Statistical Science
2002, Vol. 17, No. 1, 73–96

Inference for Superpopulation Parameters Using Sample Surveys

Barry I. Graubard and Edward L. Korn

Abstract. Sample survey inference is historically concerned with finite-population parameters, that is, functions (like means and totals) of the observations for the individuals in the population. In scientific applications, however, interest usually focuses on the "superpopulation" parameters associated with a stochastic mechanism hypothesized to generate the observations in the population rather than the finite-population parameters. Two relevant findings discussed in this paper are that (1) with stratified sampling, it is not sufficient to drop finite-population correction factors from standard design-based variance formulas to obtain appropriate variance formulas for superpopulation inference, and (2) with cluster sampling, standard design-based variance formulas can dramatically underestimate superpopulation variability, even with a small sampling fraction of the final units. A literature review of inference for superpopulation parameters is given, with emphasis on why these findings have not been previously appreciated. Examples are provided for estimating superpopulation means, linear regression coefficients and logistic regression coefficients using U.S. data from the 1987 National Health Interview Survey, the third National Health and Nutrition Examination Survey and the 1986 National Hospital Discharge Survey.

Key words and phrases: Cluster sampling, complex survey data, design-based inference, model-based inference, random effects, stratified sampling.

1. INTRODUCTION

In the context of sample surveys, Deming and Stephan (1941) considered a "superpopulation" to be a hypothetical infinite population from which the finite population is itself a sample. An investigator samples the finite population and draws inferences from the sampled values. In classical sampling theory, the targets of inference are finite-population parameters, for example, the mean $\overline{Y}$ of the N unit values in the population. A stochastic model for the finite-population values is sometimes used to evaluate and suggest sample designs and estimators; see, for example, Cochran (1939, 1946), Scott and Smith (1969), Ericson (1969), Hartley and Sielken (1975), Cassel, Särndal and Wretman (1977, Chapters 4–6), Bellhouse, Thompson and Godambe (1977) and Isaki and Fuller (1982). For addressing scientific questions (as opposed to administrative and quality assurance applications), however, the parameters associated with the stochastic model are typically of more interest than the finite-population parameters. Deming (1953) refers to inference for superpopulation parameters as an "analytic" use of survey data. A simple example of superpopulation inference is when comparing two domain means, where it is of interest to ask if the superpopulation means are equal, but seldom of interest to ask if the finite-population means are equal (since they rarely would be).

If the target of inference is a superpopulation parameter, then how should the inference be made? One usual recommendation is not to use finite-population correction factors when estimating variances of parameter estimators (Fuller, 1975; Cochran, 1977, page 39;

Barry I. Graubard is Mathematical Statistician, Biostatistics Branch, EPS-8024 National Cancer Institute, Bethesda, Maryland 20892 (e-mail: bg1p@nih.gov). Edward L. Korn is Head, Clinical Trials Section, Biometric Research Branch, EPN-8128, National Cancer Institute, Bethesda, Maryland 20892.

Yates, 1981, page 178). In addition, the following heuristic argument for means, which also applies to more complex parameters, is sometimes given for ignoring the distinction between finite-population and superpopulation inference (Skinner, Holt and Smith, 1989, page 14): under a superpopulation model in which the $Y_1, \dots, Y_N$ are independent and identically distributed with mean μ and variance σ^2, one has $\overline{Y} = \mu + O_p(N^{-1/2})$ and $S^2 = \sigma^2 + O_p(N^{-1/2})$, where S^2 is the sample variance of the $Y_1, \dots, Y_N$. Therefore, there is little difference between a finite-population inference for $(\overline{Y}, S^2)$ or a superpopulation inference for (μ, σ^2) when N is large.

This heuristic argument fails, of course, if terms of order $N^{-1/2}$ are important for the inference. This would be the case if the sample size n did not satisfy $n \ll N$, since then terms of order $N^{-1/2}$ should not be ignored when estimating the variability of the sample mean $\overline{y}$. Even with $n \ll N$, the argument does not apply when, as is typically the case, there are clusters in the population. Under a reasonable superpopulation model, one might expect in this situation that the differences in finite-population parameters and superpopulation parameters would be bounded by terms of order $K^{-1/2}$, where K is the number of clusters in the population. Since one may have the number of sampled clusters k not satisfying $k \ll K$ even though $n \ll N$, the inference for finite-population and superpopulation parameters may differ even with $n \ll N$ and N large.

