37

Survey sampling: Past controversies, current orthodoxy, and future paradigms

Roderick J.A. Little
Department of Biostatistics
University of Michigan, Ann Arbor, MI

37.1 Introduction

My contribution to this historic celebration of the COPSS concerns the field of survey sampling, its history and development since the seminal paper by Neyman (1934), current orthodoxy, and a possible direction for the future. Many encounter survey sampling through the dull prism of moment calculations, but I have always found the subject fascinating. In my first sampling course, I remember being puzzled by the different forms of weighting in regression — by the inverse of the probability of selection, or by the inverse of the residual variance (Brewer and Mellor, 1973). If they were different, which was right? My early practical exposure was at the World Fertility Survey, where I learnt some real-world statistics, and where the sampling guru was one of the giants in the field, Leslie Kish (Kish et al., 1976). Kish was proud that the developing countries in the project were more advanced than developed countries in publishing appropriate estimates of standard error that incorporated the sample design. Always engaging, he shared my love of western classical music and tolerated my model-based views. More recently, I spent time helping to set up a research directorate at the US Census Bureau, an agency that was at the forefront of advances in applied sampling under the leadership of Maurice Hansen.

What distinguishes survey sampling from other branches of statistics? The genesis of the subject is a simple and remarkable idea — by taking a simple random sample from a population, reasonably reliable estimates of population quantities can be obtained with quantifiable accuracy by sampling around a thousand units, whether the population size is ten thousand or twenty million. Simple random sampling is neither optimal or even practical in many real-world settings, and the main developments in the field concerned complex sample designs, which include features like stratification, weighting and

clustering. Another important aspect is its primary focus on finite population quantities rather than parameters of models. The practical concerns of how to do probability sampling in the real world, such as the availability of sampling frames, how to exploit administrative data, and alternative modes of survey administration, are an important part of the field; valuable, since currently statistical training tends to focus on estimation and inference, neglecting designs for collecting data.

Survey sampling is notable as the one field of statistics where the prevailing philosophy is design-based inference, with models playing a supporting role. The debates leading up to this current status quo were heated and fascinating, and I offer one view of them here. I also present my interpretation of the current status quo in survey sampling, what I see as its strengths and drawbacks, and an alternative compromise between design-based and model-based inference, Calibrated Bayes, which I find more satisfying.

The winds of change can be felt in this field right now. Robert Groves, a recent Director of the US Census Bureau, wrote:

> "For decades, the Census Bureau has created 'designed data' in contrast to 'organic data' [$\cdots$] What has changed is that the volume of organic data produced as auxiliary to the Internet and other systems now swamps the volume of designed data. In 2004 the monthly traffic on the internet exceeded 1 exabyte or 1 billion gigabytes. The risk of confusing data with information has grown exponentially... The challenge to the Census Bureau is to discover how to combine designed data with organic data, to produce resources with the most efficient information-to-data ratio. This means we need to learn how surveys and censuses can be designed to incorporate transaction data continuously produced by the internet and other systems in useful ways. Combining data sources to produce new information not contained in any single source is the future. I suspect that the biggest payoff will lie in new combinations of designed data and organic data, not in one type alone." (Groves, 2011)

I believe that the standard design-based statistical approach of taking a random sample of the target population and weighting the results up to the population is not adequate for this task. Tying together information from traditional surveys, administrative records, and other information gleaned from cyberspace to yield cost-effective and reliable estimates requires statistical modeling. However, robust models are needed that have good repeated sampling properties.

I now discuss two major controversies in survey sampling that shaped the current state of the field.

37.2 Probability or purposive sampling?

The first controversy concerns the utility of probability sampling itself. A probability sample is a sample where the selection probability of each of the samples that could be drawn is known, and each unit in the population has a non-zero chance of being selected. The basic form of probability sample is the simple random sample, where every possible sample of the chosen size n has the same chance of being selected.

When the distribution of some characteristics is known for the population, a measure of representativeness of a sample is how close the sample distribution of these characteristics matches the population distribution. With simple random sampling, the match may not be very good, because of chance fluctuations. Thus, samplers favored methods of purposive selection where samples were chosen to match distributions of population characteristics. The precise nature of purposive selection is often unclear; one form is quota sampling, where interviewers are given a quota for each category of a characteristic (such as age group) and told to sample until that quota is met.

