A method for estimating sampling variances for surveys with weighting, poststratification, and raking[*]

Hao Lu[†] Andrew Gelman[‡]

December 17, 2002

Abstract

It is common practice to use weighting, poststratification, and raking to correct for sampling and nonsampling biases and to improve efficiency of estimation in sample surveys. In general, the sampling variances of the resulting estimates depend on the weighting procedures, not just on the numerical values of the weights.

In this paper we develop a method for estimating the sampling variance of survey estimates with these adjustments, using three ideas: (1) a general notation that unifies the different forms of weighting adjustment, (2) a variance decomposition to estimate sampling variances conditional and unconditional on sample sizes within poststratification categories, and (3) the delta method applied to uncertainties in sample sizes within poststrata. The resulting variance estimates are design-based and comparable to those obtained by the jackknife. We focus on estimation of population and subgroup means but also discuss more complicated summaries such as ratios and regression coefficients.

We apply our approach to the problem that motivated this research, the New York City Social Indicators Survey, a telephone survey that uses inverse-probability weighting, poststratification, and raking to correct for sampling design and nonresponse. Our variance estimates systematically differ from those obtained using methods that do not account for the design of the weighting scheme. Assuming simple random sampling leads to underestimating the sampling variance, and treating all weights as inverse-probability causes variances to be overestimated.

[*] We thank the associate editor and reviewers for helpful suggestions and the U.S. National Science Foundation for grant SES-9987748.

[†] Thales Corp., New York, NY.

[‡] Department of Statistics, Columbia University, New York, NY 10027, gelman@stat.columbia.edu.

Key words: inverse-probability weighting, iterative proportional fitting, poststratification, raking, ratio estimate, regression estimate, sample survey, variance estimation

1 Introduction

Poststratification and weighting are used to adjust for known or expected discrepancies between response group and population. We consider surveys that use inverse-probability (Horvitz-Thompson) weighting, poststratification, and raking or iterative proportional fitting (IPF). Although these methods are commonly used, it can be difficult to estimate sampling variances of the associated weighted estimates. In general, these variances depend on the weighting procedures, not just on the numerical values of the weights, and variance-estimation methods that use only the weights, without using the weighting design, cannot in general be correct. Analytic formulas exist for inverse-probability weighting (see, e.g., Hanif and Brewer, 1980) and poststratification (e.g., Little, 1993), and Taylor-series methods exist for ratio adjustment (e.g., Jones and Chromy, 1982) and raking (e.g., Sen, 1953, Wolter, 1985, Binder and Theberge, 1988, Deville et al, 1993). However, for the combination of inverse-probability weighting, poststratification, and raking, existing design-based variance estimates rely on resampling methods (see, e.g., Yung and Rao, 1996, 2000) or else can drastically underestimate the sampling variance (see Canty and Davison, 1999). Binder (1983) and Binder and Patak (1994) suggested a general approach for estimating sampling variances from complex surveys using estimating functions. But their method can be difficult to implement when sample weights are complicated functions of sample sizes within poststrata (as in the example we consider here).

In this paper, we derive design-based analytic and Taylor-series variance estimators in a general way and show how they can work even in complicated circumstances. We go beyond published Taylor-series methods (e.g., Binder and Theberge, 1988) in considering all the weighting methods together (as happens in many real surveys). Another interesting feature of our approach is a decomposition of variation conditional and unconditional on the sample sizes of poststrata. We illustrate our method with an actual survey that motivated this work.

Section 2 of this paper briefly reviews classical weighting and poststratification and describes our general framework. In Section 3 we build a procedure to find the variances of these estimators and compare these to the simple variance estimates that either ignore weights or treat them all as inverse-probability corrections. We apply our method in Section 4 to the New York City Social Indicators Survey conducted by the School of Social Work at Columbia University. Nonresponse was high in this survey, and a complicated weighting adjustment was used to match the sample to the population. We extend our results to more complex estimates such as ratios and regression coefficients in Section 5 and conclude in Section 6.

2 Weighting and Poststratification

2.1 Methods

The essential goal of weighting in sample survey estimation is for weighted averages over the sample to provide good estimates of the corresponding averages in the target population. The usual way of explaining this is that the weighted estimator will be unbiased for the population mean under repeated sampling that uses the same sampling plan. The intuitive appeal of weighting seems to be based on a more fundamental notion of creating estimates that better "represent" the target population.

Here is a brief overview (see, e.g., Lohr, 1999, and Gelman and Carlin, 2001, for more discussion and references). There are two types of classical weights: inverse-probability weights (Horvitz and Thompson, 1952) and poststratification. A basic distinction is that the former are known at the time the survey is designed whereas the latter are only determined after the data have been collected. Also they are used for different purposes: inverse-probability weights are used to correct for unequal selection probabilities whereas poststratification weights are used to correct for known discrepancies between the sample and the population. A further distinction among types of inverse-probability weights is that sometimes they are created by the survey designer, for example using probability-proportional-to-size sampling schemes (see, e.g., Yates and Grundy, 1953), and sometimes they are a byproduct of a multistage structure, as with household size weights in a survey of individuals (Gelman and Little, 1998). Raking is a poststratification method that can be used when poststrata are formed using more than one variable, but only the marginal population totals are known. It shares the same idea with poststratification as to match the sample with what is known about the target population. Iterative proportional fitting (IPF) is simply iteration of the raking procedure (Deming and Stephan, 1940, Little, 1993). Although these methods have the same intuitive interpretation, they have different statistical properties, with potential implications for the estimation of standard errors (see, Gelman and Carlin, 2000, for some simple examples). Poststratification also has application outside of sample surveys (Little, 1993, Rosenbaum and

Rubin, 1983, 1984), but we don not discuss this here.

2.2 Notation: cell weights and unit weights

We have found it useful to develop a unified notation for design-based inference under weighting and poststratification of sample surveys, following Little (1991, 1993). We shall generally follow standard practice and focus on a single survey response at a time, labeling the values on units i in the population as $Y_i, i = 1, \ldots, N$, and in the sample as $y_i, i = 1, \ldots, n$. To start with, we suppose the goal is to estimate the population mean $\theta = \overline{Y} = \sum_{i=1}^{N} Y_i/N$.

We suppose the population is divided into J stratification/poststratification cells, with population N_j and sample size n_j in each cell $j = 1, \ldots, J$, with $N = \sum_{j=1}^{J} N_j$ and $n = \sum_{j=1}^{J} n_j$. For example, if the population of U.S. adults is classified by sex, ethnicity (white or nonwhite), 4 categories of education, 4 categories of age, and 50 states, then $J = 2 \times 2 \times 4 \times 4 \times 50 = 3200$ and the cell population N_j's would be (approximately) known from the public-use subset of the long form of the U.S. Census as updated from the Current Population Survey.