This paper builds on the results of Korn and Graubard (1998), which focused on inferences for superpopulation means, by providing results (including data examples) of more complex superpopulation quantities such as ratios and regression coefficients. An investigation is made of the relationship between superpopulation inference and two-phase sampling. Additionally, this paper reviews the literature on inferences for superpopulation parameters with attention paid to why problems with finite-population variance estimators discussed in this paper have not been previously noted.

The structure of this paper is as follows. A superpopulation model without clusters is considered in Section 2. Using this model allows for a simple examination of many of the inference issues encountered. In particular, with stratified sampling it is shown that ignoring finite-population correction factors is not sufficient to yield correct inference for superpopulation parameters. The relationship between inference for superpopulation parameters and finite-population parameters associated with two-phase sampling is also discussed. The consideration of more realistic models with clusters in Section 3 allows for the investigation of the distribution of cluster effects on the inference. It is shown that even if the sampling fraction of clusters k/K is small, finite-population inference can be misleading with certain commonly seen distributions of cluster sizes. Section 4 reviews the literature on inference for superpopulation parameters. Examples using three national health surveys of the United States are given in Section 5. The results in this paper can be viewed in light of the debate between proponents of model-based versus design-based (randomization) inference for survey data. The paper ends with a discussion of this in Section 6.

2. MODEL WITHOUT CLUSTERS

Throughout this section and later sections we will consider the properties of estimators that incorporate both the randomness due to sampling of the population and the randomness due to the generation of the population by a model. Letting the subscript RS refer to the (repeated) sampling randomness, the subscript F refer to the model randomness, and no subscript refer to both sources of randomness, one has the usual decompositions for the expectation and variance of an estimator $\hat{\theta}$ that will be used throughout: $E(\hat{\theta}) = E_F[E_{\text{RS}}(\hat{\theta})]$, $\text{Var}(\hat{\theta}) = E_F[\text{Var}_{\text{RS}}(\hat{\theta})] + \text{Var}_F[E_{\text{RS}}(\hat{\theta})]$.

In this section, population models without clusters are considered. Although applications usually involve clusters in the population, most of the important ideas are already seen in the simpler unclustered model. The population values are $(Y_1, \eta_1), \dots, (Y_K, \eta_K)$, where η is a stratum indicator with range $\{1, \dots, L\}$. We assume that the (Y_i, η_i) are independent and identically distributed, each with the same distribution as the random vector (Y, η), which has bivariate distribution function F. We restrict attention to stratified simple random sampling throughout this section and consider more general sampling schemes in Section 3.

In Section 2.1 we consider the case of estimating a mean with Y univariate. It is shown that unbiased variance estimation for the superpopulation mean involves a between-strata component. This naturally suggests the alternative possibility of ignoring the stratification in order to obtain unbiased variance estimators. This is shown not to work in Section 2.2, which also considers variance estimation for a poststratified mean. Sections 2.3 considers the estimation of superpopulation ratios, and Section 2.4 considers linear regression coefficient and other parameters. Section 2.5 considers the estimation of the superpopulation mean, but via a ratio estimator. Case–control sampling can be viewed as a special case of stratified sampling, and this is discussed in Section 2.6. To address concerns that the sampling strata are not the same as the superpopulation strata, a generalization of the superpopulation model is discussed in Section 2.7. Section 2.8 indicates the relationship of superpopulation inference with finite-population inference using two-phase sampling.