In a landmark early paper on sampling, Neyman (1934) addressed the question of whether the method of probability sampling or purposive selection was better. His resolution was to advocate a method that gets the best of both worlds, stratified sampling. The population is classified into strata based on values of known characteristics, and then a random sample of size n_j is taken from stratum j, of size N_j. If $f_j = n_j/N_j$, the sampling fraction in stratum j, is a constant, an equal probability sample is obtained where the distribution of the characteristics in the sample matches the distribution of the population.

Stratified sampling was not new; see, e.g., Kaier (1897); but Neyman expanded its practical utility by allowing f_j to vary across strata, and weighting sampled cases by $1/f_j$. He proposed what is now known as Neyman allocation, which optimizes the allocations for given variances and costs of sampling within each strata. Neyman's paper expanded the practical utility of probability sampling, and spurred the development of other complex sample designs by Mahalanobis, Hansen, Cochran, Kish and others, greatly extending the practical feasibility and utility of probability sampling in practice. For example, a simple random sampling of people in a country is not feasible since a complete list of everyone in the population from which to sample is not available. Multistage sampling is needed to implement probability sampling in this setting.

There were dissenting views — simple random sampling (or equal probability sampling in general) is an all-purpose strategy for selecting units to achieve representativeness "on average" — it can be compared with randomized treatment allocation in clinical trials. However, statisticians seek optimal properties, and random sampling is very suboptimal for some specific purposes. For example, if the distribution of X is known in population, and the

objective is the slope of the linear regression of Y on X, it's obviously much more efficient to locate half the sample at each of the extreme values of X — this minimizes the variance of the least squares slope, achieving large gains of efficiency over equal probability sampling (Royall, 1970). But this is not a probability sample — units with intermediate values of X have zero chance of selection. Sampling the extremes of X does not allow checks of linearity, and lacks robustness. Royall argues that if this is a concern, choose sample sizes at intermediate values of X, rather than letting these sizes be determined by chance. The concept of *balanced sampling* due to Royall and Herson (1973) achieves robustness by matching moments of X in the sample and population. Even if sampling is random within categories of X, this is not probability sampling since there is no requirement that all values of X are included. Royall's work is persuasive, but random sampling has advantages in multipurpose surveys, since optimizing for one objective often comes at the expense of others.

Arguments over the utility of probability sampling continue to this day. A recent example concerns the design of the National Children's Study (Michael, 2008; Little, 2010), planned as the largest long-term study of children's health and development ever to be conducted in the US. The study plans to follow 100,000 children from before birth to early adulthood, together with their families and environment, defined broadly to include chemical, physical, behavioral, social, and cultural influences. Lively debates were waged over the relative merits of a national probability sample over a purposive sample from custom-chosen medical centers. In discussions, some still confused "probability sample" with "simple random sample." Probability sampling ideas won out, but pilot work on a probability sample of households did not produce enough births. The latest plan is a form of national probability sample based on hospitals and prenatal clinics.

An equal probability design is indicated by the all-purpose nature of the National Children's Study. However, a sample that includes high pollution sites has the potential to increase the variability of exposures, yielding more precise estimates of health effects of contaminants. A compromise with attractions is to do a combination — say choose 80% of the sample by equal probability methods, but retain 20% of the sample to ensure coverage of areas with high contaminant exposures.

37.3 Design-based or model-based inference?

The role of probability sampling relates to ideas about estimation and inference — how we analyze the data once we have it. Neyman (1934) is widely celebrated for introducing confidence intervals as an alternative to "inverse probability" for inference from a probability sample. This laid the foundation

for the "design-based approach" to survey inference, where population values are fixed and inferences are based on the randomization distribution in the selection of units. . . although Neyman never clearly states that he regards population values as fixed, and his references to Student's t distribution suggest that he had a distribution in mind. This leads me to the other topic of controversy, concerning design-based vs model-based inference; see, e.g., Smith (1976, 1994), Kish and Frankel (1974), Hansen et al. (1983), Kish (1995), and Chambers and Skinner (2003).