We define π_j as the probability that a unit in cell j in the population will be included in the respondent sample, assuming all units have the same probability of inclusion. (In designs with unequal sampling probabilities, cells are defined with enough specificity that inclusion probabilities can be assumed equal within each cell.) For some designs, π_j is known but, in general, when nonresponse is present, it can only be estimated.

We are combining designed selection probabilities with respondent inclusion effects such as nonavailability and unit nonresponse. Although this is not the standard notation, we will find it useful for computing design-based standard errors with poststratification. If clustering crosses the classification groups, then we would define the groups to be the intersection of the clustering and stratification categories, so that in these smaller groups, the probability of inclusion is constant.

We label the population mean and standard deviation within cell j as $\theta_j = \overline{Y}_j$ and $\sigma_j = S_j$, the sample mean within cell j as $\overline{y}_j$. The overall mean in the population is then

$$(1) \qquad \theta = \overline{Y} = \frac{\sum_{j=1}^{J} N_j \theta_j}{N},$$

which we refer to as the basic poststratification identity. We focus on weighted estimates of the form

$$\hat{\theta} = \sum_{j=1}^{J} W_j \hat{\theta}_j, \tag{2}$$

where the *cell weights* W_j sum to 1. So far, equation (2) has no restriction: the W_j's and the $\hat{\theta}_j$'s can depend in any way on the data.

We use (1) and (2) as a way to unify existing estimation procedures. The classical weighting methods we consider in this paper generally avoid any explicit modeling of the response and restrict themselves to unsmoothed estimates $\hat{\theta}_j = \bar{y}_j$ and weights W_j that depend only on the n_j's and N_j's, but not on the y_j's, thus yielding population estimates of the form,

$$\hat{\theta}_W \;=\; \sum_{j=1}^{J} W_j \bar{y}_j = \frac{\sum_{i=1}^{n} w_i y_i}{\sum_{i=1}^{n} w_i}, \tag{3}$$

where $w_i = W_{j(i)}/n_{j(i)}$ is the *unit weight* of the items i in cell j. Strictly speaking, the denominator in (3) is unnecessary since, as we have defined them, the w_i's sum to 1, but the general ratio formula is useful when considering arbitrary unnormalized unit weights. The implicit model underlying these procedures is of equal probability of inclusion in the sample for all units within any cell. This is why $\bar{y}_j$ is considered a reasonable estimate for θ_j. This is also why it is helpful to poststratify as finely as possible, so that the implicit assumption of equal probability of inclusion is reasonable within each poststratification cell.

In our three methods, the weights are defined as follows.

- Inverse-probability: $W_j = \frac{n_j/\pi_j}{\sum_{j=1}^{J} n_j/\pi_j}$ for $j = 1, \ldots, J$ and $w_i = \frac{1/\pi_i}{\sum_{j=1}^{J} n_j/\pi_j}$ for $i = 1, \ldots, n$. Problems appear if inclusion probabilities are not known; also, these weights can often lead to high-variance estimates. With single-stage unequal probability sampling, we are defining a separate cell j for each possible weighting value.

- Poststratification: $W_j = N_j/N$ for $j = 1, \ldots, J$ and $w_i = N_{j(i)}/(N n_{j(i)})$ for $i = 1, \ldots, n$. Problems arise if $n_j = 0$ (or 1 or 2) for some cell j.

- Raking and iterative proportional fitting: We suppose a population is cross classified into $J = J_1 \times \ldots \times J_D$ cells, with $K = \sum_{d=1}^{D} J_d$ being the number of marginal cells. Now instead of knowing all N_j's for $j = 1, \ldots, J$, assume only the K margins $\vec{M} = (M_{1,1}, \ldots, M_{1,J_1}, \ldots, M_{D,1}, \ldots, M_{D,J_D})^t$

are known and we want to estimate N_j from the knowledge of sample sizes and margins. Let A be the $K \times J$ indicator matrix that satisfies

$$A\vec{N} = \vec{M}. \tag{4}$$

Here $\vec{N} = (N_1, \ldots, N_J)^t$ is the vector of populations of each cell. Therefore the kth row of A indicates which elements of $\vec{N}$ belong to the kth margin from $\vec{M}$. Write $A = (A_1, \ldots, A_D)^t$; here each A_i is a $J \times J_i$ matrix. Define the vector of sample margins

$$\vec{m} = (m_{1,1}, \ldots, m_{1,J_1}, \ldots, m_{D,J_D})^t = A\vec{n}. \tag{5}$$

Here $\vec{n} = (n_1, \ldots, n_J)^t$ is the vector of sample sizes. In our notation, raking proceeds in the following steps:

- 1. For $d = 1$, calculate $\vec{r} = (\frac{M_{d,1}}{m_{d,1}}, \ldots, \frac{M_{d,J_d}}{m_{d,J_d}})^t$, and compute $\vec{w} = A_d\vec{r}$.
- 2. Update vector $\vec{n}$ by multiplying each element n_j by w_j for $j = 1, \ldots, J$. Update vector $\vec{m}$ using (5).
- 3. Repeat steps 1,2 for $d = 2, 3, \ldots, D$.
- 4. For iterative proportional fitting, repeat steps 1, 2, 3 until $\vec{m}$ converges to $\vec{M}$.

The updated n_j coming from the above procedure will be the IPF estimate of N_j, say $\hat{N}_j$, for $j = 1, \ldots, J$. Then the weights resulting from raking or IPF are $W_j = \hat{N}_j/N$ for $j = 1, \ldots, J$ and $w_i = \hat{N}_{j(i)}/(Nn_{j(i)})$ for $i = 1, \ldots, n$. For raking and also IPF, the final weight can be written as a product of weight factors for each margin.

