Journal of Official Statistics, Vol. 18, No. 2, 2002, pp. 129–154

Models in the Practice of Survey Sampling (Revisited)

Graham Kalton[1]

Although design-based inference is the standard form of inference with large-scale sample surveys, in practice some reliance on model-dependent inference is necessary. This article considers issues of model-assisted inference, the population of inference, conditional inference, the effect of measurement errors, and analytic uses of survey data, under the design-based mode of inference. It then discusses the need for model-dependent inference for small area estimation and for handling missing data (unit nonresponse, item nonresponse, and noncoverage). The article also discusses the use of models in variance estimation.

Key words: Design-based inference; model-dependent inference; missing data; small area estimation; generalized variance functions.

1. Introduction

Since the time of Neyman's classic paper (Neyman 1934), the standard mode of inference used by survey practitioners has been design-based inference based on the randomization induced by the sample design. The essence of this form of inference is the use of survey weights in producing survey estimates that are unbiased, or more usually approximately unbiased, and consistent in expectation over the distribution of all possible samples that could be selected with the given sample design. The variances of these estimates are also determined in relation to that distribution. This mode of inference does not depend on statistical models, unlike the situation in other fields of statistics.

The design-based mode of inference has been challenged from time to time. In the 1970's in particular, there were strong challenges from, for example, Royall (1970), Royall and Cumberland (1981), and Smith (1976). Hansen, Madow, and Tepping (1983) provided a major response to these challenges. As a discussant of Smith's paper, I responded to his question: ''The basic question to ask is why should finite population inference be different from inferences made in the rest of statistics?'' (Smith 1976, p. 193). Subsequently, I expanded on that discussion in a paper entitled ''Models in the practice of survey sampling'' (Kalton 1983a). This explains the parenthetical ''revisited'' in the current title.

In the first Morris Hansen lecture, Smith (1994) returned to the subject of inference from survey samples from a theoretical perspective. In this lecture, I am also revisiting the subject, but my perspective is that of a practitioner. My experience has mainly been with large-scale household surveys, and in recent years particularly with U.S. federal govern-

[1] Westat, 1650 Research Blvd., Rockville, MD 20850, U.S.A. Email: grahamkalton@westat.com
Acknowledgment: I would like to thank several Westat colleagues, and particularly Mike Brick, for their helpful comments on a draft of this article. The views expressed are, however, mine alone.

ment surveys, for which a prime objective is the production of descriptive statistics. In this context, I find the combination of probability sampling and the design-based mode of inference generally appropriate, although there are some limitations of the design-based approach that should be noted (see Section 2)[2]. The importance of large samples for both probability sampling and design-based inference should be made clear here. As Kish (1965, p.29) notes, ''Probability sampling for randomization is not a dogma, but a strategy, especially for large numbers.'' This statement also applies for design-based inference, as discussed later.

Most large-scale surveys are characterized by both large numbers of sampled cases and large numbers of estimates to be produced from the survey data. This latter feature is an important justification for the use of design-based inference. Survey reports often contain thousands of estimates. The development of model-dependent methods to produce good estimates for all of them is not feasible, whereas design-based methods can be applied universally. However, when there is a key estimate that is of paramount importance, then the rigorous development and thorough testing of a model-dependent estimator may be justified by the increased precision that it might have. The general approach of using methods that have widespread applicability at the cost of some loss of optimal properties for specific estimates is applied in other areas of survey sampling practice as well. For instance, the methods of weighting adjustment and imputation for handling missing data are general-purpose strategies, that may be suboptimal for a particular analysis (Kalton 1983b, pp. 17–18).

Although design-based inference may serve well for descriptive statistics from large-scale surveys, even in this case model-dependent methods are generally needed to some degree. In fact, many of the developments in survey sampling in the past quarter century have been concerned with the application of model-dependent methods to address such problems as missing data and small area estimation. These applications are reviewed in Section 3. Another area where model-dependent approaches are needed is that of the ''analytic'' uses of survey data, that is survey analyses that seek to identify and measure causal mechanisms. Some remarks on this complex subject are given in Section 2.5.

It is axiomatic that all models are false. The attraction of the design-based approach to survey inference for most descriptive estimation from large-scale surveys is its avoidance of reliance on models. Nevertheless, as well expressed in the popular quotation from Box (1979), ''All models are wrong, but some are useful.'' Models are indeed useful, and needed, for some problems in survey analysis. My general approach to the use of model-dependent methods for descriptive estimation is to treat the model as a crutch, to be used only to the extent that the survey data cannot fully support the desired estimates. If the sample is strong enough, and if there is no weakness from missing data, then design-based inferences alone will serve well. In practice, however, there are virtually always some missing data, and the sample sizes for subdomains may sometimes be too small

[2] The situation is somewhat different for establishment surveys because of the highly skewed distributions of many of the variables of interest, leading to small numbers of establishments dominating the survey estimates. Furthermore, the sampling frames for establishment surveys often contain auxiliary variables related to the sizes of the establishments that can play an important role in design and estimation. Although design-based inference is still the generally preferred mode for establishment surveys, the case for model-dependent methods is stronger in this area. See, for example, the predictive approach to inference advocated by Valliant, Dorfman, and Royall (2000).

to support subdomain estimates of adequate reliability. In consequence, it is necessary to rely on model-dependent methods to some extent. It should be noted that the reference here is to ''model-dependent'' inference, not to ''model-assisted'' inference. Models are widely used within the design-based mode of inference, both in sample design and in estimation, but in a ''model-assisted'' manner so that the validity of the survey estimates does not depend on the validity of the model assumptions (see Section 2.1).

Most of the issues discussed in this article have a long history and the arguments advanced by the early researchers still hold. To demonstrate this point, some of the early references are cited.

2. Issues in Design-Based Inference

Design-based inference (also known as randomization inference) is concerned with inferences about a finite population of size N, with fixed values for element i (Y_i, X_i, Z_i, etc.), based on data collected from a probability sample. Inferences are made about finite population parameters on the basis of the random selection process (e.g., about $\bar{Y} = \Sigma Y_i/N$ or $\mathrm{Cov}\,(X_i, Y_i) = \Sigma(X_i - \bar{X})(Y_i - \bar{Y})/N$). This section considers some issues that arise in applying the design-based paradigm in practice.

2.1. *Model-assisted versus model-dependent inference*

From the early days, models have been widely used in the design and analysis of survey samples. Models are used in designing samples in a variety of ways, such as modeling stratum variances in establishment surveys to determine strata sampling fractions, modeling cluster homogeneity for area samples to determine effective methods of clustering, and modeling variances for use in determining sample sizes. The suitability of the model chosen affects the efficiency of the sample design and hence the precision of the survey estimates but, when probability sampling is used, the models do not affect the validity of design-based inferences. The use of models for sample design will not be treated further here.

In the analysis of survey data either model-assisted or model-dependent methods may be used, using the terminology of Särndal, Swensson, and Wretman (1992). To illustrate the difference, consider the estimation of a population total Y based on a probability sample of size n in which element i is selected with probability π_i. The Horvitz-Thompson estimator $\hat{Y} = \Sigma^n y_i/\pi_i$ is a consistent design-unbiased estimate of Y that makes no use of a model. Suppose now that a vector of auxiliary variables $\boldsymbol{X}_i = (X_{1i}, X_{2i}, \ldots, X_{pi})$ is known for all population elements ($i = 1, 2, \ldots, N$), and that the relationship between Y_i and $\boldsymbol{X}_i$ can be modeled by $Y_i = f(\boldsymbol{X}_i) + e_i$, with $Y = \Sigma^N f(\boldsymbol{X}_i) + \Sigma^N e_i$. For simplicity we assume that $f(\boldsymbol{X}_i)$ does not include unknown parameters that need to be estimated from the sample. Then, without any modeling assumptions, Y may be estimated by $\hat{Y}_R = \Sigma^N f(\boldsymbol{X}_i) + \Sigma^n e_i/\pi_i$. Like $\hat{Y}$, $\hat{Y}_R$ is a consistent design-unbiased estimator of Y. Whether the model-assisted estimator $\hat{Y}_R$ has a lower variance than $\hat{Y}$ depends on the suitability of the model.

The difference estimator is a simple example of the above model-assisted estimator. Here $f(\boldsymbol{X}_i) = kX_i$, with k a predetermined constant, often set at 1. Then $\hat{Y}_R = kX + \Sigma e_i/\pi_i$, where $e_i = y_i - kx_i$, which may alternatively be expressed in the more usual

form as $\hat{Y}_R = \hat{Y} + k(X - \hat{X})$. The optimum value for k that minimizes $V(\hat{Y}_R)$, the variance of $\hat{Y}_R$, is $B = \Sigma^N(Y_i - \bar{Y})(X_i - \bar{X})/\Sigma^N(X_i - \bar{X})^2$. When k is estimated from the sample by b, the estimator of B, then $\hat{Y}_R$ becomes the regression estimator. Model-assisted estimators of this type have a long history. See, for example, Hansen, Hurwitz, and Madow (1953, Vol. 1, p. 457). For recent developments, see Särndal, Swensson, and Wretman (1992, Chapter 6).

