written $d = (s, \mathbf{y}_s)$, where s is the set or sequence of units selected and y_s represents the y-values in the sample. Let p denote the sampling design giving for every possible sample the probability that it is the one selected. Now in general, the design can depend

on auxiliary variables $\mathbf{x}$ that are known for the whole population and even on the variable of interest $\mathbf{y}$. For example, in surveys that rely on volunteers or that involve nonresponse, the probability of volunteering or of responding, and hence being in the sample, is often related to the variable of interest. The adaptive sampling designs in the last part of this book also depend on the variable of interest. Thus, the sampling design can be written $p(s|\mathbf{y})$.

The likelihood function is thus the probability that the sample s is selected and the values $\mathbf{y}_s$ are observed and can be written

$$L_d(\theta) = \sum p(s|\mathbf{y}) f(\mathbf{y}; \theta)$$

where the sum is over possible realizations of the population $\mathbf{y}$ that are consistent with the observed data d. Since the y-values in the sample are fixed by the data, the sum is over all possible values $\mathbf{y}_{\bar{s}}$ for the units not in the sample.

An important point to note is that in general the likelihood function depends on both the design and the model. A prevalent mistake in statistics and other fields is to analyze data through careful modeling but without considering the procedure by which the sample is selected. The "likelihood" based on the model only, without consideration of the design, was termed the *face-value likelihood* by Dawid and Dickey (1977) because inference based on it alone takes the data at face value without considering how the data were selected.

There are certain conditions, however, under which the design can be ignored for inference. For any design in which the selection of the sample depends on y-values only through those values $\mathbf{y}_s$ included in the data, the design probability can be moved out of the sum and forms a separate factor in the likelihood. Then the likelihood can be written

$$L_d(\theta) = p(s|\mathbf{y}_s) \sum_{\mathbf{y}_{\bar{s}}} f(\mathbf{y}; \theta)$$

The design then does not affect the value of estimators or predictors based on direct likelihood methods such as maximum likelihood or Bayes estimators. For any such "ignorable" design, the sum in the likelihood above, over all values of $\mathbf{y}$ leading to the given data value, is simply the marginal probability of the y and values associated with the sample data. This marginal distribution depends on what sample was selected but does not depend on how that sample was selected. For likelihood-based inference with a design ignorable in this sense, the face-value likelihood gives the correct inference.

Likelihood-based inference, such as maximum likelihood estimation or prediction and Bayes methods, is simplified if the design can be ignored at the inference stage. The fact that the sampling design does not affect the value of a Bayes or likelihood-based estimator in survey sampling was noted by Godambe (1966) for designs that do not depend on any values of the variable of interest and by Basu (1969) for designs that do not depend on values of the variable of interest outside the sample. Scott and Smith (1973) showed that the design could become relevant to inference when the data lacked information about the labels of the units in the

sample. Rubin (1976) gave exact conditions for a missing data mechanism—of which a sampling design can be viewed as an example—to be relevant in frequentist and likelihood-based inference. For likelihood-based methods such as maximum likelihood and Bayes methods, the design is "ignorable" if the design or mechanism does not depend on values of the variable of interest outside the sample or on any parameters in the distribution of those values. For frequency-based inference such as design- or model-unbiased estimation, however, the design is relevant if it depends on any values of the variable of interest, even in the sample. Scott (1977) showed that the design is relevant to Bayes estimation if auxiliary information used in the design is not available at the inference stage. Sugden and Smith (1984) gave general and detailed results on when the design is relevant in survey sampling situations. Thompson and Seber (1996) discuss the underlying inference issues for adaptive designs, in which the selection procedure deliberately takes advantage of observed values of the variable of interest (and see the descriptions of these designs in later chapters of this book).

The concept of design ignorability thus depends on the model assumed, the design used, and the data collected. It is important to underscore that a design said to be "ignorable" for likelihood-based inference might not be ignorable for a frequentist-based inference, such as model-unbiased estimation, and that even though a design may be ignorable at the inference stage, in that, for example, the way an estimator is calculated does not depend on the design used, the design is still relevant a priori to the properties of the estimator. Ironically, in the real world, it is quite possible that the only data sets for which the designs are truly "ignorable" for inference purposes are those that were obtained through deliberately planned and carefully implemented sampling designs.