3 Variance estimates

3.1 Decomposing the variance of the weighted estimate

We treat the vector of sample sizes $\vec{n}$ as random (see, e.g., Holt and Smith, 1979, Little, 1993, for discussion); the variance decomposition gives some sense as to the importance of this choice. We assume that the poststrata are fine enough so that all units in each cell have equal probabilities of inclusion in the respondent sample (this is the implicit assumption underlying the weighted average estimators). With enough samples in each cell, the population mean and standard deviation θ_j and

σ_j can be estimated by the sample mean and standard deviation. Based on the notation of the previous section and these assumptions, the variances of the "classical estimates" $\hat{\theta} = \sum_{j=1}^J W_j \bar{y}_j$ can always be written as

$$
\begin{aligned}
\text{var}(\hat{\theta}) &= \text{E}(\text{var}(\hat{\theta}|\vec{n})) + \text{var}(\text{E}(\hat{\theta}|\vec{n})) \\
&= \text{E}\left[\sum_{j=1}^J W_j^2 \text{var}(\bar{y}_j|\vec{n})\right] + \text{var}\left[\sum_{j=1}^J W_j \text{E}(\bar{y}_j|\vec{n})\right] \\
&= \text{E}\left[\sum_{j=1}^J W_j^2 \frac{\sigma_j^2}{n_j}\right] + \text{var}\left[\sum_{j=1}^J W_j \theta_j\right] \\
&= \sum_{j=1}^J \text{E}\left(\frac{W_j^2}{n_j}\right)\sigma_j^2 + \theta^t \text{var}(W)\theta \qquad (6)
\end{aligned}
$$

Here var(W) represents the variance matrix of $W = (W_1, \ldots, W_J)^t$, where W is considered a function of the random variable $\vec{n}$. This is also classical: W is determined solely from $\vec{n}$, not from y. In (6) we need to ensure that $W_j = 0$ whenever $n_j = 0$: that is, since no data are available in such cells, they are assigned zero weights. To use equation (6) to find the variance, we need the distribution of $\vec{n}$. Without further information about the survey, it would be appropriate to assume $\vec{n} \sim \text{multinomial}(n; \frac{\pi_1 N_1}{\sum_{j=1}^J \pi_j N_j}, \ldots, \frac{\pi_J N_J}{\sum_{j=1}^J \pi_j N_j})$, where the J categories include all cross-classifications of strata and poststrata. In (6), the correlations between poststrata are zero because we are assuming that the design is single-stage (or that any clustering in the design occurs within poststrata).

Both terms in (6) are important; as Canty and Davison (1999) discuss, if the unit weights w_i are treated as fixed, this is equivalent to ignoring the sampling variance in $\vec{n}$ and thus ignoring the second term in (6), causing the variance to be underestimated.

3.2 Computation using delta method

The first term in (6) can be estimated by $\sum_{j=1}^J \frac{\widehat{W}_j^2}{n_j} s_j^2$, where s_j^2 is the sample variance of cell j and $\widehat{W}_j$ is the weight we get. For the second term in (6), we can estimate θ by the vector of sample means. For poststratification, var$(\vec{W})$ is simply zero. For inverse-probability weights and raking and IPF, the vector of cell weights $\vec{W}$ is a continuous function of $\vec{n}$, and we can write

$$
\vec{W} = h(\vec{n}) \approx h(\vec{n}_0) + B(\vec{n} - \vec{n}_0) \qquad (7)
$$

for a proper matrix B by Taylor expansion. Here $\vec{n}_0$ is the true sample size vector, and $h(\vec{n}_0)$ is the weight vector for this sample size. Then

$$\text{(8)} \qquad \text{var}(\vec{W}) \approx B\text{var}(\vec{n})B^t.$$

Since $\vec{n}$ has a multinomial distribution, $\text{var}(\vec{n})$ can be estimated by observed data. Therefore we only need to figure out the matrix B to get an estimate of $\text{var}(\vec{W})$. We estimate B using a perturbation method. From (7), if we change $\vec{n}$ by a small amount, $\vec{W}$ is also changed:

$$\text{(9)} \qquad \triangle\vec{W} \approx B\triangle\vec{n}.$$

The perturbation procedure is a sort of delta method: in step 1, we increase the first element of $\vec{n}$, n_1 by a small amount ϵ. Then obviously $\triangle\vec{n} = (\epsilon, 0, \ldots, 0)^t$. Now use the new $\vec{n}$, say $\vec{n}'$, repeat the IPF procedure and get new weights for all cells, say $\vec{W}'$. Let $\triangle\vec{W}$ be the difference of the two weights $\vec{W}' - \vec{W}$. Apply (9), the first column of B can be approximately expressed as $\triangle\vec{W}/\epsilon$. In the next step, we increase the second element of $\vec{n}$, n_2 by a small amount ϵ, and similarly we can find the estimate of the second column of B. Repeating J times, we can find the estimate of B, say $\hat{B}$. We need only calculate $\text{var}(\vec{W})$ once for the entire survey. For different estimates, we only need to change the first term in (6) and θ.

This method works well when there is a sufficiently large sample size in each poststratification or raking category. It also can be used for other kinds of weights that are generated based on $N_j, n_j, \pi_j, j = 1, \ldots, J$, for example, smoothed weighting, or combinations of inverse-probability and IPF weighting. One of the advantage of this methods is that it will also work even with complicated weighting rules, as we illustrate in the examples.

3.3 Comparison to simpler approaches

We shall compare our variance estimates to two commonly-used approximations based on simplified assumptions about the weighting. We find that for our example, the simplest approximation—the assumption of simple random sampling—underestimates the sampling variance. The next-simplest approximation of inverse-probability weighting overestimates the sampling variance.

3.3.1 Analysis as if the data had arisen from simple random sampling

We first consider the simplest approximation, which is to compute the variance of the weighted estimate under simple random sampling. Although not based on a realistic model, this is a reasonable computation because it is so simple to do, especially for binary outcomes, that it is usually computed in practice. The actual sampling variance, compared to this approximation, can be used as a design effect for correcting simple analyses.

The estimator is $\hat{\theta} = \sum_{i=1}^n w_i y_i$, with $\sum_{i=1}^n w_i = 1$. We can estimate its variance by the sample variance of the data, with the weights w_i treated as constants. Then,

$$\text{var}_{\text{SRS}}(\hat{\theta}) = \sum_{i=1}^n w_i^2 \text{var}(y_i) = (\sum_{i=1}^n w_i^2)\text{var}(y_1), \tag{10}$$

and $\text{var}(y_1)$ can be estimated as the weighted sampling variance of the data;

$$\widehat{\text{var}}(y_1) \approx \sum_{i=1}^n w_i (y_i - \hat{\theta})^2.$$

(For binary outcomes, this is just $\hat{\theta}(1-\hat{\theta})$.) Method (10) will tend to underestimate the variance because it ignores the variance in the weights, w_i. These calculations are from a design-based perspective, under which sample values y_i and sample means $\bar{y}$ are random variables (in contrast to population values Y_i and population means $\bar{Y}$ which are fixed); see, e.g., Cochran (1977).