2.1 Estimating a Superpopulation Mean

The target parameter is $\mu = E_F(Y)$, which is to be estimated using stratified simple random sampling without replacement. Let K_h be the known number of observations in the hth stratum in the finite population, $h = 1, \dots, L$. Let k_h be the number of sampled observations in the hth stratum, which can be a function of $\{K_1, \dots, K_L\}$. [Note that $\sum_{h=1}^{L} K_h E_F(Y|\eta = h)$ is not defined as the target parameter, as this quantity is not a superpopulation parameter because it depends on the K_h.] The total number of observations sampled, $k = k_1 + \cdots + k_L$, is fixed. The stratified mean is

$$\bar{y} = \sum_{h=1}^{L} \frac{K_h}{K} \bar{y}_h, \tag{2.1}$$

where $\bar{y}_h$ is the mean of the sampled observations in stratum h. Under repeated sampling of the same finite population, the stratified mean is an unbiased estimator of the population mean $\overline{Y}$. The repeated-sampling variance estimator of $\bar{y}$ is

$$\widehat{\text{var}}_{\text{wo}}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{K_h - k_h}{K_h} \frac{s_h^2}{k_h}, \tag{2.2}$$

where s_h^2 is the sample variance of the observations in the hth stratum. Under repeated sampling of the same finite population, $\widehat{\text{var}}_{\text{wo}}(\bar{y})$ is an unbiased estimator of the variance of $\bar{y}$. In what follows, it will be useful to consider the repeated-sampling variance estimator treating the sample as if it had been a with-replacement sample, namely,

$$\widehat{\text{var}}_{\text{wr}}(\bar{y}) = \sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{s_h^2}{k_h}.$$

Incorporating the randomness due to the sampling and the distribution function F, we have $\bar{y}$ is unbiased for μ, and (in obvious notation)

$$\begin{aligned}
\text{Var}(\bar{y}) &= E_F[\text{Var}_{\text{RS}}(\bar{y})] + \text{Var}_F[E_{\text{RS}}(\bar{y})] \\
&= E_F\Biggl[\sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{K_h - k_h}{K_h} \frac{S_h^2}{k_h}\Biggr] + \text{Var}_F[\overline{Y}] \\
&= \frac{1}{K^2} \sum_{h=1}^{L} \sigma_h^2 E_F\biggl[\frac{K_h(K_h - k_h)}{k_h}\biggr] \\
&\quad + \frac{1}{K^2} \sum_{h=1}^{L} \sigma_h^2 E_F(K_h) \\
&\quad + \frac{1}{K^2} \text{Var}_F\Biggl(\sum_{h=1}^{L} K_h \mu_h\Biggr) \\
&= \frac{1}{K^2} \sum_{h=1}^{L} \sigma_h^2 E_F\biggl[\frac{K_h^2}{k_h}\biggr] + \frac{1}{K} \Delta_{\text{betw}},
\end{aligned} \tag{2.3}$$

where S_h^2 is the variance of the population values in the hth stratum, μ_h and σ_h^2 are the mean and variance of Y in the hth stratum with respect to the F distribution and the between-strata variability of the μ_h is

$$\Delta_{\text{betw}} = \sum_{h=1}^{L} \pi_h (\mu_h - \mu)^2,$$

where $\pi_h \equiv E(K_h/K) = P(\eta = h)$. (Note that $\mu = \sum_{h=1}^{L} \pi_h \mu_h$.) The expected values of the variance estimators are given by

$$\begin{aligned}
E[\widehat{\text{var}}_{\text{wo}}(\bar{y})] &= E_F\Biggl[E_{\text{RS}}\Biggl[\sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{K_h - k_h}{K_h} \frac{s_h^2}{k_h}\Biggr]\Biggr] \\
&= E_F\Biggl[\sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{K_h - k_h}{K_h} \frac{S_h^2}{k_h}\Biggr]
\end{aligned} \tag{2.4}$$

$$= \frac{1}{K^2} \sum_{h=1}^{L} \sigma_h^2 E_F\biggl[\frac{K_h(K_h - k_h)}{k_h}\biggr] \tag{2.5}$$

and

$$\begin{aligned}
E[\widehat{\text{var}}_{\text{wr}}(\bar{y})] &= E_F\Biggl[\sum_{h=1}^{L} \frac{K_h^2}{K^2} \frac{S_h^2}{k_h}\Biggr] \\
&= \frac{1}{K^2} \sum_{h=1}^{L} \sigma_h^2 E_F\biggl[\frac{K_h^2}{k_h}\biggr].
\end{aligned} \tag{2.6}$$