The difference estimator and the other estimators of this type are model-assisted but not model-dependent because their design unbiasedness does not depend on any model assumptions. However, if assumptions are made about the e_i, then an alternative estimator may be used. For example, if the e_i are assumed to be *iid* variables with a mean of zero, then the model-dependent estimator $\hat{Y}_m = \Sigma^n y_i + \Sigma^{N-n} f(\mathbf{X}_i)$ may be used to estimate Y. In the case of the model corresponding to the difference estimator, $\hat{Y}_m$ reduces to $\Sigma^n y_i + \Sigma^{N-n} kX_i$. These estimators are design unbiased if $\Sigma^{N-n} e_i = 0$, and model-design unbiased if $E_\xi E_p(e_i) = 0$, where E_ξ and E_p denote expectation with respect to the model and the sample design, respectively.

2.2. *Population of inference*

Design-based inference is concerned with inferences about parameters of the finite population from which the sample was drawn. The restrictive nature of this inference needs to be acknowledged. One important area of concern relates to the analysis of survey data to measure causal influences, often called the ''analytic uses'' of survey data. That area is treated subsequently. For the present, the discussion is confined to the descriptive uses of survey data that simply aim to measure characteristics of the population, such as the proportion of children in poverty, the number of public schools with Internet access, and the national crop acreage for cotton. Even here the design-based mode of inference may be restrictive because of its failure to account for the dynamic nature of populations over time.

Most surveys are cross-sectional and relate, in principle at least, to a specific point in time (although in practice data collection may be spread over several months). After data collection, the production of survey estimates may take a year or more, by which time the composition and characteristics of the population will have changed to some degree. Although there are cases where the specific population at the time of data collection is of direct analytic interest (for example, a survey of a store's inventory as of December 30, 2001 may be of accounting interest), analysts generally use the survey data to infer the characteristics of the current population. To make estimates for the current population, they often make the assumption that negligible change has occurred since the data collection. Thus they employ an implicit model for change over time. For many populations and survey subjects, this model may be a reasonable one, but not necessarily for all. Some examples of the complications caused by the time dimension are given below.

First, consider inferences from educational assessment surveys, such as the U.S. National Assessment of Educational Progress (NAEP) or the Third International Mathematics and Science Study (TIMSS), for, say, mathematics proficiency at 8th grade. Suppose that the survey was conducted in 2001, and that the results are published in 2002. Consider a small state in which most schools and 8th graders are sampled. For

design-based inferences about the achievement levels of the population of 8th graders in 2001, the standard errors incorporate finite population corrections (fpc's). Ignoring measurement errors (see below), if all students were surveyed in the state, the standard errors would in fact be zero. When this standard approach to variance estimation is adopted, analysts are left to attach their own subjective degrees of uncertainty to the 2001 estimates when using these estimates to describe achievement levels in 2002.

Note that the population of 8th graders in 2002 has changed almost completely from that in 2001. For inferences about the 2002 population, a model is therefore needed. One way to construct this model is to assume a superpopulation model, from which the 2001 and 2002 student populations are drawn as random realizations. The schools and environments may be modeled as fixed, as constant features of the superpopulation. This form of model results in fpcterms being used for the sampling of schools but not being used for the students within schools. Chromy (1998) presents a motivation that supports this practice.

A special feature of the above example is that the composition of the population changes almost completely between the date of the survey and the date for inference. In most cases, this feature does not hold: there is, for example, a substantial overlap in the memberships of the adult populations of last year and of this year. It is mainly the characteristics of the population members that may change between the years. As a result, the procedure outlined above is not generally applicable.

As a second example, consider continuing surveys, such as the U.S. National Health Interview Survey (NHIS) that has been collecting data from nationally representative samples of individuals each week for many years, the proposed U.S. American Community Survey (ACS) that will collect data from national representative samples of households each month starting in 2003, and the U.S. Continuing Survey of Food Intakes by Individuals (CSFII) that collected data from nationally representative samples of individuals for each quarter in the three-year period 1994-96. These surveys are designed so that their data can be aggregated over time to provide larger sample sizes and, in the case of NHIS and CSFII, to provide annual estimates that average over seasonal variations. When data are aggregated in this way, what is the relevant population of inference? One approach is to incorporate the time dimension into the population definition, so that, for example, mean food intakes from three years of the CSFII refer to averages over individuals and over the three years. In this case, the survey weights are determined to produce representations of the population $\times$ time matrix. Hence, if the CSFII third year sample had been much larger than the samples for the first and second years (as had been contemplated but was not implemented), its sampled individuals would have been assigned lower weights in the analysis in order to produce the right annual balance of roughly one-third weight to each year. An alternative approach would be to employ a modeling assumption that food intakes do not vary across the time period for the survey. In this case the data can be pooled without the need for rebalancing to give roughly equal annual representation. From the perspective of using the survey data to estimate current food intakes, the latter approach has the attraction in this particular case of placing larger weight on the third, most recent, year, as well as producing estimates with lower sampling errors. Indeed, from this perspective, it might anyway be desirable to place larger weight on the most recent data (see also Kish 1999).

As a third example, consider the U.S. Current Population Survey (CPS) that collects

labor force data each month, primarily to produce monthly estimates. The data collected relate to a specific reference week in each month (the 7-day period including the 12th of the month). A problem with this procedure is that there are a fair number of occasions when the reference week is unusual because of the weather, holidays, etc., so that the survey estimates do not accurately reflect the population of interest to analysts.

Another feature of the CPS is that it is a repeated survey with a long time series of estimates available. Time series methods may be used with repeated surveys of this type to identify trends and seasonal components, to improve the precision of current estimates, and to seasonally adjust those current estimates (see, for example, Bell and Hillmer 1990; Scott, Smith, and Jones 1977). The use of time series models provides a means of forecasting current (and future) parameters, without making the assumption of no change across short time periods.

Finally, consider panel surveys such as the U.S. Survey of Income and Program Participation (SIPP) and the U.S. National Education Longitudinal Study of 1988 (NELS: 88). The unique value of panel surveys is that they provide measures of gross change, that is, change — or lack of change — at the individual level. Thus, for example, the 1996 SIPP panel can produce an estimate of the proportion of individuals who were continuously in poverty between 1996 and 1999. Such estimates have important policy implications, yet they are historical data by the time they are available, when generally the measure of main interest is the proportion of individuals in poverty in 2001 who will still be in poverty in 2004. The model assumption of no change in this proportion between the 1996-99 and 2001-04 periods is a tenuous one given the length of time in between, and particularly so if economic conditions are changing. As a check on the constancy of the proportion remaining in poverty over any 3-year period over time, a rotating panel design could usefully be employed, say starting a new panel every year. A finding that the estimates for different 3-year periods produced from the different panels are similar supports the ''no change'' assumption. A finding that the estimates for different periods differ markedly highlights the difficulty in forecasting the future. Even when the estimates for different time periods are very different, any understanding that can be obtained of the reasons for the differences may be useful in making predictions for a later period.

2.3. Conditional inference

In applying the design-based approach in analyzing survey data, an issue that arises in a number of different contexts is whether inferences should be made conditional on some aspects of the realized sample or whether the inferences should be unconditional, that is based on the full set of possible samples that the sample design could have generated. For example, for a sample design in which the sample size is a random variable, should the variances of the survey estimates be computed conditional on the realized sample size or should they be computed unconditionally over the distribution of all possible sample sizes?

To examine this issue from a design-based perspective, consider an estimator z for a population parameter Z. The design-based approach seeks estimators that are design unbiased, or more commonly approximately so, for the finite population parameter given the sample design employed. Let z be unbiased for Z, denoted by $E(z) = Z$, where the

expectation is over all possible samples. Let the variance of z be $V(z) = E(z - Z)^2$. Now consider a feature of the realized sample, a condition C, that has the property that $E(z \mid C) = Z$, i.e., z is conditionally unbiased for Z, and let $V(z \mid C) = E(z - Z \mid C)^2$. The issue is whether the precision of z should be measured by the unconditional variance $V(z)$ or by the conditional variance $V(z \mid C)$. (In passing, it is interesting to note that if $V(z)$ is an unbiased estimate of $V(z \mid C)$ from the realized sample, $v(z)$ is also unbiased for $V(z)$. This result holds because $V(z) = E_c V(z \mid C) + V_c E(z \mid C)$ and $V_c E(z \mid C) = 0$, where E_c and V_c denote expectation and variance over the possible outcomes of the condition.)