3.3.2 Assuming inverse-probability weighting

The other natural simplifying assumption is to pretend that the weighting is all inverse-probability, with independent sampling where the probability that unit i is selected is proportional to $1/w_i$. An extensive review of variance estimation for inverse-probability weighting is Hanif and Brewer (1980). To appropriately compute the variance for inverse-probability sampling, we must acknowledge the ratio form of the weighted mean:

$$\hat{\theta} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i}.$$

The denominator of this expression is 1, but only after the weights have been normalized. If we define the population quantities $\mu_u = \text{E}(w_i y_i)$, $\mu_w = \text{E}(w_i)$, $u_i = w_i y_i$ and $\theta = \mu_u/\mu_w$, $\bar{u} = \sum_{i=1}^n w_i y_i/n$, $\bar{w} = \sum_{i=1}^n w_i/n$, then the estimator can be written as

$$\hat{\theta} = \frac{\bar{u}}{\bar{w}} = \frac{\mu_u(1+\epsilon_{\bar{u}})}{\mu_w(1+\epsilon_{\bar{w}})},$$

where $\epsilon_{\bar{u}} \sim \mathrm{N}(0, \mathrm{var}(\bar{u})/\mu_u^2)$, and $\epsilon_{\bar{w}} \sim \mathrm{N}(0, \mathrm{var}(\bar{w})/\mu_w^2)$. The Taylor expansion yields approximation,

$$\hat{\theta} \approx \theta(1+\epsilon_{\bar{u}})(1-\epsilon_{\bar{w}}+\epsilon_{\bar{w}}^2-\ldots).$$

Its variance can be estimated by

$$\begin{aligned} \mathrm{var}(\hat{\theta}) &\approx \theta^2 \mathrm{var}(\epsilon_{\bar{u}}-\epsilon_{\bar{w}}) = \frac{1}{\mu_w^2}\mathrm{var}(\bar{u}-\theta\bar{w}) \\ &\approx \frac{1}{\bar{w}^2}\mathrm{var}(\bar{u}-\hat{\theta}\bar{w}) = \frac{1}{(\sum_{i=1}^n w_i)^2}\mathrm{var}(\sum_{i=1}^n z_i), \end{aligned}$$

where $z_i = u_i - \hat{\theta} w_i$, $i = 1, \ldots, n$. Under the assumption of independent sampling, $\mathrm{var}(\sum_{i=1}^n z_i)$ can be estimated by $\sum_{i=1}^n z_i^2$ since $\mathrm{E}(\sum_{i=1}^n z_i) \approx 0$. Therefore we can estimate the variance of $\hat{\theta}$ by

$$\widehat{\mathrm{var}}_{\mathrm{HT}}(\hat{\theta}) = \sum_{i=1}^n w_i^2 (y_i - \hat{\theta})^2; \tag{11}$$

here we are assuming the weights have been normalized, so that $\sum_{i=1}^n w_i = 1$.

These approximate estimates are not in general correct since they do not account for the design of the weighting procedure; in particular, the individual w_i's are not fixed with the units but actually depend on the selected sample. In the example in the next section, we compare our variance estimates to these simpler formulas, focusing on the comparison with (10) since it is the natural point of comparison.

3.4 Jackknife variance estimator

We now consider an alternative approach to estimating sampling variances—the jackknife—which is similar to our method in that it is based on recomputing the weights based on perturbations of the data. Compared to our approach, the jackknife has the advantage of simplicity and the disadvantage of not having an analytic decomposition as in (6). The two methods give similar variance estimates, as we discuss in Sections 4.3 and 4.4.

Jackknife variance estimation for stratified multistage sampling is discussed by Rao and Shao (1992) and Yung and Rao (1996, 2000). Using notation from our Section 2.2, the procedure goes as follows: for stratum j, delete the hth sample. Repeat the raking procedure for the remaining samples and generate the jhth replicate estimate, $\hat{\theta}_{jh}$. Repeat the above steps for all h and j. Then the jackknife variance estimate of $\hat{\theta}$ is,

$$\widehat{\text{var}}_{\text{Jack}}(\hat{\theta}) = \sum_{j=1}^{J} \frac{n_j - 1}{n_j} \sum_{h=1}^{n_j} (\hat{\theta}_{jh} - \hat{\theta})^2.$$

4 Application to the Social Indicators Survey (SIS)

4.1 Survey design

In this section, we apply our method to the 1997 New York City Social Indicators Survey (NYCSIS) by the Columbia University School of Social Work in 1997 (Garfinkel and Meyers, 1999). This survey is designed to assess the individual and family well-being of area residents to better study the citywide effects of current national and state government policies on social services and welfare reform. NYCSIS is based on repeated samples of New York City proper and non-New York City metro area counties. In the 1997 wave of the survey which we analyze here, 2224 respondent families were sampled. In collecting this data, two independent telephone samples were taken; a sample of 1477 families with children and a sample of 747 individuals who were asked about their families (defined as the respondent, spouse/partner, and any children cared for by the respondent or spouse/partner).

Both samples were based on random digit dialing. In the caregiver survey, the telephone interviewer identified for each household the number of children under age 18 living at that location. One child was randomly chosen from all resident children. This child's caregiver was selected as the respondent. In the individual version, the telephone interviewer identified the number of people 18 or older residing at the household and randomly chose one of these adults as the respondent. If the respondent was also a primary caregiver, then a random child was selected from the available children. Here, the chosen focal child became the subject of child care-giving topics addressed by the respondent during the interview. If the respondent was not a primary caregiver, the survey questionnaire omitted questions specific to child rearing. The sampling design had four strata: New York City families with children, New York City families without children, non-New York City families with children, and non-New York City families without children.

4.2 Weighting adjustments

The NYCSIS used a complicated weighting scheme to adjust for several demographic variables. Here we describe the weighting procedure and then illustrate how we used our method to compute sampling variances of weighted averages.

A ten-step weighting procedure based on inverse-probability weighting and raking was applied to correct for unequal sampling and nonresponse rates. First, inverse-probability weighting was used to correct for differences in the probabilities of selection. Then poststratification and iterative proportional fitting were used to adjust based on demographics. The variables considered in the weighting procedure are selection probabilities at the family level, selection probabilities based on telephone availability and intermittent phone service at the household level, selection probabilities based on child caregiver status relative to the city and metro populations, educational differences by race or ethnicity, and family composition discrepancies. To correct such variations between the sample and population distributions, accurate population estimates from the 1996 Current Population Survey (CPS) were used for weighting controls. See Becker (1998) for details.

In step 1, sample cases were weighted at the family level by inverse probability of their selection so as to adjust for selection bias due to family and household size. Step 1 weights were calculated as the square root of the number of adults or children in the household divided by the number of adults or children in the family, capped at 4. Square roots were used because inverse-probability weights for household size tend to overcorrect in telephone surveys (Gelman and Little, 1998).