The without-replacement variance estimator $\widehat{\text{var}}_{\text{wo}}(\bar{y})$ underestimates the (total) variability of $\bar{y}$ by the model variability of $\overline{Y}$, $\text{Var}_F[\overline{Y}]$. The underestimation by the with-replacement estimator $\widehat{\text{var}}_{\text{wr}}(\bar{y})$ is less, Δ_{betw}/K. Therefore, if $\Delta_{\text{betw}} = 0$, or if there is only one stratum, $\widehat{\text{var}}_{\text{wr}}(\bar{y})$ unbiasedly estimates the variability of $\bar{y}$. If the sampling fractions in the strata are not small and

Δ_{betw} is not small compared to the within-stratum variability of Y, then the underestimation of the variance estimators is not small. The negative biases of $\widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y})$ and $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y})$ are given by

$$\begin{aligned}\text{Negative bias of } \widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y}) &= \frac{\mathrm{Var}(\bar{y}) - E_F[\widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y})]}{\mathrm{Var}(\bar{y})} \\ &= \frac{\Delta_{\mathrm{betw}}}{\Delta_{\mathrm{betw}} + \sum_{h=1}^{L} \pi_h \sigma_h^2 / f_h}\end{aligned}$$

and

$$\begin{aligned}&\text{Negative bias of } \widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y}) \\ &\quad = \frac{\Delta_{\mathrm{betw}} + \sum_{h=1}^{L} \pi_h \sigma_h^2}{\Delta_{\mathrm{betw}} + \sum_{h=1}^{L} \pi_h \sigma_h^2 / f_h},\end{aligned} \tag{2.7}$$

where f_h is the sampling fraction in stratum h (assumed to be nonrandom for these expressions). For example, suppose the sampling fractions are 25% in each of the strata and $\Delta_{\mathrm{betw}} = \sum_{h=1}^{L} \pi_h \sigma_h^2$. Then the negative bias of $\widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y})$ is 20% and of $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y})$ is 40%.

To obtain an unbiased variance estimator of $\bar{y}$, one can add an unbiased estimator of $\mathrm{Var}_F[\overline{Y}]$ to $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y})$ or, equivalently, add an unbiased estimator of Δ_{betw}/K to $\widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y})$. In particular, let

$$\begin{aligned}\widehat{\mathrm{var}}[\overline{Y}] &= \frac{1}{K-1} \sum_{h=1}^{L} \frac{K_h}{K} (\bar{y}_h - \bar{y})^2 \\ &\quad + \frac{1}{K} \sum_{h=1}^{L} \frac{K_h}{K} \left[1 - \frac{K - K_h}{(K-1)k_h}\right] s_h^2\end{aligned} \tag{2.8}$$

and

$$\begin{aligned}\hat{\Delta}_{\mathrm{betw}} &= \frac{K}{K-1} \sum_{h=1}^{L} \frac{K_h}{K} (\bar{y}_h - \bar{y})^2 \\ &\quad - \sum_{h=1}^{L} \frac{K_h(K - K_h)}{K(K-1)} \frac{s_h^2}{k_h}.\end{aligned} \tag{2.9}$$

Then

$$\begin{aligned}E[\widehat{\mathrm{var}}[\overline{Y}]] &= E_F\left[\frac{1}{K-1} \sum_{h=1}^{L} \frac{K_h}{K} (\overline{Y}_h - \overline{Y})^2\right] \\ &\quad + E_F\left[\frac{1}{K} \sum_{h=1}^{L} \frac{K_h - 1}{K-1} S_h^2\right] \\ &= \mathrm{Var}_F[\overline{Y}]\end{aligned} \tag{2.10}$$

and $E[\hat{\Delta}_{\mathrm{betw}}] = \Delta_{\mathrm{betw}}$. Therefore,

$$\begin{aligned}\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y}) &\equiv \widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y}) + \widehat{\mathrm{var}}[\overline{Y}] \\ &= \widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y}) + \hat{\Delta}_{\mathrm{betw}}/K\end{aligned} \tag{2.11}$$

is an unbiased estimator for $\mathrm{Var}(\bar{y})$.