In my view, if a subset of the sampling distribution in which z is conditionally approximately unbiased for Z can be identified, then a conditional analysis should be employed in analyzing an actual sample. Under some outcomes of the condition, the conditional approach will yield smaller variance estimates than the unconditional approach, whereas under other outcomes the reverse will hold. Using the conditional approach, the analyst should employ the conditional variance in either case. It is invalid to use the smaller of the conditional and unconditional variances as has sometimes seemed to be suggested. A number of survey statisticians support the conditional approach. However, Smith (1994, p.18) argues for the unconditional approach on the grounds that: ''Once one level of conditioning has been accepted, it opens the Pandora's box of all other possible conditions.'' In his discussion of Smith's paper, Royall (1994) presents a counter-argument (see also Valliant, Dorfman, and Royall 2000).

Often the condition that is considered is the set of achieved sample sizes in various subclasses. In this case, the issue is whether the precision of the survey estimates should be assessed conditional on the realized subclass sample sizes (see Rao 1985). Two common examples are the estimation of a subclass mean, when the subclass sample size is a random variable, and the application of poststratification, when the sample sizes in the poststrata are random variables. Consider first the case of estimating a subclass mean or a poststratified mean from a simple random sample. In both these cases, the sample mean is unbiased for the population mean conditional on the realized subclass sample sizes (assuming that there is at least one sampled element in each poststratum).

In the case of the subclass mean, at the design stage the subclass sample size that will be realized is not known. Hence, for design purposes the unconditional variance may be used for determining the overall sample size (however, see below). When the sample has been selected, should the inference be based on the realized sample size or not? Suppose that a sample of 1,000 persons is selected, and that the subclass represents ten percent of the population. The expected subclass sample size is thus 100, but the realized sample size may be smaller or larger than that. If the realized sample size is 85, then the estimate of the subclass mean is clearly less precise than if the sample size were 115. Thus, I believe that correct inference about the subclass population mean should be based on the achieved sample size.

As noted above, at the design stage the realized subclass sample size will not be known. Thus, it is natural to consider the use of the unconditional variance in planning the overall sample size to yield a subclass mean of desired precision. However, if this is done, the realized precision will fall short of the level desired almost half the time (and will otherwise generally exceed it). To guard against this outcome, it may be appropriate to increase

the overall sample size, with the increase being based on the distribution of the conditional variances (a procedure adopted in the CSFII for producing estimates for subclasses defined in terms of age, sex, and poverty status).

The case for the use of conditional inference for the analysis of poststratified simple random samples (conditioning on the realized samples in the poststrata) has been well made by Holt and Smith (1979). They also provide references to indicate that many other statisticians also support this position. One early statement is that by Williams (1964, p. 1060): ''After the sample has been selected, the conditional estimator, given the realized distribution of the sample, is frequently the most relevant.''

The discussion above applies only for simple random sampling. The situation is different for complex sample designs (Rao 1985). Consider, for example, a two-stage sample in which the PSUs are selected with probabilities proportional to estimated sizes, and the sample sizes in the PSUs are allowed to vary in order to produce an overall equal probability sample of elements. The sample mean for the total sample or for a subclass is estimated by $r = \Sigma y_\alpha / \Sigma x_\alpha$, where y_α is the total of the y-values and x_α is the sample size in sampled PSU α. Standard practice treats r as a ratio mean, with a random sample size (see, for example, Kish 1965) and the variance of r is computed unconditionally. The unconditional approach is applied in this case because in general, when conditioning on the sample size x, $E(r \mid x) \neq \bar{Y}$, thus not satisfying the conditional unbiasedness requirement of the conditional approach.

Failure to satisfy the requirement of conditional unbiasedness (or rather conditional approximate unbiasedness) also applies to poststratified estimates from complex designs, when conditioned on realized sample sizes (Rao 1985; Zhang 2000). For this reason, variance estimates of poststratified estimates from complex designs should be computed unconditionally. The required variance estimates can be produced by recomputing the poststratified weights for each replicate with replication methods of variance estimation or by the use of an appropriate procedure with the Taylor Series linearization approach.

Another issue arises with poststratification. With the design-based approach the sampling distribution of a survey estimator is generally established under a design and estimation scheme that is prespecified. Thus, the theory treats the poststrata as predetermined; they are not affected by the realized sample. However, as Alexander (1994) notes, in practice the choice of poststrata is sometimes affected by the sample outcome. There are dangers here since the sampling distribution is altered in ways that can be determined only by considering the poststratification adjustments that would be chosen for each possible sample outcome, and this is impossible. Determining poststrata based on sample outcomes can lead to biased estimates with variances that differ from the textbook formulae.

2.4. Measurement errors

Standard design-based theory, like most statistical theory, ignores the problem of measurement error. Thus, there is a true value for each element in the population and that is the value reported if that element is sampled. In practice, there are of course measurement errors that could arise, for instance, because the respondent provides an inaccurate answer, the interviewer records the answer incorrectly, the answer is coded incorrectly, or another type of processing error occurs.

A seminal paper on measurement errors in surveys is that of Hansen, Hurwitz, and Bershad (1961). Their model separates the measurement error for each individual into bias and variable error components. Variable errors occur when the individual gives different responses to a survey question over conceptually repeated trials of the survey under the same essential survey conditions. Let y_{it} be the response of an individual on trial t, and $y_{it} = \mu_i' + d_{it}$ where $E_t(d_{it}) = 0$, with the expectation being taken over trials. The individual response bias is then $\beta_i = \mu_i' - \mu_i$, where μ_i is that individual's true value. The individual response variance is $\sigma_i^2 = E_t(d_{it}^2)$.

Under this model, the sample mean for a particular trial from an equal probability sample is $\bar{y}_t = n^{-1}\Sigma y_{it} = \bar{\mu}' + \bar{d}_t = \bar{\mu} + \bar{\beta} + \bar{d}_t$, where the averages are for the n elements in the sample. The overall expectation of $\bar{y}_t$ is taken over both samples (s) and trials (t). Thus $E(\bar{y}_t) = E_s E_t(\bar{y}_t \mid s) = \mu + \beta$, where μ is the true population mean and β is the population average of the individual response biases. The standard computations of the accuracy of survey estimates make no allowance for bias.

The variance of $\bar{y}_t$ is similarly given by $V_s E_t(\bar{y}_t \mid s) + E_s V_t(\bar{y}_t \mid s)$. The first term here is $V_s(\bar{\mu}')$ that conforms to standard design-based theory for the sample design employed, but that relates to the μ_i' rather than the true μ_i. The second term reflects the individual response variances and possibly covariances. Assuming that d_{it} and d_{jt} are uncorrelated, this second term reduces to $n^{-1}\bar{\sigma}^2$ where $\bar{\sigma}^2 = N^{-1}\Sigma\sigma_i^2$.

Measurement error variance is sometimes reflected in standard variance estimates, but this is not always so. If measurement errors are uncorrelated and the finite population correction (fpc) is negligible, then measurement error variance is automatically included. However, if the fpc is non-negligible, measurement error variance is not fully covered. In the limit, with a complete census with $n = N$, the standard variance estimate for a sample mean is zero, whereas the response error gives a variance of $N^{-1}\bar{\sigma}^2$.

In practice, measurement errors are often correlated for sets of sampled elements as, for instance, when they occur from interviewer effects leading to interviewer variance. In this case, correlated measurement variance is generally not covered by standard variance estimates. However, even here there are exceptions. For instance, interviewer variance is covered in a multistage design with a negligible fpc at the first stage if each interviewer collects data in only one PSU.

In summary, standard design-based theory does not adequately address measurement errors. It does not deal with the effect of measurement bias on survey estimates or their measures of accuracy and it often does not properly reflect measurement variance in assessing the precision of survey estimates.

2.5. Analytic uses of survey data

Much of the literature on design-based inference relates to the estimation of descriptive measures for the finite population. However, survey data are also widely used for analytic purposes. Deming (1950, 1953) identified the importance of the distinction between ''enumerative'' and ''analytic'' studies both for design and analysis, where enumerative studies — here termed descriptive studies — deal with questions of ''How many?'' and analytic studies deal with questions of ''Why?''

As a simple example of an analytic inference, consider a test of significance of the

difference between the means of two subgroups. The null hypothesis for such a test is that the population subgroup means are equal, but that hypothesis is clearly always false for the finite subgroup populations sampled; if a complete census were taken, the two subgroup means would not be identical (except in exceptional cases). In general, such a test is meaningless for the finite population. A solution to this problem is to view the finite population as a random realization of a process that generates the population, or as a random sample from a superpopulation (Deming and Stephan 1941). The inference then is made to the superpopulation.