In step 2, cases with multiple phone lines were weighted downward by the inverse of the number of phone lines to correct for their higher probabilities of telephone selection. Conversely, cases with interrupted telephone service were weighted upward in direct proportion to the number of months during which they had no telephone service. This adjusted for their lower probability of telephone availability.

In step 3, weights were constructed relative to the city population to adjust for differences in selection probability resulting from variations in child caregiver status. The sample was stratified according to child caregiver status and weighted totals for families with and without children were calculated. Two ratios were obtained by comparing the CPS New York City totals for families with and without children to their corresponding NYCSIS weighted totals. From these ratios, all family cases were weighted upward in accordance with 1996 CPS New York City child caregiver population proportions.

In constructing step 4 weights, our sample was stratified both by highest educational attainment per family and by the racial or ethnic identity of the families' survey respondents. NYCSIS weighted proportions for educational attainment by racial or ethnic representation were produced. Sixteen ratios were calculated by comparing 1996 CPS New York City education by race or ethnicity proportions to our weighted NYCSIS proportions. From these ratio weights, our sample strata were adjusted for deviations in educational attainment and racial or ethnic composition relative to the underlying New York City population.

Step 5 involved the building of family composition weights, and our sample was stratified into categories based on family organization. Parallel to the method in step 4, NYCSIS weighted proportions for family composition were obtained. Six ratios were computed by setting corresponding 1996 CPS New York City family composition proportions over the weighted NYCSIS proportions. With these ratio weights, our sample strata were accurately stabilized for differences in family composition relative to the general New York City population.

Steps 6, 7, and 8 of the weighting procedure involved iterations back to weight steps 3, 4, and 5. Weights adjusting for selection probability due to child caregiver status, weights adjusting for differences in educational attainment by race or ethnicity, and weights adjusting for family composition by poverty status, all converged after two iterations.

In step 9, final weights for all NYCSIS sample families were calculated as the products of the appropriate individual weights produced by steps 1 through 8. Lastly, with step 10, a weight was created to adjust the NYCSIS data to the aggregate population total for New York City from the 1996 CPS. The weights converged well with two iterations: for example, in step 8, all of the weight updating factors are between 0.99 to 1.02.

The final weighting corrections are shown in Tables 1–4. Table 1 shows that most Social Indicators Survey respondents live in one-family households, have one telephone line, and have complete telephone service. These cases are simply weighted with a value of 1 in the first two steps. Table 2 shows that weights are largest for families without children. These weights are comparatively large because the representation of these strata in the sample of 2224 cases is small. Table 3 shows that the minority respondents tend to have higher educations than the general minority population and are weighted down significantly. The SIS sample did not cover enough minorities and whites with low educations, thus, the weights for these cases are comparatively high. The weights in Table 4 show that the SIS oversampled single male householders without children and couples households with children, while sampling too few single female householders without children and single male householders with children. A brief summary of the weights appears in Table 4.

From the histograms of the final weights (Figure 1), we observe that the weights for the two strata of New York City families with children and non-New York City families with children are balanced primarily around 1; however, a few outlying larger weights remain in each strata. With the final weights for the two strata of New York City families without children and non-New York City families without children, the histograms show that these weights are spread in larger distributions with the majority of the cases balanced around 5 and 10 respectively. A few heavy outlying final weights exist for each of these strata.

4.3 Examples of weighted means and variance estimates

Given these weights, we can compute weighted estimates, and by applying methods we described in Section 3 we can find the variances of these estimates. Some examples of results are shown in Tables 5–10, with comparison with CPS estimates. In these calculations, "don't know" responses and refusals are not included in the weighted estimates.

We know that applying poststratification weights can decrease the bias of the total survey error, and at the same time this can increase the variance due to an increase in the variance of the respondent weights (Kish, 1965). Therefore for estimators not based on poststratification strata, we calculate the design effects, the inflation factor to the variance of the estimators assuming simple random sampling. Tables 6–10 show that the design effect is around 1.7^2 for New York City families for various estimands of interest. That is, the approximation based on simple random sampling drastically underestimates the variances (see Canty and Davison, 1999, for a similar point). (We write design effects as d^2, so that d represents the factor by which confidence intervals must be widened.)

Conversely, the approximate variance estimates based on the assumption of inverse-probability sampling are all much too high in these examples. Jackknife variance estimates are also included in Tables 6–10. They are very close to estimates using decomposition (6).

Description	Weight (frequency)
4+ lines, complete service	0.25 (27)
3 lines, complete service	0.33 (81)
2 lines, complete service	0.50 (271)
1 line, complete service	1.00 (1765)
1 line, < 1 month intermittent service	1.09 (34)
1 line, 1-3 months intermittent service	1.20 (31)
1 line, 4-6 months intermittent service	1.71 (7)
1 line, 7+ months intermittent service	4.00 (8)

Table 1: Inverse-probability weights (with frequencies in the sample in parentheses) adjusting for differences in selection associated with telephone use at the household level.

	with children	without children
NYC families	1.00 (1176)	6.75 (306)
non-NYC metro families	3.14 (584)	10.36 (158)

Table 2: Final weight factors for four different strata with frequencies in the sample in parentheses.

	16 years	13–15 years	12 years	1–11 years
White	0.87 (521)	0.75 (241)	2.10 (185)	6.16 (27)
Black	0.48 (137)	0.55 (184)	1.05 (171)	1.77 (68)
Hispanic	0.31 (116)	0.36 (167)	0.93 (126)	2.89 (82)
Asian	0.48 (95)	0.56 (30)	1.98 (22)	14.31 (7)

Table 3: Final weight factors stratified by ethnicity and educational attainment with frequencies in the sample in parentheses. There were 45 unidentified people not included in this table; they were given weight factors of 1 for these variables.

	with children	without children
Coupled families	0.87 (1224)	1.07 (174)
Female householder	1.05 (471)	1.20 (164)
Male householder	1.40 (55)	0.78 (123)

Table 4: Final weight factors for different family compositions with frequencies in the sample in parentheses. There were 13 people who don't know or refuse to answer; they were given weight factors of 1 for these variables.