In design-based inference, population values are fixed, and inference is based on the probability distribution of sample selection. Obviously, this assumes that we have a probability sample (or "quasi-randomization," where we pretend that we have one). In model-based inference, survey variables are assumed to come from a statistical model. Probability sampling is not the basis for inference, but is valuable for making the sample selection ignorable; see Rubin (1976), Sugden and Smith (1984), and Gelman et al. (1995). There are two main variants of model-based inference: Superpopulation modeling, where frequentist inference is based on repeated samples from a "superpopulation" model; and Bayesian modeling, where fixed parameters in the superpopulation model are assigned a prior distribution, and inferences about finite population quantities or parameters are based on their posterior distributions. The argument about design-based or model-based inference is a fascinating component of the broader debate about frequentist versus Bayesian inference in general: Design-based inference is inherently frequentist, and the purest form of model-based inference is Bayes.

37.3.1 Design-based inference

More formally, for $i \in \{1, \ldots, N\}$, let y_i be the survey (or outcome) variable of the ith unit, where $N < \infty$ is the number of units in the population, and let I_i be the inclusion indicator variable of the ith unit. Let Z represent design information, such as stratum or cluster indicators. We consider inference about a finite population quantity $Q(Y, Z)$, for example the population total $Q(Y, Z) = y_1 + \cdots + y_N$, where $Y = (y_1, \ldots, y_N)$.

In the design-based or randomization approach as described, e.g., by Cochran (1977), inferences are based on the distribution of $I = (I_1, \ldots, I_N)$, and the outcome variables $y_1, \ldots, y_N$ are treated as fixed quantities. Inference involves (a) the choice of an estimator for Q, $\widehat{q} = \widehat{q}(Y_{\text{inc}}, I, Z)$, where Y_{inc} is the included part of Y; and (b) the choice of a variance estimator $\widehat{\nu} = \widehat{\nu}(Y_{\text{inc}}, I, Z)$ that is unbiased or approximately unbiased for the variance of $\widehat{q}$ with respect to the distribution of I. Inferences are then generally based on normal large-sample approximations. For example, a 95% confidence interval for Q is $\widehat{q} \pm 1.96\sqrt{\widehat{\nu}}$.

Estimators $\widehat{q}$ are chosen to have good design-based properties, such as

(a) *Design unbiasedness*: $\mathrm{E}(\widehat{q}|Y) = Q$, or

(b) *Design consistency*: $\widehat{q} \to Q$ as the sample size gets large (Brewer, 1979; Isaki and Fuller, 1982).

It is natural to seek an estimate that is design-efficient, in the sense of having minimal variance. However, it became clear that that kind of optimality is not possible without an assumed model (Horvitz and Thompson, 1952; Godambe, 1955). Design-unbiasedness tends to be too stringent, and design-consistency is a weak requirement (Firth and Bennett, 1998), leading to many choices of estimates; in practice, choices are motivated by implicit models, as discussed further below. I now give some basic examples of the design-based approach.

Example 1 (Estimate of a population mean from a simple random sample): Suppose the target of inference is the population mean $Q = \overline{Y} = (y_1 + \cdots + y_N)/N$ and we have a simple random sample of size n, $(y_1, \ldots, y_n)$. The usual unbiased estimator is the sample mean $\widehat{q} = \overline{y} = (y_1 + \cdots + y_n)/n$, which has sampling variance $V = (1 - n/N)S_y^2$, where S_y^2 is the population variance of Y. The estimated variance $\widehat{v}$ is obtained by replacing S_y^2 in V by its sample estimate s_y^2. A 95% confidence interval for $\overline{Y}$ is $\overline{y} \pm 1.96\sqrt{\widehat{v}}$.

Example 2 (Design weighting): Suppose the target of inference is the population total $T = (y_1 + \cdots + y_N)$, and we have a sample $(y_1, \ldots, y_n)$ where the ith unit is selected with probability π_i, $i \in \{1, \ldots, n\}$. Following Horvitz and Thompson (1952), an unbiased estimate of T is given by

$$\widehat{t}_{\mathrm{HT}} = \sum_{i=1}^{N} w_i y_i I_i,$$

where $w_i = 1/\pi_i$ is the sampling weight for unit i, namely the inverse of the probability of selection. Estimates of variance depend on the specifics of the design.