With the superpopulation approach, the sample can be viewed as a two-phase sample, a first-phase sample that produces the finite population and a second-phase sample that produces the survey sample. If the survey produces an estimate of the difference in sample means of d that is unbiased for the finite population difference D, it follows that d is also unbiased for δ, the difference in the means in the superpopulation since $E(d) = E_1E_2(d) = E_1(D) = \delta$. The variance of d is:

$$V(d) = E_1V_2(d) + V_1E_2(d) = E_1V_2(d) + V_1(D)$$

As discussed above, if $v(d)$ is an unbiased estimator of $V_2(d)$, the variance of d in the finite population, it is also unbiased for $E_1V_2(d)$. Under the superpopulation model, $V_1(D)$ may be estimated by $v_1(d) = \hat{N}_1^{-1}s_1^2 + \hat{N}_2^{-1}s_2^2$, where $\hat{N}_1$ and $\hat{N}_2$ are the estimated finite population subgroup population sizes and s_1^2 and s_2^2 are the estimated element variances in these two subgroups.

As a simple case, consider the estimate $\bar{y}$ of the superpopulation mean μ from a simple random sample of size n. From above, its variance is:

$$V(\bar{y}) = E_1[(n^{-1} - N^{-1})S^2] + N^{-1}\sigma^2$$

where S^2 and σ^2 are the element variances in the finite population and the superpopulation, respectively. Since $E_1(S^2) = \sigma^2$, $V(\bar{y})$ reduces to $n^{-1}\sigma^2$, without the finite population correction. See Korn and Graubard (1994) for superpopulation inference with complex designs.

A further extension of the above line of reasoning occurs when survey data are used to try to identify causal mechanisms and the magnitude of causal effects. The inferential issues can be demonstrated by considering multiple regression models, models that are widely used for this purpose. Since the extensive literature on this topic is too complex to be thoroughly reviewed, only a few general observations will be made. See Pfeffermann (1996) for a more detailed treatment.

Consider a multiple regression equation developed to assess the effect of x_1 on y after controlling for a set of confounding variables $(x_2, \ldots, x_p)$. The superpopulation regression coefficient β_1 measures this effect under the assumption that all the relevant confounding variables are incorporated in the model. Many, probably most, social scientists carry out regression analyses of this type with survey data without regard to the survey weights or the complex sample design. They rely on the model assumptions for their analyses, including the assumption that any differences in selection probabilities and response propensities do not give rise to selection bias. Assuming homoskedasticity, the regression analysis is performed unweighted to yield estimates of the regression coefficients ($\hat{\boldsymbol{\beta}}$), and the variance of $\hat{\boldsymbol{\beta}}$ is computed using standard least squares theory.

In contrast, under the design-based approach the regression coefficients are estimated using survey weights and the complex sample design is taken into account in estimating the variances of these estimates. With this approach, the sample regression coefficients ($\boldsymbol{b}$) estimate the finite population parameters ($\boldsymbol{B}$) that would be obtained had a complete census been taken. Given the chosen form of the regression equation (that is, given the chosen set of predictors), the finite population parameters $\boldsymbol{B}$ are defined as the quantities that minimize the average squared residuals for that specific finite population (see, for example, Kish and Frankel 1974). As DuMouchel and Duncan (1983) note, this representation involves no model assumptions. It is simply a definition of the descriptive finite population parameters $\boldsymbol{B}$ given the chosen form of equation. Whether $\boldsymbol{B}$ is of interest, of course, depends on a sensible choice of predictors for the regression equation.

If the standard regression modeling assumptions are valid, then the census value B_1 estimates β_1, the causal influence of x_1 in the superpopulation. Thus b_1 is a consistent estimate of β_1. This raises the issue of whether $\hat{\beta}_1$ or b_1 should be used to estimate β_1. If the model assumptions hold, b_1 will be less precise than $\hat{\beta}_1$, and hence $\hat{\beta}_1$ is preferred. However, b_1 provides some protection against a misspecified model in a particular sense. As noted above, the regression equation estimated with regression coefficients $\boldsymbol{b}$ provides the best predictive equation of the specific form for the *given* finite population. Note, however, that this protection does not hold if the regression equation is to be applied to a different population (for example, the protection does not hold if the survey is conducted in Maryland and generalization to the U.S. is sought). No matter whether $\boldsymbol{b}$ or $\hat{\boldsymbol{\beta}}$ is used to estimate $\boldsymbol{\beta}$, the regression model needs to be carefully developed and evaluated if the focus of the analysis is on the interpretation of the regression coefficients (this is somewhat less crucial if the regression is to be used only for predictions).

The protection afforded by the design-based approach to regression analysis seems to be valuable in many cases. However, if it leads to estimated regression coefficients of inadequate precision, it may be necessary to rely fully on the model and its assumptions and use the estimates $\hat{\boldsymbol{\beta}}$ (see Korn and Graubard 1999 for further discussion). An alternative strategy, in line with the reasoning for small area estimates, would be to use a composite estimate that is a weighted combination of $\boldsymbol{b}$ and $\hat{\boldsymbol{\beta}}$. Note, however, that any sizeable difference between $\boldsymbol{b}$ and $\hat{\boldsymbol{\beta}}$ is problematic since it suggests model misspecification (see the example in DuMouchel and Duncan 1983).

A final comment in this section concerns the use of surveys for the evaluation of an intervention. For example, health care interventions may be introduced in several purposively chosen underserved areas with the aim of improving the health status of the populations of those areas. Evaluation of the interventions may be conducted by means of a before-after design in which the health statuses of the populations are established by surveys conducted in each area before and after the interventions. The change in health status of an area's population is then used to measure the effect of the intervention, usually after taking into account the corresponding change in a control population. Frequently the same sample size is used in all areas, irrespective of their population sizes, so that the effect can be ascertained with the same level of precision in each area. However, analysts often also want to combine the results obtained in different areas to increase the power of the analyses. One approach that is sometimes suggested is to pool the survey data for the different areas, using the survey weights in the combined analysis. This approach produces

valid results for the population that consists of the union of the area populations. However, that population is seldom of interest. Instead, a meta-analytic approach that analyses the individual surveys separately and then combines the results appears preferable for the intended inference to all such areas.

3. Model-Dependent Methods

Even though the design-based approach is generally applied for descriptive analyses of large-scale surveys, it cannot fully address all problems of making inferences from survey samples. This section considers the need for model-dependent methods in the areas of small area estimation, missing data, and variance estimation.

3.1. Small area estimation

As stated above, design-based inference is applicable for large-scale surveys with estimates based on large samples. However, even with large-scale surveys, the sample sizes for small domains may be too small to support design-based estimates of adequate precision. The increasing demand for estimates for small domains in recent years, particularly for small geographical domains, has led to substantial research on methods to produce model-dependent methods to satisfy that demand (see, for example, Ghosh and Rao 1994; Rao 1999, 2000, 2002). A common terminology is to describe model-dependent estimators as indirect estimators, in contrast to design-based direct estimators (Schaible 1996).

What is interpreted as a ''small area'' depends on sample size. Estimates are often wanted for states in the U.S., yet sample sizes are generally too small to produce direct estimates of adequate precision for most, if not all, states. Thus, states are often treated as small areas for which indirect estimates are produced. When county estimates are required from national surveys, they are almost always produced by indirect small area estimation methods. Indirect estimates are, in fact, produced for states and/or counties by a number of U.S. federal statistical agencies. See Schaible (1996) for an excellent description of the estimates at the time that report was written, the U.S. Census Bureau's web site http://www.census.gov/hhes/www/saipe/overview/html for a description of the Bureau's Small Area Income and Poverty Estimates (SAIPE) program, and Folsom and Judkins (1997) for a description of the methods used to make small area estimates of substance abuse. Hansen, Hurwitz, and Madow (1953, Vol. 1, pp. 483-486) describe an early application of model-dependent small area estimation methods to produce estimates of radio listening in more than 500 county areas from an interview survey conducted in 85 areas.

The distinction that is widely made to distinguish between direct and indirect estimators is that the former rely on sample data only for the small area and the time period in question whereas the latter ''borrow strength'' from data for other areas and/or time periods. Although this phraseology is appealing, I think that it fails to fully capture the essence of the distinction. Indirect estimators borrow strength through statistical models that make use of predictive auxiliary data at the small area level. The key to the ability to produce good indirect estimators is the availability of such auxiliary data, and this point needs to be emphasized. Another concern that I have with the above distinction is that it is

closely related to, but not identical to, the distinction between design-based and model-dependent estimators. For example, along the lines of the discussion in Section 2.1, the model-assisted design consistent regression estimator for a small area $\hat{Y}_{Ra} = \Sigma^{N_a} b_j X_{ji} + \Sigma^{n_a} e_i/\pi_i$ is classified as an indirect estimator if the regression coefficients b_j are estimated using some sample data that come from outside the area. Singh, Gambino, and Mantel (1994) term this type of estimator a modified indirect estimator. My preference is to equate direct estimators with design-based estimators and indirect estimators with model-dependent estimators, and these terms are so used in what follows.