Family type	SIS unweighted %	SIS weighted % ± s.e.	CPS %
Coupled families, children	49.7	16.5 ± 0.014	16.5
Coupled families, no children	6.9	18.8 ± 0.016	19.0
Female householder, children	26.3	14.4 ± 0.012	14.5
Female householder, no children	7.5	29.8 ± 0.025	29.9
Male householder, children	2.9	1.7 ± 0.001	1.7
Male householder, no children	6.1	18.3 ± 0.016	18.4
Don't know, refuse	0.6	0.5 ± 0.001	0

Table 5: Estimated populations (in proportion) of different family types among New York City families. Here since we used this information in the poststratification, the estimated variances are essentially zero.

	estimates (%)			standard error ests (%)			
Household	SIS	SIS	CPS	design-	assuming	assuming	jackknife
tenure	raw	weighted		based	SRS	inv-prob	
Owner	30.0	24.8	26.9	1.5	1.1	2.2	1.7
Renter	65.3	68.1	72.0	1.8	1.2	2.4	1.9
Staying there/rent free/other	4.7	7.1	1.1	1.2	0.7	1.4	1.3

Table 6: Proportions of different household tenures among New York City families. Design-based standard errors (computed using our method developed in Sections 3.1 and 3.2) can be compared with standard error estimates under two simplifying approximations and the jackknife estimate. Design effects would be computed as the ratio of the design-based and SRS variances and are in the range 1.3^2 *to* 1.8^2 *for this example.*

	estimates (%)			standard error ests (%)			
Place of	SIS	SIS	CPS	design-	assuming	assuming	jackknife
birth	raw	weighted		based	SRS	inv-prob	
United States	56.4	61.2	60.0	1.9	1.3	2.4	1.9
Other country	43.6	38.8	40.0	1.9	1.3	2.4	1.9

Table 7: Proportion of people born in United States and other countries of New York City families. Design-based standard errors (computed using our method developed in Sections 3.1 and 3.2) can be compared with standard error estimates under two simplifying approximations and the jackknife estimate. Design effects, compared to simple random sampling, are 1.5^2 *for this example.*

	estimates (%)		standard error ests (%)			
Rating of	SIS	SIS	design-	assuming	assuming	jackknife
NYC	raw	weighted	based	SRS	inv-prob	
Very good	15.8	16.4	1.5	1.0	1.7	1.5
Pretty good	39.5	44.3	2.1	1.3	2.5	2.2
Only fair	37.1	30.8	1.9	1.2	2.0	2.0
Poor	7.6	8.5	1.1	0.7	1.3	1.1

Table 8: Adult rates New York City as a place to live .... New York City families only. Design-based standard errors (computed using our method developed in Sections 3.1 and 3.2) can be compared with standard error estimates under two simplifying approximations and the jackknife estimate. Design effects, compared to simple random sampling, are in the range 1.5^2 *to* 1.7^2 *for this example.*

	estimates (%)			standard error ests (%)		
Opinion of	SIS	SIS	design-	assuming	assuming	jackknife
NYC	raw	weighted	based	SRS	inv-prob	
Become a better place	32.8	35.8	2.1	1.2	2.5	2.1
Remained the same	29.7	29.2	1.9	1.2	2.2	1.9
Gotten worse	37.5	35.0	1.9	1.2	2.4	2.0

Table 9: Adult thinks that in the last few years New York has … New York City families only. Design-based standard errors (computed using our method developed in Sections 3.1 and 3.2) can be compared with standard error estimates under two simplifying approximations and the jackknife estimate. Design effects, compared to simple random sampling, are in the range 1.5^2 *to* 1.7^2 *for this example.*

	estimates (%)			standard error ests (%)		
Rating of	SIS	SIS	design-	assuming	assuming	jackknife
police	raw	weighted	based	SRS	inv-prob	
Very good	21.9	23.2	1.9	1.1	2.1	1.9
Pretty good	35.0	39.5	2.1	1.3	2.5	2.1
Only fair	31.6	24.8	1.7	1.1	2.0	1.7
Poor	11.5	12.5	1.4	0.9	2.0	1.4

Table 10: Adult rates police protection … New York City families only. Design-based standard errors (computed using our method developed in Sections 3.1 and 3.2) can be compared with standard error estimates under two simplifying approximations and the jackknife estimate. Design effects, compared to simple random sampling, are in the range 1.5^2 *to* 1.7^2 *for this example.*

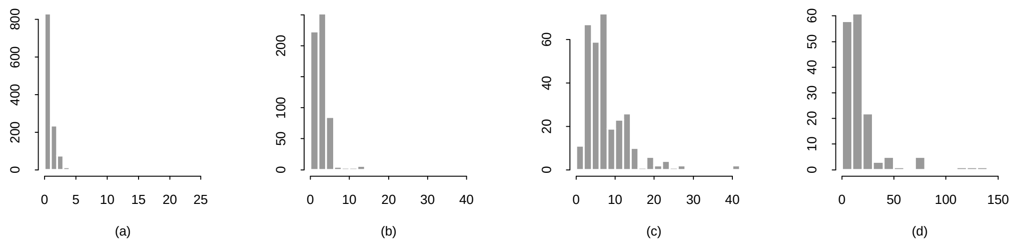

Figure 1: (a) Weights for New York City families with children. (b) Weights for non-New York City metro families with children. (c) Weights for New York City families without children. (d) Weights for non-New York City metro families without children.

4.4 A simulation study

We conducted a simulation study based on the NYCSIS data. For each cell, we calculate the mean response to a single question (for example, household tenure) as the "truth." Then we simulate 1000 fake data sets: simulating $\vec{n}$ from the multinomial distribution, then simulating responses y within each cell using the binomial distribution, given the "true" probabilities. For each simulation, we compute the survey weights and the weighted estimate, which we call $\hat{\theta}_{\text{sim}}$. For each simulation, compute the 5 different standard error estimates: assuming simple random sampling; assuming inverse-probability weighting; assuming simple poststratification weighting; using our design-based method; and using the jackknife. Then we put all the simulations together: we compute the standard deviations of the 1000 $\hat{\theta}_{\text{sim}}$'s: this is the "true" standard error. We compare the true standard error to the average standard error estimates obtained in the 5 different ways. The results are shown in Tables 11-15.

Household	true	assuming	assuming	assuming	design-	jackknife
tenure	sd	SRS	inv-prob	post-strat	based	
Owner	1.9	1.1	2.1	1.5	1.8	1.8
Renter	2.0	1.2	2.3	1.7	1.9	1.9
Staying there/rent free/other	1.2	0.7	1.3	1.0	1.2	1.1

Table 11: From simulation study: true standard error and five different standard error estimates (in percentages) for the proportions of different household tenures among New York City families.

Place of	true	assuming	assuming	assuming	design-	jackknife
birth	sd	SRS	inv-prob	post-strat	based	
United States	2.0	1.3	2.5	1.7	1.9	1.9
Other country	2.0	1.3	2.5	1.7	1.9	1.9

Table 12: From simulation study: true standard error and five different standard error estimates (in percentages) for the proportion of people born in United States and other countries of New York City families.