2.2 Stratified Simple Random Sampling with Proportional Allocation and Simple Random Sampling with Poststratification

Since the superpopulation variance estimator (2.11) contains a between-strata component, a natural question to ask is whether one can estimate this variance by using a standard with-replacement variance estimator that ignores the stratification (since this standard estimator also contains a between-strata component). This approach does not work, as easily seen by considering the case of stratified simple random sampling without replacement with proportional allocation, that is, stratified simple random sampling with $k_h/K_h \equiv k/K$. The superpopulation variance estimator (2.11) reduces to

$$\begin{aligned}\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y}) &= \frac{1}{K-1} \sum_{h=1}^{L} \frac{K_h}{K} (\bar{y}_h - \bar{y})^2 + \frac{1}{k} \sum_{h=1}^{L} \frac{K_h - 1}{K-1} s_h^2 \\ &\approx \frac{1}{K} \sum_{h=1}^{L} \frac{K_h}{K} (\bar{y}_h - \bar{y})^2 + \frac{1}{k} \sum_{h=1}^{L} \frac{K_h}{K} s_h^2,\end{aligned}$$

whereas the with-replacement simple-random-sampling estimator is given by

$$\begin{aligned}\frac{s^2}{k} &= \frac{1}{k} \frac{k}{k-1} \sum_{h=1}^{L} \frac{K_h}{K} (\bar{y}_h - \bar{y})^2 \\ &\quad + \frac{1}{k} \frac{k}{k-1} \sum_{h=1}^{L} \left(\frac{K_h}{K} - \frac{1}{k}\right) s_h^2 \\ &\approx \frac{1}{k} \sum_{h=1}^{L} \frac{K_h}{K} (\bar{y}_h - \bar{y})^2 + \frac{1}{k} \sum_{h=1}^{L} \frac{K_h}{K} s_h^2,\end{aligned}$$

which seriously overestimates the between-strata component. [In some applications where the number of sampled first-stage units is small, this overestimation may be considered acceptable to be able to estimate the variance with reasonable precision (Korn and Graubard, 1995).]

With simple random sampling and poststratification, the sampled units are classified into the strata after the sampling. As the population sizes of the strata are assumed known, the estimator (2.1) of the superpopulation mean can still be calculated. We refer to this estimator as the poststratified mean and denote it by $\bar{y}_{\mathrm{ps}}$. The repeated-sampling variance of $\bar{y}_{\mathrm{ps}}$ is given by (Hansen, Hurwitz and Madow, 1953, page 232)

$$\begin{aligned}\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y}_{\mathrm{ps}}) &= \frac{1}{k} \frac{K-k}{K} \sum_{h=1}^{L} \frac{K_h}{K} s_h^2 \\ &\quad + \frac{1}{k^2} \frac{K-k}{K} \sum_{h=1}^{L} \frac{K - K_h}{K} s_h^2.\end{aligned} \tag{2.12}$$

Note that the first term is the same as the variance estimator in the case of stratified simple random sampling without replacement with proportional allocation. The second term is of smaller order and also only involves within-stratum variability. Therefore, the bias of $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y}_{\mathrm{ps}})$ as an estimator of the (superpopulation) variability of $\bar{y}_{\mathrm{ps}}$ would be expected to be similar to the stratified sampling situation with proportional allocation.