From the design-based perspective, the approach to estimation for small areas is first to seek direct estimators of adequate precision, making full use of auxiliary information in a model-assisted way. When that approach fails, it becomes necessary to resort to indirect estimators that depend on statistical models. No attempt is made here to review the extensive literature on models for indirect estimation. Instead, this discussion is confined to some general observations about the subject, using the U.S. Census Bureau's SAIPE program for illustrative purposes.

The SAIPE program was introduced in the early 1990s to produce updated estimates of median household income and numbers of poor people in various age ranges for U.S. states (annually) and for counties (biennially). The focus here is on the estimates of the numbers of poor related children aged 5–17, which are needed for the distribution of about 7 billion USD each year for programs for educationally disadvantaged children under Title I of the Elementary and Secondary Education Act. Reauthorization of that act in 1994 also required the U.S. Census Bureau to produce these estimates for school districts. Given the importance of the county and school district estimates of poor school-age children, the 1994 act authorized a National Research Council panel to review the estimates to assess their suitability as the basis for the Title I allocations. The U.S. Census Bureau and that panel have conducted extensive evaluations of the SAIPE program's estimates of poor school-age children (see, for example, National Research Council 2000a,b). These evaluations bring out some of the requirements for producing good small area estimates, as discussed below.

The procedures used to produce the SAIPE state, county, and school district estimates are highly complex, but for present purposes a simplified account that ignores many of the complexities will suffice. In essence, the SAIPE state and county estimates are based on regression models in which the dependent variable is a direct estimate from the Current Population Survey (available for all states, but only for counties that have some CPS sampled households) and in which the independent variables are obtained from 1990 Census data, tax data, and food stamp data. The regression models have two error components: a model error that would occur had the model been fitted with the population value rather than the direct estimate as the dependent variable, and the sampling error in the direct estimate. The estimate for a given state from the regression model is then combined with the direct estimate for that state as a weighted average, with weights that are inversely proportional to estimates of model error and sampling error, respectively. The same procedure is applied to produce the county estimates from the county regression model and, where available, the county direct estimates. These composite estimates are the empirical best linear unbiased prediction (EBLUP) estimates. The school district estimates are obtained by a ''shares approach'' that partitions the county estimates of numbers of poor school-age

children between the school districts in proportion to the numbers of poor school-age children that the districts had in the 1990 Census.

The SAIPE estimates of poor school-age children illustrate well the importance of auxiliary data. The state and county models make use of Internal Revenue Service tax return data on child exemptions reported by families in poverty and on people receiving food stamps, both of which are available at state and county levels and are predictive of the dependent variables in the regressions. However, these variables cannot be used in a model for school districts because they cannot currently be geocoded sufficiently well at that level. The simplistic shares model for school districts is in fact used because of the lack of suitable auxiliary variables at the school district level. Data on free and reduced-price lunches under the National School Lunch Program are compiled at the school and school district levels, but they are not currently nationally available and there are concerns about the comparability of the data across the country.

In considering potential predictor variables for the regression models, comparability across areas is an important concern that needs careful examination. For example, the income and allowable deductions for food stamp eligibility are higher in Alaska and Hawaii, thus affecting comparability. In the more recent estimates, the U.S. Census Bureau has made adjustments to compensate for this fact.

In general, the literature on small area estimation methodology pays too little attention to the development of appropriate predictor variables from the available data sources. Careful construction of predictor variables can improve the effectiveness of small area models and avoid the distortion of the estimates for some areas. Thus, for example, the U.S. Census Bureau conducted research on how best to use the food stamp data in the state poverty models, and as a result modified the predictor variable to be a monthly average over a 12-month period centered on January 1 of the following year, to subtract those who received food stamps due to specific natural disasters, and to smooth outliers from a time-series analysis, in addition to the adjustments in Alaska and Hawaii noted above. As another example, the predictor variable ''Estimated population under age 21'' that was used in the county model for the first round of estimates was changed to the ''Estimated population under age 18'' in subsequent rounds as a result of evaluations conducted on the first round model. Efforts to develop and refine predictor variables should be an important component of any small area estimation program.

When small area estimates are to be used for such consequential purposes as fund allocation, the importance of thorough evaluation cannot be overstressed. These evaluations should include internal evaluations using a range of regression diagnostics at each round of production of estimates and, to the extent possible, external evaluations that compare the estimates with estimates or values obtained from other sources and likely from other time periods. See National Research Council (2000b) for a description of the extensive evaluations of the SAIPE estimates of poor school-age children that the U.S. Census Bureau and the panel conducted. Such evaluations need to be repeated at each round of production to check that the predictor variables still operate as they have in the past and that the regression model still works well. (There is, for example, the possibility that, as a result of welfare reform, food stamp data may operate differently in the future.)

It should be noted that small area estimation methods are not a panacea. Even after careful modeling efforts, the resulting estimates are often still likely to be subject to substantial

levels of error. This is, for example, the case with the county estimates of poor school-age children, even with the use of predictor variables that appear better than those available in some other applications of small area estimation methods. Also, in common with direct survey estimates, small area estimates are out-of-date to some degree. Indeed, they will be less timely than direct estimates if the auxiliary data take a long time to compile, and time also has to be allowed for the small area modeling and evaluation work.

Finally, I note an issue about the composite estimator that is used to combine the model estimate and the direct estimate. As noted above, in the SAIPE program, as in most small area applications, the weights assigned to the two estimates are chosen to maximize the precision of the composite estimator under the assumption that the model holds. The weights are based on estimates of sampling error variance and model error variance, and the serious difficulties in estimating these variances can lead to problems. In the SAIPE state model for poor school-age children, for example, the model error was estimated as zero for six of the first seven years for which the model was estimated. As a result, in these years the composite estimate reduces to total reliance on the model estimate for all states, even though many states have direct estimates of reasonable precision (see Bell 1999 for some ways that this problem may be addressed). As a more general issue, the use of the EBLUP estimates appears to be logically inconsistent with the design-based perspective on the use of models that I proposed earlier. From that perspective, models are used only to the extent necessary. If this perspective were applied rigorously, the model estimates would be assigned the minimum weights needed to produce the required levels of precision for the composite estimates. Even if this perspective is not fully applied, in view of the uncertainty about the model errors, it may still be appropriate to reduce the weights assigned to the model estimates below those employed in the EBLUP estimates.

3.2. Missing data

Missing data present an increasingly serious problem in survey research. Missing data may be usefully divided into four main categories: noncoverage, which occurs when elements in the target population are not included on the sampling frame and hence have no chance of selection for the sample; total, or unit nonresponse, when sampled elements fail to provide any survey responses; partial nonresponse, when sampled elements respond to an appreciable number of items but also fail to respond to many others (as occurs in panel and multi-phase surveys when sample elements respond at early waves, but not at later waves, of data collection); and item nonresponse, when respondents fail to provide acceptable answers for some items.

Whenever there are missing data, models are needed in the survey analysis. Even when the missing data are ignored and the analysis is simply applied to the respondents, there is an implicit model involved. That model employs a missing completely at random (MCAR) assumption that is generally untenable. The adjustment procedures that are widely used in an attempt to compensate for the various forms of missing data listed above employ alternative, hopefully more plausible, assumptions. Compensation for noncoverage is made by weighting adjustments to make weighted sample totals conform to population control totals. Compensation for total nonresponse and often for partial nonresponse is made by nonresponse weighting adjustments that make weighted respondent sample totals conform

to full sample totals. Compensation for item nonresponse, and sometimes partial nonresponse, is made by imputation, that is, by assigning values for the missing responses (see, for example, Brick and Kalton 1996). An important feature to note about all these compensation procedures is that they are general-purpose strategies, intended to enable analysts to perform any form of analysis. The models underlying these compensation procedures are developed to this end; a different model for missing data may be more suitable for a specific analysis. The following subsections briefly review the models used for noncoverage adjustments, nonresponse adjustments, and imputation.

3.2.1. Noncoverage adjustments

In its basic form, a noncoverage adjustment is a population weighting adjustment that uses the same procedures as poststratification. An external source is used to produce a cross-classified table of a set of auxiliary variables with population totals in the cells. Sample estimates of these totals — incorporating weights for unequal selection probabilities and adjustments to those weights for total nonresponse — are generated, and the weights of respondents within each cell are then inflated or deflated to make the resultant sample totals conform to the external control totals. Although discussed here primarily as a noncoverage adjustment, this form of adjustment can also compensate for total nonresponse by taking account of auxiliary variables not already covered by the nonresponse adjustments, and can serve to improve the precision of survey estimates associated with the auxiliary variables.