Rating of NYC	true sd	assuming SRS	assuming inv-prob	assuming post-strat	design- based	jackknife
Very good	1.6	1.0	1.7	1.4	1.6	1.6
Pretty good	2.5	1.3	2.6	1.9	2.2	2.2
Only fair	2.0	1.2	2.2	1.7	1.9	1.9
Poor	1.1	0.7	1.2	0.9	1.1	1.1

Table 13: From simulation study: true standard error and five different standard error estimates (in percentages) for the proportion of adults who rate New York City as a place to live ....

Opinion of NYC	true sd	assuming SRS	assuming inv-prob	assuming post-strat	design- based	jackknife
Become a better place	2.2	1.2	2.5	1.9	2.1	2.1
Remained the same	2.0	1.2	2.3	1.6	1.9	1.9
Gotten worse	2.0	1.2	2.4	1.7	2.0	2.0

Table 14: From simulation study: true standard error and five different standard error estimates (in percentages) for the proportion of adults who think that in the last few years New York has ... New York City families only.

Rating of police	true sd	assuming SRS	assuming inv-prob	assuming post-strat	design- based	jackknife
Very good	2.1	1.1	2.2	1.7	1.9	1.9
Pretty good	2.2	1.3	2.4	1.9	2.1	2.1
Only fair	1.9	1.1	2.1	1.6	1.8	1.7
Poor	1.5	0.9	1.9	1.2	1.4	1.4

Table 15: From simulation study: true standard error and five different standard error estimates (in percentages) for the proportion of adults who rate police protection ....

5 Other estimators: ratios and regression coefficients

So far we have focused on estimating the population mean or subgroup means. In general, however, one may be interested in more complex estimands, most notably ratios and regression estimates (see, e.g., Cochran, 1977).

5.1 Ratios

Ratios arise in various ways, perhaps the most common being means of subgroups with unknown population proportions. For example, suppose we are interested in θ, the average income of whites. If we let Y_i be the income response and U_i be the indicator with $U_i = 1$ if the respondent is white and 0 otherwise, then

$$\theta = \frac{\sum_{i=1}^{N} U_i Y_i}{\sum_{i=1}^{N} U_i} = \frac{\overline{V}}{\overline{U}}, \tag{12}$$

where $V_i = Y_i U_i$. The bias correction and standard error of a ratio estimate can be estimated using Taylor expansion (see, Lohr, 1999). As a particular example, the inverse-probability weighting estimator can be viewed as a ratio estimator.

In general, the ratio (12) can be estimated in a weighted analysis by

$$\hat{\theta} = \frac{\sum_{i=1}^{n} w_i u_i y_i}{\sum_{i=1}^{n} w_i u_i} = \frac{\bar{v}_w}{\bar{u}_w}, \tag{13}$$

where y_i, u_i are observed valuez, and w_i is the unit weight for case i. We define v_i as $u_i y_i$, and $\bar{v}_w = \sum_{i=1}^{n} w_i v_i$, $\bar{u}_w = \sum_{i=1}^{n} w_i u_i$. The variance of this estimate can be approximated by

$$\begin{aligned} \mathrm{var}(\hat{\theta}) &\approx \mathrm{E}(\hat{\theta} - \theta)^2 \approx \frac{1}{\overline{U}^2}\mathrm{E}(\bar{v}_w - \theta \bar{u}_w)^2 \\ &\approx \frac{1}{\bar{u}_w^2}\mathrm{var}(\bar{v}_w - \theta \bar{u}_w) \\ &= \frac{1}{\bar{u}_w^2}\mathrm{var}(\bar{z}_w) \end{aligned}$$

where $z_i = u_i y_i - \hat{\theta} u_i$ and $\bar{z}_w = \sum_{i=1}^{n} w_i z_i = \bar{v}_w - \hat{\theta}\bar{u}_w$. Now we can use the method described in Section 3 to find $\mathrm{var}(\bar{z}_w)$ and hence the variance of the ratio estimator.

We illustrate using the survey described in Section 4. Suppose that in estimating the adult opinions of police protection (see Table 10), we are interested in the different opinions of different ethnicities. As explained in last paragraph, this is a ratio estimate, and its variance can be estimated as explained. The result is shown in Table 16.

Adult thinks %	White±s.e.(d.eff)	Black±s.e.(d.eff)	Hispanic±s.e.(d.eff)	Other±s.e.(d.eff)
Very good	$29.2 \pm 3.4\ (1.6^2)$	$12.2 \pm 2.0\ (1.4^2)$	$18.9 \pm 3.6\ (1.9^2)$	$38.1 \pm 5.3\ (1.2^2)$
Pretty good	$45.7 \pm 3.6\ (1.6^2)$	$32.1 \pm 3.8\ (1.8^2)$	$40.3 \pm 4.7\ (2.0^2)$	$28.8 \pm 4.4\ (1.1^2)$
Only fair	$14.1 \pm 2.2\ (1.4^2)$	$35.7 \pm 3.6\ (1.7^2)$	$32.3 \pm 4.4\ (1.9^2)$	$26.9 \pm 5.0\ (1.2^2)$
Poor	$11.0 \pm 2.3\ (1.6^2)$	$20.0 \pm 2.9\ (1.6^2)$	$8.5 \pm 3.6\ (2.7^2)$	$6.2 \pm 2.6\ (1.2^2)$
Total	100	100	100	100

Table 16: Estimates and standard errors for the proportion of New York City adults of different ethnic groups who rate police protection ...The design effects compared to simple random sampling are in parentheses.

5.2 Regression estimates

Regression estimates commonly arise in analytical studies of sample survey responses that attempt to understand what variables are predictive of an outcome of interest Y. Suppose that we are interested in finding a relation between Y_i and a p-dimensional vector of explanatory variables X_i, where $X_i = (X_{i1}, \ldots, X_{ip})^t$. We want to estimate the p-dimensional vector of population parameters, $\beta = (\beta_1, \ldots, \beta_p)$, in the linear model $Y = \beta X$. Now given the sample $y_s = (y_1, \ldots, y_n)^t$ and $x = (x_1, \ldots, x_n)$, the regression coefficient can be estimated by

$$\hat{\beta} = (xWx^t)^{-1} xWy, \tag{14}$$

where W is a diagonal matrix of the unit weights w_i. Let

$$z_i = x_i^t (y_i - \hat{\beta} x_i).$$

Then, using linearization (see, Shah et al, 1977),

$$\text{var}(\hat{\beta}) = (xWx^t)^{-1} \text{var}(\sum_{i=1}^{n} w_i z_i)(xWx^t)^{-1} \tag{15}$$

where $\text{var}(\sum_{i=1}^{n} w_i z_i)$ can be estimated using the method described in Section 3, so can $\text{var}(\hat{\beta})$.