Example 3 (Estimating a population mean from a stratified random sample: For a stratified random sample with selection probability $\pi_j = n_j/N_j$ in stratum j, the Horvitz–Thompson estimator of the population mean $Q = \overline{Y} = (y_1 + \cdots + y_N)/N$ is the stratified mean, viz.

$$\overline{y}_{\mathrm{HT}} = \frac{1}{N} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \frac{N_j}{n_j} y_{ij} = \overline{y}_{\mathrm{st}} = \sum_{j=1}^{J} P_j \overline{y}_j,$$

where $P_j = N_j/N$ and $\overline{y}_j$ is the sample mean in stratum j. The corresponding

estimate of variance is

$$\widehat{v}_{\rm st} = \sum_{j=1}^{J} \left(1 - \frac{n_j}{N_j}\right) \frac{s_j^2}{n_j},$$

where s_j^2 is the sample variance of Y in stratum j. A corresponding 95% confidence interval for $\overline{Y}$ is $\overline{y}_{\rm st} \pm 1.96\sqrt{\widehat{v}_{\rm st}}$.

Example 4 (Estimating a population mean from a PPS sample): In applications such as establishment surveys or auditing, it is common to have measure of size X available for all units in the population. Since large units often contribute more to summaries of interest, it is efficient to sample them with higher probability. In particular, for probability proportional to size (PPS) sampling, unit i with size $X = x_i$ is sampled with probability cx_i, where c is chosen to yield the desired sample size; units that come in with certainty are sampled and removed from the pool. Simple methods of implementation are available from lists of population units, with cumulated ranges of size. The Horvitz–Thompson estimator

$$\widehat{t}_{\rm HT} = c \sum_{i=1}^{N} \frac{y_i}{x_i} I_i$$

is the standard estimator of the population total in this setting.

The Horvitz–Thompson estimator often works well in the context of PPS sampling, but it is dangerous to apply it to all situations. A useful guide is to ask when it yields sensible predictions of nonsampled values from a modeling perspective. A model corresponding to the HT estimator is the HT model

$$y_i \overset{\rm iid}{\sim} \mathcal{N}(\beta x_i, \sigma^2 x_i^2), \tag{37.1}$$

where $\mathcal{N}(\mu,\tau^2)$ denotes the Normal distribution with mean μ and variance τ^2. This leads to predictions $\widehat{\beta} x_i$, where

$$\widehat{\beta} = n^{-1} \sum_{i=1}^{N} \frac{y_i}{x_i} I_i,$$

so $\widehat{t}_{\rm HT} = \widehat{\beta}(x_1 + \cdots + x_N)$ is the result of using this model to predict the sampled and nonsampled values. If the HT model makes very little sense, the HT estimator and associated estimates of variance can perform poorly. The famous elephant example of Basu (1971) provides an extreme and comic illustration.

Models like the HT estimator often motivate the choice of estimator in the design-based approach. Another, more modern use of models is in model-assisted inference, where predictions from a model are adjusted to protect

against model misspecification. A common choice is the generalized regression (GREG) estimator, which for a total takes the form:

$$\widehat{t}_{\mathrm{GREG}} = \sum_{i=1}^{N} \widehat{y}_i + \sum_{i=1}^{N} \frac{y_i - \widehat{y}_i}{\pi_i},$$

where $\widehat{y}_i$ are predictions from a model; see, e.g., Särndal et al. (1992). This estimator is design-consistent whether or not the model is correctly specified, and foreshadows "doubly-robust" estimators in the mainline statistics literature.

37.3.2 Model-based inference

The model-based approach treats both $I = (I_1, \ldots, I_N)$ and $Y = (y_1, \ldots, y_N)$ as random variables. A model is assumed for the survey outcomes Y with underlying parameters θ, and this model is used to predict the nonsampled values in the population, and hence the finite population total. Inferences are based on the joint distribution of Y and I. Rubin (1976) and Sugden and Smith (1984) show that under probability sampling, inferences can be based on the distribution of Y alone, provided the design variables Z are conditioned in the model, and the distribution of I given Y is independent of the distribution of Y conditional on the survey design variables. In frequentist superpopulation modeling, the parameters θ are treated as fixed; see, e.g., Valliant et al. (2000). In Bayesian survey modeling, the parameters are assigned a prior distribution, and inferences for $Q(Y)$ are based on its posterior predictive distribution, given the sampled values; see, e.g., Ericson (1969), Binder (1982), Rubin (1987), Ghosh and Meeden (1997), Little (2004), Sedransk (2008), Fienberg (2011), and Little (2012). I now outline some Bayesian models for the examples discussed above.