Absent problems of missing data, poststratification is used to improve the precision of the survey estimates. No modeling assumptions are needed. With noncoverage, this form of adjustment invokes the assumption that those noncovered are missing at random (MAR) within the cells. Thus, for example, in the U.S. Current Population Survey, weighted sample totals by age, race, and sex are made to conform to current population estimates. The most substantial adjustment required is for young black males: for instance, the coverage ratio — the inverse of the weighting adjustment — for the cell of black males aged 20–29 was 0.66 in early 1996 (U.S. Bureau of Labor Statistics and U.S. Census Bureau 2000). The adjustment makes the assumption that the sample in this cell can represent those missed. This assumption is clearly a dubious one, but it is probably preferable to the alternative assumption that those missed are represented by the total population (as is implied by not making the adjustment). An obvious course of action to produce a more realistic assumption is to seek to control for more auxiliary variables. However, the opportunities are generally severely limited by the shortage of matching control totals from auxiliary data. Also sample sizes in cells can become too small to produce sufficiently stable adjustments.

A variety of alternative calibration methods have been developed to address the small cell sample size problem and also the problem that sometimes the full population joint distribution of the auxiliary variables is not known (see, for example, Deville and Särndal 1992; Deville, Särndal, and Sautory 1993). These methods provide a means to control for more auxiliary variables, but in doing so employ additional assumptions. Thus, for example, a two-way raking of sample marginal totals to population marginal totals assumes that the missing data rate in a cell of the two-way table is the product of a row and column effect (Kalton and Maligalig 1991). This assumption is untestable without knowledge of the population joint distribution.

A general concern in making weighting adjustments for missing data is that the resultant weights may become highly variable, thus potentially reducing bias in some estimates at a cost of a large increase in variance. With the cell-by-cell reweighting approach, this problem is often handled by collapsing cells and/or trimming weights. When cell collapsing is used, the operation is usually done by a visual inspection to determine which cells should be combined based on their likely similarity in terms of the survey variables. With some calibration methods, restrictions can be placed on the upper and lower limits of the weights, thus automatically avoiding this problem. However, this automatic solution also involves a redistribution of the weighting adjustments that deserves examination to assess how well it works for the particular survey (Kalton and Flores-Cervantes 1998).

Calibration methods and automatically constrained weights are important developments that extend the possibilities for compensating for missing data and facilitate the application of such adjustments. However, they do involve modeling assumptions that, where possible, should be examined.

3.2.2. Nonresponse adjustments

Two different modeling approaches are used for total nonresponse. One approach treats the population as composed of two strata, one of respondents and the other of nonrespondents. Over conceptually repeated trials of the survey, this deterministic approach assumes that respondents always respond and nonrespondents never do so (see, for example, Cochran 1977). The other approach, known as the quasi-randomization approach, assumes that every population element has a probability of responding if sampled, say ϕ_i $(i = 1, 2, \ldots, N)$ (Oh and Scheuren 1983). Furthermore, this approach assumes that $\phi_i > 0$ for all elements in the population, a necessary but doubtful assumption. The two procedures lead to broadly similar prescriptions for developing weighting adjustments for total nonresponse. An attraction of the quasi-randomization approach is that it enables responding to be treated as another phase of sampling, albeit one with unknown selection probabilities that need to be estimated from models.

Failure to make nonresponse adjustments implicitly assumes that the nonrespondents are MCAR ($\phi_i = \phi$, a constant). Nonresponse adjustment procedures aim to improve on this assumption. The basic nonresponse adjustment method is similar to the cell-by-cell noncoverage adjustment method described above. The difference is that noncoverage adjustments need external control totals, whereas nonresponse adjustments are generally based on internal data in the sample. The requirement for an auxiliary variable to be used for nonresponse adjustments is that its value be known for both respondents and nonrespondents. Within each cell defined in terms of a set of such auxiliary variables, the nonresponse adjustment consists of inflating the weights of responding sample members by the inverse of the base-weighted response rate in the cell. By inflating the weights in this way, the respondents in a cell represent the nonrespondents in that cell.

Consider an estimate of the population mean incorporating this form of nonresponse adjustment, $\bar{y}_w$. Under the quasi-randomization approach, the bias of $\bar{y}_w$ is approximately

$$B(\bar{y}_w) \approx N^{-1} \Sigma\Sigma \bar{\phi}_c^{-1} (Y_{ci} - \bar{Y}_c)(\phi_{ci} - \bar{\phi}_c)$$

where ci indexes element i in adjustment cell c and $\bar{\phi}_c$ is the average response probability in cell c (Brick and Kalton 1996). Inspection of this equation shows that $\bar{y}_w$ is

approximately unbiased if Y_{ci} and ϕ_{ci} are uncorrelated, and in particular if the nonrespondents are MAR within each cell, i.e., when $\phi_{ci} = \bar{\phi}_c$. Indeed, when the MAR assumption holds, weighted estimates corresponding to unbiased sample estimates with complete response are approximately unbiased for their population parameters. This result holds for all survey estimates. Since surveys are generally multi-purpose, collecting data on many variables and producing numerous estimates, the strategy of seeking nonresponse adjustment cells for which the MAR assumption is plausible is usually the primary one adopted.

When a survey is designed to produce estimates from only one variable, or only a few highly related variables, an alternative strategy is to form adjustment cells that are homogeneous in the variable or variables. This strategy has the attraction of producing estimates that have lower variance than when the cells are formed on the basis of the MAR assumption (Little 1986). It seldom occurs, however, that a survey is concerned with only one variable or a few related variables. Thus this strategy is rarely applied in dealing with total nonresponse (but see below for its use with item nonresponse).

In the case of total nonresponse — as distinct from partial nonresponse — there are usually very few auxiliary variables for which information is available for the nonrespondents. For area samples, generally all that is known about the nonrespondents is the area where they are located (stratum, primary sampling unit, and segment), and the characteristics of the area (e.g., urban/rural). For random digit dialing telephone surveys, all that is known for many nonrespondents is their telephone exchange and the characteristics of that exchange. In this situation, weighting adjustments may be based on all the limited information available, thus obviating the need to make a choice of which auxiliary variables to use. The resulting adjustments may be useful in reducing nonresponse bias, but the assumption that they will eliminate bias in the survey estimates is a heroic one.

In contrast to the dearth of information available for total nonrespondents, there is often a wealth of information available for partial nonrespondents from the responses they did provide. For example, information about partial nonrespondents who drop out of a panel survey at the second wave is available from their responses at the first wave. The challenge here is to make the most effective use of all this auxiliary information in developing weighting adjustments for the partial nonrespondents. To meet this challenge, models may be developed to predict partial response status from the auxiliary variables using such techniques as logistic regression or classification trees (see, for example, Rizzo, Kalton, and Brick 1996).

Suppose that a logistic regression model is developed for this purpose. Then the predicted probability of being a respondent (versus being a partial nonrespondent) can be computed from the regression for each respondent, and the inverse of that probability — with minor adjustments — can be used as a weighting adjustment. As a simple example, consider the case with only two categorical auxiliary variables in the model (each represented by a set of indicator variables) and assume that no interaction terms are included because none were significant. In this case, the weights could be computed from the model as indicated or, alternatively, they could be computed for each cell of the two-way table. Even if the model were a valid one, these two procedures will yield somewhat different adjustments. How should a choice be made between them? The adjustments produced from the model are likely to be less variable and, as such, they will cause

less inflation to the variance of the survey estimates. However, the model provides a general prediction rather than reflecting the outcome in the specific sample. In line with the discussion of conditional inference in Section 2.3, I favor matching the adjustment to the specific sample. This position argues in principle for the cell-by-cell adjustment procedure, although the variance inflation effect needs to be considered.

3.2.3. Imputation

Nonresponse weighting adjustments and imputation are closely related, and indeed in many cases one can be converted into the other. Yet, there is also a major difference between the two procedures that affects the modeling assumptions that need to be made.

Consider first the similarities. With cell weighting, each nonrespondent's record in a cell could be divided into a set of records equal in number to the number of respondents in that cell, and each subrecord could be assigned the full record of responses from one of the respondents. The nonrespondents' weight would be allocated across the subrecords in proportion to the respondent's weights (Kalton 1983b). This imputation procedure is identical in its effect to a nonresponse weighting adjustment. In a number of imputation schemes — including the widely used hot deck method — a missing response to an item for one element (the recipient) is assigned the value from another element (the donor). Recipients and donors are matched in some way in terms of a set of auxiliary variables that are responses to other survey items. For univariate analyses of an item for which imputations have been performed, such imputation schemes are equivalent to adding the weight of the recipient to that of the donor and dropping the recipient from the analysis.

The major difference between a nonresponse weighting adjustment and imputation is that the former compensates for nonresponse to all the survey items in a single operation whereas the latter is item specific. With weighting adjustments, the full set of responses of selected respondents is substituted for the full set of missing responses for the total nonrespondents. With imputation, a value is assigned for an item nonresponse in a respondent's record, with the responses to other items in the record being those provided by that respondent. Weighting adjustments thus maintain the associations between the survey variables that are present in the respondents' records, but imputation may distort the associations.

Imputation procedures use auxiliary variables to try to satisfy two objectives: to satisfy the MAR assumption and to preserve the associations between all the variables in the data set. The latter objective is of key concern because imputation is intended to be a multipurpose solution that will enable analysts to conduct whatever analyses they choose, and most analyses involve interrelationships between variables.