6 Discussion

In practice, classical weighting methods such as inverse-probability and iterative proportional fitting of poststratification weighting are extensively used. We showed in this paper how sampling variances of the resulting weighted estimators can be found in a rather straightforward way, even with large numbers of poststrata as in the NYCSIS example in Section 4. For this application, we recommended that the users compute standard errors assuming simple random sampling and then multiply by 1.7 to correct for design and weighting.

We hope these methods will allow researchers to better account for the design of weighting schemes in survey inferences. By comparison, simpler estimates of standard errors based on standard inverse-probability or poststratification formulas can be far off (see, for example, Tables 6–10, where our design-based estimates are compared to these approximations).

But there are difficulties with classical weighted estimates. As (3) indicates, if the unit weights are too variable, then $\hat{\theta}$ will itself have an unacceptable high variance, and this will occur if the n_j's are small. This is thus a tension between two extremes: (a) keeping the number of weighting cells small, so that the individual n_j's will be reasonably large and the weighted estimate not too variable; and (b) increasing the number of cells, which makes more plausible the implicit assumption of equal-probability sampling within cells but leads to more variable estimates. A commonly-used compromise is to keep a large number of weighting cells but to "smooth the weights": that is, to set the units weights so that they are less variable than would arise from simply setting $w_i \sim N_{j(i)}/n_{j(i)}$. In practice, this means that the units from cells with small sample sizes do not get such large weights as they would receive under the unbiased estimate. Raking and IPF can be thought of as sophisticated methods for smoothing weights (Elliott and Little, 1999). In settings where cell sizes are small, estimated variances within cells can be unstable as well.

An extreme version of the instability problem occurs with non-structural zero cells: that is, cells for which $n_j = 0$ but $N_j > 0$. This can obviously happen; for example, if $J = 3200$ in the second paragraph of Section 2.2 and n is 1500, say, which is typical in national polls, then by necessity most of the cells will be empty. In this case the classical solution is to pool weighting cells, or to adjust for partial information (as with raking). The choice of which cells to pool or which margins to rake over is somewhat arbitrary and contradicts the goal of including in the analysis all variables that affect the probability of inclusion, which is a basic principle in classical sampling inference.

Further work is needed to define ways which will overcome these difficulties while incorporating the information used in classical weighting adjustments that are currently used in practice.

7 References

Becker, D. E., 1998. The New York City Social Indicators Survey: an analysis of the weighting procedure. Technical report, School of Social Work, Columbia University.

Binder, D. A., 1983. On the variance of asymptotically normal estimators from complex surveys. *International Statistical Review*, **51**, 279-292.

Binder, D. A., and Theberge, A., 1988. Estimating the variance of raking ratio estimators. *Canadian Journal of Statistics*, **16** (supplement), 47-55.

Binder, D. A., and Patak, Z., 1994. Use of estimating functions for estimation from complex surveys. *Journal of American Statistical Association*, **89**, 1035-1043.

Canty, A. J., and Davison, A. C., 1999. Resampling-based variance estimation for labour force surveys. *The Statistician*, **48**, 379-391.

Cochran, W. G., 1977. *Sampling Techniques,* 3d ed. New York: Wiley.

Deming, W. E., and Stephan, F. F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *The Annuals of Mathematical Statistics*, **11**, 427-444.

Deville, J., Särndal, C., and Sautory, O., 1993. Generalized raking procedures in survey sampling. *Journal of American Statistical Association*, **88**, 1013-1020.

Elliott, M. R., and Little, R. J. A., 1999. Model-based alternatives to trimming survey weights. Technical report, Department of Biostatistics, University of Michigan, Ann Arbor.

Garfinkel, I., and Meyers, M. K., 1999. New York City social indicators 1997: a tale of many cities. School of Social Work, Columbia University.

Gelman, A., and Carlin, J. B., 2001. Poststratification and weighting adjustments in survey nonresponse. In *Survey Nonresponse*, ed. Groves, R., Dillman, D., Eltinge, J., and Little, R..

Gelman, A., and Little, T. C., 1998. Improving on probability weighting for household size. *Public Opinion Quarterly,* **62**, 398-404.

Hanif, M., and Brewer, K. R. W., 1980. Sampling with unequal probabilities without replacement: a review. *International Statistical Review*, **48**, 317-335.

Horvitz, D. G. and Thompson, D. J., 1952. A generalization of sampling without replacement from a finite population. *Journal of American Statistical Association*, **47**, 663-685.

Holt, D., and Smith, T. M. F., 1979. Post-stratification. *Journal of the Royal Statistical Society, Ser. A,* **142**, 33-46.

Jones, S. M., and Chromy, J. R., 1982. Improved variance estimators using weighting class adjustments for sample survey nonresponse. In *Proceedings of the Survey Research Methods Section, American Statistical Association,* 105-110.

Kish, L., 1992. Weighting for unequal P_i. *Journal of Official Statistics*, **8**, 183-200.

Little, R. J. A., 1986. Survey nonresponse adjustments for estimates of means. *International Statistics Review*, **54**, 139-157.

Little, R. J. A., 1991. Inference with survey weights. *Journal of Official Statistics*, **7**, 405-424.

Little, R. J. A., 1993. Post-stratification: a modeler's perspective. *Journal of the American Statistical Association*, **88**, 1001-1012.

Little, T. C., 1996. Models for nonresponse adjustment in sample surveys. Ph.D. thesis, Department of Statistics, University of California, Berkeley.

Lohr, S. L., 1999. *Sampling: Design and Analysis.* Pacific Grove, Ca.: Brooks-Cole.

Rao, J. N. K., and Shao, J., 1992. Jackknife variance estimation with survey data under hot deck imputation. *Biometrika*, **79**, 811-822.

Rosenbaum, P. R., and Rubin, D. B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41-55.

Rosenbaum, P. R., and Rubin, D. B., 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, **79**, 516-524.

Sen, A. R., 1953. On the estimate of variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics,* **5**, 119-127.

Shah, B. V., Holt, M. M., and Folsom, R.E., 1977. Inference about regression models from sample survey data. *Bulletin of the International Statistical Institute*, **47**, 43-57.

Yates, F. and Grundy, P. M., 1953. Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society*, **B15**, 253-261.

Yung, W., and Rao, J. N. K., 1996. Jackknife linearization variance estimators under stratified multi-stage sampling. *Survey Methodology*, **22**, 23-31.

Yung, W., and Rao, J. N. K., 2000. Jackknife variance estimation under imputation for estimators using poststratification information. *Journal of the American Statistical Association*, **95**, 903-915.

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