Example 1 continued (Bayes inference for a population mean from a simple random sample): A basic model for simple random sampling is

$$y_i|\mu,\sigma^2 \overset{\text{iid}}{\sim} \mathcal{N}(\mu,\sigma^2),$$

with a Jeffreys' prior on the mean and variance $p(\mu, \log\sigma^2) =$ a constant. A routine application of Bayes theorem yields a t distribution for the posterior distribution of $\overline{Y}$, with mean $\overline{y}$, scale $s\sqrt{1-n/N}$ and degrees of freedom $n-1$. The 95% credibility interval is the same as the frequentist confidence interval above, except that the normal percentile, 1.96, is replaced by the t percentile, as is appropriate since the variance is estimated. Arguably this interval is superior to the normal interval even if the data is not normal, although better models might be developed for that situation.

Example 2 continued (Bayesian approaches to design weighting): Weighting of cases by the inverse of the probability of selection is not really a model-based tool, although (as in the next example) model-based estimates correspond to design-weighted estimators for some problems. Design weights are conceived more as covariates in a prediction model, as illustrated in Example 4 below.

Example 3 continued (Estimating a population mean from a stratified random sample): For a stratified random sample, the design variables Z consist of the stratum indicators, and conditioning on Z suggests that models need to have distinct stratum parameters. Adding a subscript j for stratum to the normal model for Example 1 leads to

$$y_i|\mu_j, \sigma_j^2 \overset{\text{ind}}{\sim} \mathcal{N}(\mu_j, \sigma_j^2),$$

with prior $p(\mu_j, \log \sigma_j^2) =$ a constant. The resulting posterior distribution recovers the stratified mean as the posterior mean and the stratified variance for the posterior variance, when the variances σ_j^2 are assumed known. Estimating the variances leads to the posterior distribution as a mixture of t distributions. Many variants of this basic normal model are possible.

Example 4 continued (Estimating a population mean from a PPS sample): The posterior mean from the HT model (37.1) is equivalent to the HT estimator, aside from finite population. Zhen and Little (2003) relax the linearity assumption of the mean structure, modeling the mean of Y given size X as a penalized spline; see also and Zheng and Little (2005). Simulations suggest that this model yields estimates of the total that have superior mean squared error than the HT estimator when the HT model is misspecified. Further, posterior credible intervals from the expanded model have better confidence coverage.

37.3.3 Strengths and weakness

A simplified overview of the two schools of inference is that weighting is a fundamental feature of design-based methods, with models playing a secondary role in guiding the choice of estimates and providing adjustments to increase precision. Model-based inference is much more focused on predicting nonsampled (or nonresponding) units with estimates of uncertainty. The model needs to reflect features of the design like stratification and clustering to limit the effects of model misspecification, as discussed further below. Here is my personal assessment of the strengths and weaknesses of the approaches.

The attraction of the design-based perspective is that it avoids direct dependence on a model for the population values. Models can help the choice of estimator, but the inference remains design-based, and hence somewhat nonparametric. Models introduce elements of subjectivity — all models are

wrong, so can we trust results? Design-based properties like design consistency are desirable since they apply regardless of the validity of a model. Computationally, weighting-based methods have attractions in that they can be applied uniformly to a set of outcomes, and to domain and cross-class means, whereas modeling needs more tailoring to these features.

A limitation of the design-based perspective is that inference is based on probability sampling, but true probability samples are harder and harder to come by. In the household sample setting, contact is harder — there are fewer telephone land-lines, and more barriers to telephonic contact; nonresponse is increasing, and face-to-face interviews are increasingly expensive. As Groves noted in the above-cited quote, a high proportion of available information is now not based on probability samples, but on ill-defined population frames.