All imputation procedures use a model of some type and hot deck imputation is no exception. In many cases that model may be represented by a multiple regression equation that predicts the values of the variable to be imputed, y, from a set of auxiliary variables formed from responses to other survey items, $\boldsymbol{x}$ (Kalton and Kasprzyk 1986). The value imputed for a missing response is generally obtained by adding a residual to the regression prediction in order to preserve the variability in the distribution of y. Hot deck imputation is of this form. It employs a regression equation in which the predictors $\boldsymbol{x}$ are indicator variables that index the hot deck cells and the residuals are taken from respondents in those cells. The outcome of this process is simply to assign the value from a donor to the

recipient in a cell. A special case of imputation occurs when the regression model provides a perfect fit to the respondent data (e.g., all the y-values are the same within each hot deck cell). If the MAR assumption holds, the imputation becomes a deductive one that could be treated as an edit, with the missing responses being deduced from other responses.

Corresponding to the two approaches for forming cells for weighting adjustments, there are two possible approaches to forming imputation cells for matching donors and recipients for hot deck imputation. One approach is to attempt to construct cells in terms of the auxiliary variables such that the item nonresponses are MAR, as is generally done in forming weighting adjustment cells. The other is to form cells such that the y-values are homogeneous within cells. The latter approach is of primary importance with imputation because it addresses the objective of maintaining the associations in the data set.

In general, the association between y and x is preserved if x is used as an auxiliary variable in making imputations for y, but the association is attenuated towards zero if x is not so used. In the context of a hot deck imputation scheme in which x is always reported and y is randomly assigned from donors within a cell, the expectation of the conditional covariance between x and y in a cell (i.e., the ''within covariance'') is zero, where the expectation is over the random assignment of donors. As a result, for the imputed cases, the only contribution to the total covariance of x and y comes from the ''between covariance'' term in the analysis of covariance decomposition, that is, from the covariance of the cell means of x and y. Thus the covariance of x and y is incorrectly estimated unless x and y have a true conditional covariance of zero within imputation cells. A conditional covariance of zero is achieved by using x as an auxiliary variable to form the imputation cells such that $x = k_c$, a constant in each cell. Thus, a major consideration in forming imputation cells is to preserve covariances by choosing auxiliary variables that are associated with y, or equivalently by choosing auxiliary variables to form cells that are homogeneous in y.

Imputation focuses on incorporating all the major predictors of y in the imputation model in order to give effective imputations that maintain associations. In practice, of course, there are limits on the number of variables that can be included and on the efforts that can be devoted to developing imputation models. As a result, some associations will be attenuated. The magnitude of the attenuation may not be of serious concern if the level of missing data is low or if the association itself is low and hence probably of little substantive interest. However, analysts need to recognize that imputation models are imperfect and that they should conduct their analyses of imputed data sets with this in mind. In addition to concerns about the effects of imputation on associations, its effect on the variances of the survey estimates also needs to be taken into account. A considerable amount of research is underway on methods for estimating variances from imputed data sets, and here also it needs to be recognized that these methods depend on models (Kim 2001).

A particular application of imputation occurs in statistical matching, as is widely used in microsimulation modeling. With statistical matching, there are two (or more) data sets, with one containing variables ($\boldsymbol{X}$, $\boldsymbol{Y}$) and the other ($\boldsymbol{X}$, $\boldsymbol{Z}$), and interest centers on analyzing ($\boldsymbol{X}$, $\boldsymbol{Y}$, $\boldsymbol{Z}$) and particularly ($\boldsymbol{Y}$, $\boldsymbol{Z}$). In one version of statistical matching, the $\boldsymbol{Z}$ variables are imputed to the ($\boldsymbol{X}$, $\boldsymbol{Y}$) data set using the $\boldsymbol{X}$ variables as the auxiliary variables. In general statistical matching employs the assumption that $\boldsymbol{Y}$ and $\boldsymbol{Z}$ are conditionally independent given $\boldsymbol{X}$, as discussed above (see, for example, Rodgers 1984). With $\boldsymbol{X}$ often not providing powerful predictors of $\boldsymbol{Z}$ (or $\boldsymbol{Y}$) and with 100 percent of the $\boldsymbol{Z}$ values being imputed, there

is a very real danger that statistical matching will produce seriously biased estimates of the ($\boldsymbol{Y}$, $\boldsymbol{Z}$) associations.

3.4. Variance estimation

In addition to their use in forming survey estimates, models are also often used for estimating the standard errors of those estimates. These models are usually in the form of generalized variance functions (GVFs) relating the relvariance (R_i) of an estimate of a proportion or total (x_i) to the estimate itself in the form $R_i = \alpha + \beta x_i^{-1}$ (Wolter 1985, Ch. 5) or in terms of a model based on design effects (Kish 1965, Ch. 14). Such models are used for four main purposes: (1) they provide standard error estimates for which direct computations have not been made; (2) they give a concise presentation of sampling errors for survey reports; (3) the model standard errors may be preferable to those computed directly in that they have lower standard errors; and (4) the models can help in planning future sample designs (Kalton 1977).

The last of these four purposes fits in the category of the use of models for planning future surveys, as referred to earlier. The other three relate to the use of models in analysis, the focus here. Until recently, Purpose (1) was the main thrust for modeling sampling errors. Computations of sampling errors for estimates based on complex sample designs are complicated and were difficult to perform without the computing power and sampling error software that are now available. For this reason, sampling errors were calculated for only a few estimates. Based on results of these calculations, models were developed to predict sampling errors for other survey estimates. Although sampling errors can now be computed readily for many estimates, this reasoning still appears to be prevalent with a number of surveys. I believe that current computing capabilities should be used to calculate direct estimates of sampling errors for many survey estimates in all surveys, even if models are to be used for Purposes (2) and (3).

Concise presentation of sampling errors in survey reports is a challenge, and providing sampling error models from which the analyst can derive sampling errors is one solution. However, this approach has the unattractive feature that the analyst has to perform calculations to obtain the sampling errors for the estimates of interest. Also, the model estimates of sampling errors are often reasonable for some estimates, but poor for others. With the move towards presenting estimates from federal surveys on CD-ROMs and the Internet, other approaches may replace the need for models for presentation purposes. For instance, a link may be incorporated with each estimate in a table to enable the analyst to readily obtain the estimate's sampling error from a pop-up window.

Links that provide direct sampling error estimates for entries in table cells may replace model estimators. However, these links still do not fully satisfy all user needs. As a simple example, an analyst may want to obtain the sampling error of the difference between two cell entries (say, means or percentages), which is not readily available from links or most sampling error models. For some survey data sets available on the Internet, in the near future it is likely that users will be able to request both specific estimates and their associated sampling errors, thus considerably enhancing the provision of direct estimates of sampling errors to users.

The future should see a shift away from the use of sampling error for Purposes (1) and

(2), but Purpose (3) will remain. Like the survey estimates themselves, their variances are sample estimates that are subject to sampling error. This feature should be taken into account in the analysis. Thus, for example, if the number of degrees of freedom for the variance estimate of a sample mean is small, the confidence interval for the population mean should be obtained from the t distribution rather than the normal distribution. With multi-stage samples, the number of degrees of freedom for a variance estimate depends on the number of primary sampling units (PSUs) and the stratification. Some surveys have few PSUs, in which case the variance estimates will be imprecise even for national estimates (Kalton 1995). For other surveys, national variance estimates may be reasonably precise, but analyses relating to subsets of PSUs, such as regional analyses or analyses of rural populations, will likely be based on few PSUs and hence have imprecise variance estimates, even when the direct survey estimates themselves have adequate precision. Particularly in this latter case, model-based variance estimates may be preferable to direct variance estimates.

Finally, mention should be made of the need for models in variance estimation with nonmeasurable probability sample designs. Model-based methods are required for designs with a single primary selection per stratum and for designs using controlled selection. The collapsed strata technique is often used in these cases; it leads to an overestimation of variance unless a model of equal means in the collapsed strata holds. Collapsing can be beneficial even with measurable designs if they have few primary selections. In this case, the more precise variance estimates that come with the use of collapsed strata can sometimes outweigh the bias associated with collapsing (Rust and Kalton 1987).

4. Concluding Remarks

As indicated in the introduction, I find the design-based approach to inference generally suitable for most large-scale surveys, and particularly for household surveys for which the distributions of survey variables are generally well behaved. This is not to argue that other schools of thought are not also acceptable. In some cases, a different approach may yield an alternative and informative insight, with the possibility of some improvement in current methods. However, the design-based approach has been practiced by many thoughtful survey statisticians for many years and it has withstood the test of time. Thus, if a different approach produces results that are markedly at variance with the current ones, it probably makes sense to critically question the validity of the new results.