Another limitation of design-based inference is that it is basically asymptotic, and provides limited tools for small samples, such as for small area estimation. The asymptotic nature leads to (in my opinion) too much emphasis on estimates and estimated standard errors, rather than obtaining intervals with good confidence coverage. This is reflected by the absence of t corrections for estimating the variances in Examples 1 and 3 above.

On a more theoretical level, design-based inference leads to ambiguities concerning what to condition on in the "reference set" for repeated sampling. The basic issue is whether to condition on ancillary statistics — if conditioning on ancillaries is taken seriously, it leads to the likelihood principle (Birnbaum, 1962), which design-based inference violates. Without a model for predicting non-sampled cases, the likelihood is basically uninformative, so approaches that follow the likelihood principle are doomed to failure.

As noted above, design-based inference is not explicitly model-based, but attempting design-based inference without any reference to implicit models is unwise. Models are needed in design-based approach, as in the "model-assisted" GREG estimator given above.

The strength of the model-based perspective is that it provides a flexible, unified approach for all survey problems — models can be developed for surveys that deal with frame, nonresponse and response errors, outliers, small area models, and combining information from diverse data sources. Adopting a modeling perspective moves survey sample inference closer to mainstream statistics, since other disciplines like econometrics, demography, public health, rely on statistical modeling. The Bayesian modeling requires specifying priors, but has that benefit that it is not asymptotic, and can provide better small-sample inferences. Probability sampling justified as making sampling mechanism ignorable, improving robustness.

The disadvantage of the model-based approach is more explicit dependence on the choice of model, which has subjective elements. Survey statisticians are generally conservative, and unwilling to trust modeling assumptions, given the consequences of lack of robustness to model misspecification. Developing good models requires thought and an understanding of the data, and models have the potential for more complex computations.

37.3.4 The design-model compromise

Emerging from the debate over design-based and model-based inference is the current consensus, which I have called the Design-Model Compromise (DMC); see Little (2012). Inference is design-based for aspects of surveys that are amenable to that approach, mainly inferences about descriptive statistics in large probability samples. These design-based approaches are often model assisted, using methods such as regression calibration to protect against model misspecification; see, e.g., Särndal et al. (1992). For problems where the design-based approach is infeasible or yields estimates with insufficient precision, such as small area estimation or survey nonresponse, a model-based approach is adopted. The DMC approach is pragmatic, and attempts to exploit the strengths of both inferential philosophies. However, it lacks a cohesive overarching philosophy, involving a degree of "inferential schizophrenia" (Little, 2012).

I give two examples of "inferential schizophrenia." More discussion and other examples are given in Little (2012). Statistical agencies like the US Census Bureau have statistical standards that are generally written from a design-based viewpoint, but researchers from social science disciplines like economics are trained to build models. This dichotomy leads to friction when social scientists are asked to conform to a philosophy they view as alien. Social science models need to incorporate aspects like clustering and stratification to yield robust inferences, and addressing this seems more likely to be successful from a shared modeling perspective.

Another example is that the current paradigm generally employs direct design-based estimates in large samples, and model-based estimates in small samples. Presumably there is some threshold sample size where one is design based for larger samples and model based for smaller samples. This leads to inconsistency, and ad-hoc methods are needed to match direct and model estimates at different levels of aggregation. Estimates of precision are less easily reconciled, since confidence intervals from the model tend to be smaller than direct estimates because the estimates "borrow strength." Thus, it is quite possible for a confidence interval for a direct estimate to be wider than a confidence interval for a model estimate based on a smaller sample size, contradicting the notion that uncertainty decreases as information increases.

37.4 A unified framework: Calibrated Bayes

Since a comprehensive approach to survey inference requires models, a unified theory has to be model-based. I have argued (Little, 2012) that the appropriate framework is calibrated Bayes inference (Box, 1980; Rubin, 1984; Little, 2006), where inferences are Bayesian, but under models that yield inferences with

good design-based properties; in other words, Bayesian credibility intervals when assessed as confidence intervals in repeated sampling should have close to nominal coverage. For surveys, good calibration requires that Bayes models should incorporate sample design features such as weighting, stratification and clustering. Weighting and stratification is captured by included weights and stratifying variables as covariates in the prediction model; see, e.g., Gelman (2007). Clustering is captured by Bayesian hierarchical models, with clusters as random effects. Prior distributions are generally weakly informative, so that the likelihood dominates the posterior distribution.

Why do I favor Bayes over frequentist superpopulation modeling? Theoretically, Bayes has attractive properties if the model is well specified, and putting weakly informative prior distributions over parameters tends to propagate uncertainty in estimating these parameters, yielding better frequentist confidence coverage than procedures that fix parameters at their estimates. The penalized spline model in Example 4 above is one example of a calibrated Bayes approach, and others are given in Little (2012). Here is one more concluding example.

Example 5 (Calibrated Bayes modeling for stratified sampling with a size covariate): A common model for estimating a population mean of a variable Y from a simple random sample $(y_1, \ldots, y_n)$, with a size variable X measured for all units in the population, is the simple ratio model

$$y_i|x_i, \mu, \sigma^2 \overset{\text{ind}}{\sim} \mathcal{N}(\beta x_i, \sigma^2 x_i),$$

for which predictions yield the ratio estimator $\overline{y}_{\text{rat}} = \overline{X} \times \overline{y}/\overline{x}$, where $\overline{y}$ and $\overline{x}$ are sample means of Y and X and $\overline{X}$ is the population mean of X. Hansen et al. (1983) suggest that this model is deficient when the sample is selected by disproportionate stratified sampling, yielding biased inferences under relatively minor deviations from the model. From a calibrated Bayes perspective, the simple ratio model does not appropriately reflect the sample design. An alternative model that does this is the separate ratio model

$$y_i|x_i, z_i = j, \mu_j, \sigma_j^2 \overset{\text{ind}}{\sim} \mathcal{N}(\beta_j x_i, \sigma_j^2 x_i),$$

where $z_i = j$ indicates stratum j. Predictions from this model lead to the separate ratio estimator

$$\overline{y}_{\text{sep}} = \sum_{j=1}^{J} \frac{\overline{y}_j}{\overline{x}_j} P_j \overline{X}_j,$$

where P_j is the proportion of the population in stratum j. This estimator can be unstable if sample sizes in one or more strata are small. A Bayesian modification is to treat the slopes β_j as $\mathcal{N}(\beta, \tau^2)$, which smooths the estimate towards something close to the simple ratio estimate. Adding prior distributions for the variance components provides Bayesian inferences that incorporate errors for estimating the variances, and also allows smoothing of the stratum-specific variances.

37.5 Conclusions

I am a strong advocate of probability sampling, which has evolved into a flexible and objective design tool. However, probability samples are increasingly hard to achieve, and the strict design-based view of survey inference is too restrictive to handle all situations. Modeling is much more flexible, but models need to be carefully considered, since poorly chosen models lead to poor inferences. The current design-model compromise is pragmatic, but lacks a coherent unifying principle. Calibrated Bayes provides a unified perspective that blends design-based and model-based ideas. I look forward to further development of this approach, leading to more general acceptance among survey practitioners. More readily-accessible and general software is one area of need.

Hopefully this brief traverse of survey sampling in the last eighty years has piqued your interest. It will be interesting to see how the field of survey sampling evolves in the next eighty years of the existence of COPSS.

Acknowledgements

This work was supported as part of an Interagency Personnel Agreement with the US Census Bureau. The views expressed on statistical, methodological, technical, or operational issues are those of the author and not necessarily those of the US Census Bureau.

References

Basu, D. (1971). An essay on the logical foundations of survey sampling, part I (with discussion). In *Foundations of Statistical Inference* (V.P. Godambe and D.A. Sprott, Eds.). Holt, Rinehart and Winston, Toronto, pp. 203–242.

Binder, D.A. (1982). Non-parametric Bayesian models for samples from finite populations. *Journal of the Royal Statistical Society*, *Series B*, 44:388–393.

Birnbaum, A. (1962). On the foundations of statistical inference (with discussion). *Journal of the American Statistical Association*, 57:269–326.

Box, G.E.P. (1980). Sampling and Bayes inference in scientific modelling and robustness (with discussion). *Journal of the Royal Statistical Society, Series A*, 143:383–430.

Brewer, K.R.W. (1979). A class of robust sampling designs for large-scale surveys. *Journal of the American Statistical Association*, 74:911–915.