As detailed throughout the article the design-based approach has its limitations. One aspect is its focus on a clearly defined population that generally does not correspond to the population of concern to analysts. In this case, analysts are left to make their own subjective inferences from the surveyed population to the population of interest to them.

Another limitation of the design-based approach is that it does not handle the imperfections in the sample caused by missing data. Necessarily, models are required for this purpose. With the levels of noncoverage and total nonresponse that are now common, this limitation is a significant one, affecting a substantial proportion of the sample. The models that are employed to handle noncoverage and total nonresponse require suitable external data and data on the nonrespondents, respectively. With the limited amount of such data

available, these models may be beneficial, but they cannot be expected to compensate fully for the missing data.

Design-based methods are also limited in their ability to produce small area estimates of adequate precision, even with the use of auxiliary information in model-assisted estimation procedures. Model-dependent methods are therefore often required in such situations. Here, again, the effectiveness of the model-dependent approach depends on the availability of effective auxiliary data. Also, careful attention needs to be paid to the development of an appropriate model and its evaluation.

5. References

Alexander, C.H. (1994). Discussion of: Smith, T.M.F., Sample Surveys 1975-1990; An Age of Reconciliation? International Statistical Review, 62, 22–28.

Bell, W.R. (1999). Accounting for Uncertainty about Variances in Small Area Estimation. Bulletin of the International Statistical Institute, 52nd Session, Invited Papers, Tome LVIII (2), 57–60.

Bell, W.R. and Hillmer, S.C. (1990). The Time Series Approach to Estimation for Repeated Surveys. Survey Methodology, 16, 195–215.

Box, G.E.P. (1979). Robustness in the Strategy of Scientific Model Building. In Robustness in Statistics, R.L. Launer and G.N. Wilkinson eds. New York: Academic Press.

Brick, J.M. and Kalton, G. (1996). Handling Missing Data in Survey Research. Statistical Methods in Medical Research, 5, 215–238.

Chromy, J.R. (1998). The Effects of Finite Sampling on State Assessment Sample Requirements. NAEP Validity Studies (NVP). Palo Alto, CA: American Institutes for Research.

Cochran, W.G. (1977). Sampling Techniques. New York: Wiley.

Deming, W.E. (1950). Some Theory of Sampling. New York: Wiley.

Deming, W.E. (1953). On the Distinction Between Enumerative and Analytic Surveys. Journal of the American Statistical Association, 48, 244–255.

Deming, W.E and Stephan, F.F. (1941). On the Interpretation of Censuses as Samples. Journal of the American Statistical Association, 36, 45–49.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association, 87, 376–382.

Deville, J.-C., Särndal, C.-E., and Sautory, O. (1993). Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association, 88, 1013–1020.

DuMouchel, W.H. and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. Journal of the American Statistical Association, 78, 535–543.

Folsom, R.E. and Judkins, D.R. (1997). Substance Abuse in States and Metropolitan Areas: Model Based Estimates from the 1991-1993 NHDSA — Methodology Report. Substance Abuse and Mental Health Services Administration, Rockville, MD.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal. Statistical Science, 9, 55–93.

Hansen, M.H., Hurwitz, W.N., and Bershad, M.A. (1961). Measurement Errors in Censuses and Surveys. Bulletin of the International Statistical Institute, 38(2), 359–374.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). Sample Survey Methods and Theory. Vol. I. Methods and Applications. Vol. II. Theory. New York: Wiley.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association, 78, 776–793.

Holt, D. and Smith, T.M.F. (1979). Post Stratification. Journal of the Royal Statistical Society, Series A, 142, 33–46.

Kalton, G. (l977). Practical Methods for Estimating Survey Sampling Errors. Bulletin of the International Statistical Institute, 47, 495–514.

Kalton, G. (l983a). Models in the Practice of Survey Sampling. International Statistical Review, 51, 175–188.

Kalton, G. (1983b). Compensating for Missing Survey Data. Ann Arbor: Institute for Social Research, University of Michigan.

Kalton, G. (1995). Variance Estimation with Few Degrees of Freedom. Bulletin of the International Statistical Institute, 56(4), 1642–1645.

Kalton, G. and Flores-Cervantes, I. (1998). Weighting Methods. In New Methods for Survey Research, A. Westlake, J. Martin, M. Rigg, and C. Skinner (eds.). Chesham, Bucks, U.K.: Association for Survey Computing, 77–92.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Survey Data. Survey Methodology, 12, 1–16.

Kalton, G. and Maligalig, D.S. (1991). A Comparison of Methods of Weighting Adjustment for Nonresponse. Proceedings of the U.S. Census Bureau 1991 Annual Research Conference, 409–428.

Kim, J.-K. (2001). Variance Estimation After Imputation. Survey Methodology, 27, 75–83.

Kish, L. (l965). Survey Sampling. New York: Wiley.

Kish, L. (1999). Cumulating/Combining Population Surveys. Survey Methodology, 25, 129–138.

Kish, L. and Frankel, M.R. (1974). Inference from Complex Surveys (With Discussion). Journal of the Royal Statistical Society, Series B, 36, 1–37.

Korn, E.L. and Graubard, B.I. (1994). Variance Estimation for Superpopulation Parameters: Should One Use With-Replacement Estimators? Proceedings of the Section on Survey Research Methods, American Statistical Association, 124–132.

Korn, E.L. and Graubard, B.I. (1999). Analysis of Health Surveys. New York: Wiley.

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157.

National Research Council (2000a). Small-Area Income and Poverty Estimates: Priorities for 2000 and Beyond. In Panel on Estimates of Poverty for Small Geographic Areas, C.F. Citro and G. Kalton (eds.). Committee on National Statistics. Washington, D.C.: National Academy Press.

National Research Council (2000b). Small-Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology. In Panel on Estimates of Poverty for Small Geographic Areas, C.F. Citro and G. Kalton (eds.). Committee on National Statistics. Washington, D.C.: National Academy Press.

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, Series A, 97, 558–606.

Oh, H.L. and Scheuren, F. (1983). Weighting Adjustments for Unit Nonresponse. In Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies, W.G. Madow, I. Olkin, and D. Rubin (eds.). New York: Academic Press, 143–184.

Pfeffermann, D. (1996). The Use of Sampling Weights for Survey Data Analysis. Statistical Methods in Medical Research, 5, 239–261.

Rao, J.N.K. (1985). Conditional Inference in Survey Sampling. Survey Methodology, 11, 15–31.

Rao, J.N.K. (1999). Some Recent Advances in Model-Based Small Area Estimation. Survey Methodology, 25, 175–186.

Rao, J.N.K. (2000). Statistical Methodology for Indirect Estimations in Small Areas. International Statistical Seminars, Eustat–The Basque Statistics Institute.

Rao, J.N.K. (2002). Small Area Estimation. New York: Wiley.

Rizzo, L., Kalton, G., and Brick, J.M. (1996). A Comparison of Some Weighting Adjustment Methods for Panel Nonresponse. Survey Methodology, 22, 43–53.

Rodgers, W. L. (1984). An Evaluation of Statistical Matching. Journal of Business and Economic Statistics, 2, 91–102.

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

Royall, R.M. (1994). Discussion of: Smith, T.M.F., Sample Surveys 1975-1990; An Age of Reconciliation? International Statistical Review, 62, 19–21.

Royall, R.M. and Cumberland, W.G. (1981). An Empirical Study of the Ratio Estimator and Its Variance. Journal of the American Statistical Association, 76, 66–88.

Rust, K. and Kalton, G. (1987). Strategies for Collapsing Strata for Variance Estimation. Journal of Official Statistics, 3, 69–81.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992). Model Assisted Survey Sampling. New York: Springer-Verlag.

Schaible, W.L. (ed.) (1996). Indirect Estimators in U.S. Federal Programs. New York: Springer-Verlag.

Scott, A.J., Smith, T.M.F., and Jones, R.G. (1977). The Application of Time Series Methods to the Analysis of Repeated Surveys. International Statistical Review, 45, 13–28.

Singh, M.P., Gambino, J., and Mantel, H.J. (1994). Issues and Strategies for Small Area Data. Survey Methodology, 20, 3–22.

Smith, T.M.F. (1976). The Foundations of Survey Sampling: a Review. Journal of the Royal Statistical Society, Series A, 139, 183–204.

Smith, T.M.F. (1994). Sample Surveys 1975-1990; An Age of Reconciliation? International Statistical Review, 62, 5–34.

U.S. Bureau of Labor Statistics and U.S. Census Bureau (2000). Current Population Survey. Design and Methodology. Technical Paper 63. Washington, DC: Economics and Statistics Administration.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference. A Prediction Approach. New York: Wiley.

Williams, W.H. (1964). Sample Selection and the Choice of Estimator in Two-Way Stratified Populations. Journal of the American Statistical Association, 59, 1054–1062.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Zhang, L.-C. (2000). Post-Stratification and Calibration — A Synthesis. The American Statistician, 54, 178–184.

Received March 2002