Brewer, K.R.W. and Mellor, R.W. (1973). The effect of sample structure on analytical surveys. *Australian Journal of Statistics*, 15:145–152.

Chambers, R.L. and Skinner, C.J. (2003). *Analysis of Survey Data*. Wiley, New York.

Cochran, W.G. (1977). *Sampling Techniques*, 3rd edition. Wiley, New York.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *Journal of the Royal Statistical Society, Series B*, 31:195–233.

Fienberg, S.E. (2011). Bayesian models and methods in public policy and government settings. *Statistical Science*, 26:212–226.

Firth, D. and Bennett, K.E. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society, Series B*, 60:3–21.

Gelman, A. (2007). Struggles with survey weighting and regression modeling (with discussion). *Statistical Science*, 22:153–164.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (1995). *Bayesian Data Analysis*. Chapman & Hall, London.

Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population Sampling*. Chapman & Hall, London.

Godambe, V.P. (1955). A unified theory of sampling from finite populations. *Journal of the Royal Statistical Society, Series B*, 17:269–278.

Groves, R.M. (2011) The future of producing social and economic statistical information, Part I. *Director's Blog*, www.census.gov, September 8, 2011. US Census Bureau, Department of Commerce, Washington DC.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An evaluation of model-dependent and probability-sampling inferences in sample surveys (with discussion). *Journal of the American Statistical Association*, 78:776–793.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–685.

Isaki, C.T. and Fuller, W.A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77:89–96.

Kaier, A.N. (1897). *The Representative Method of Statistical Surveys* [1976, English translation of the original Norwegian]. Statistics Norway, Oslo, Norway.

Kish, L. (1995). The hundred years' wars of survey sampling. *Statistics in Transition*, 2:813–830. [Reproduced as Chapter 1 of *Leslie Kish: Selected Papers* (G. Kalton and S. Heeringa, Eds.). Wiley, New York, 2003].

Kish, L. and Frankel, M.R. (1974). Inferences from complex samples (with discussion). *Journal of the Royal Statistical Society*, *Series B*, 36:1–37.

Kish, L., Groves, L.R., and Krotki, K.P. (1976). Standard errors from fertility surveys. *World Fertility Survey Occasional Paper 17*, International Statistical Institute, The Hague, Netherlands.

Little, R.J.A. (2004). To model or not to model? Competing modes of inference for finite population sampling. *Journal of the American Statistical Association*, 99:546–556.

Little, R.J.A. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, 60:213–223.

Little, R.J.A. (2010). Discussion of articles on the design of the National Children's Study. *Statistics in Medicine*, 29:1388–1390.

Little, R.J.A. (2012). Calibrated Bayes: An alternative inferential paradigm for official statistics (with discussion). *Journal of Official Statistics*, 28:309–372.

Michael, R.T. and O'Muircheartaigh, C.A. (2008). Design priorities and disciplinary perspectives: the case of the US National Children's Study. *Journal of the Royal Statistical Society*, *Series A*, 171:465–480.

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*, 97:558–606.

Rao, J.N.K. (2011). Impact of frequentist and Bayesian methods on survey sampling practice: A selective appraisal. *Statistical Science*, 26:240–256.

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

Royall, R.M. and Herson, J.H. (1973). Robust estimation in finite populations, I and II. *Journal of the American Statistical Association*, 68:880–893.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 53:581–592.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12:1151–1172.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

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

Sedransk, J. (2008). Assessing the value of Bayesian methods for inference about finite population quantities. *Journal of Official Statistics*, 24:495–506.

Smith, T.M.F. (1976). The foundations of survey sampling: A review (with discussion). *Journal of the Royal Statistical Society, Series A*, 139:183–204.

Smith, T.M.F. (1994). Sample surveys 1975–1990: An age of reconciliation? (with discussion). *International Statistical Review*, 62:5–34.

Sugden, R.A. and Smith, T.M.F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71:495–506.

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

Zheng, H. and Little, R.J.A. (2003). Penalized spline model-based estimation of the finite population total from probability-proportional-to-size samples. *Journal of Official Statistics*, 19:99–117.

Zheng, H. and Little, R.J.A. (2005). Inference for the population total from probability-proportional-to-size samples based on predictions from a penalized spline nonparametric model. *Journal of Official Statistics*, 21:1–20.