Using the decomposition

$$\mathrm{Var}(\bar{y}_{\mathrm{ps}}) = E_F[\mathrm{Var}_{\mathrm{RS}}(\bar{y}_{\mathrm{ps}})] + \mathrm{Var}_F[E_{\mathrm{RS}}(\bar{y}_{\mathrm{ps}})],$$

one can calculate $\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y}_{\mathrm{ps}})$ as the sum of $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y}_{\mathrm{ps}})$ and an estimator of $\mathrm{Var}(\overline{Y})$. One such estimator of $\mathrm{Var}(\overline{Y})$ is s^2/k, although one might also consider the poststratified estimator

$$\widehat{\mathrm{var}}(\overline{Y}) = \frac{1}{K}\frac{k}{k-1}\left[\left(\sum_{h=1}^{L}\frac{K_h}{K}\bar{y}_{2h}\right) - \left(\sum_{h=1}^{L}\frac{K_h}{K}\bar{y}_h\right)^2\right],$$

where $\bar{y}_{2h}$ is the mean of the squares of the sampled y observations in stratum h.

2.3 Estimating a Superpopulation Ratio

The above results can be applied to statistics more complex than a mean by using linearization. The estimation of a ratio is considered here, and the estimation of linear regression coefficients and other parameters is considered in Section 2.4. The population values (U_i, X_i, η_i), $i = 1, \ldots, K$, are assumed to be independent and identically distributed, each with the same distribution as the random vector (U, X, η), which has trivariate distribution function F. The target parameter is the ratio $\rho = \mu_u/\mu_x \equiv E_F(U)/E_F(X)$, with μ_x assumed to be greater than 0. With stratified simple random sampling without replacement, the estimator of the ratio is

$$\bar{r} = \frac{\bar{u}}{\bar{x}} \equiv \frac{\sum_{h=1}^{L}(K_h/K)\bar{u}_h}{\sum_{h=1}^{L}(K_h/K)\bar{x}_h},$$

where $\bar{x}_h(\bar{u}_h)$ is the mean of the sampled $X(U)$ observations in stratum h. For the sampled observations, the linearized variate is defined by $z_i = (u_i - \bar{r}x_i)/\bar{x}$. Analogously to (2.2), the repeated-sampling variance estimator of $\bar{r}$ is

$$\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{r}) = \sum_{h=1}^{L}\frac{K_h^2}{K^2}\frac{K_h - k_h}{K_h}\frac{s_{zh}^2}{k_h}, \tag{2.13}$$

where s_{zh}^2 is the sample variance of the z's in the hth stratum. [An alternative derivation defines $z_i = [u_i - (\overline{U}/\overline{X})x_i]/\overline{X}$ where $\overline{U}$ and $\overline{X}$ are the population means, calculates the variance of $\bar{z}$, and then substitutes $\bar{r}$ for $\overline{U}/\overline{X}$ and $\bar{x}$ for $\overline{X}$ to obtain the variance estimator (2.13).] Under repeated sampling of the same finite population, $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{r})$ is an approximately unbiased estimator of the variance of $\bar{r}$ when the strata sample sizes are large; see below. The repeated-sampling variance estimator, treating the sample as if it had been a with-replacement sample, is

$$\widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{r}) = \sum_{h=1}^{L}\frac{K_h^2}{K^2}\frac{s_{zh}^2}{k_h}. \tag{2.14}$$

We consider asymptotic properties of the estimators with L fixed as k, $K \to \infty$, and $k_h/K_h \to \gamma_h$, $h = 1, \ldots, L$. Formally, a sequence of populations, samples and estimators is implicitly indexed by $\alpha \to \infty$. The approximately equal sign "$\cong$" below should be interpreted as meaning that k times the difference of the quantities on its right- and left-hand sides approaches 0 as $\alpha \to \infty$. (The limit of $K\,\mathrm{Var}(\bar{r})$ is also the variance of the asymptotic distribution of $\sqrt{K}\bar{r}$ if F has finite first and second moments.) Incorporating the randomness due to the sampling and the distribution function F, we have, analogously to (2.3), (2.5) and (2.6),

$$\mathrm{Var}(\bar{r}) \cong \frac{1}{K^2}\sum_{h=1}^{L}\sigma_{zh}^2 E_F\left[\frac{K_h^2}{k_h}\right] + \frac{1}{K}\Delta_{z\mathrm{betw}}, \tag{2.15}$$

$$E[\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{r})] \cong \frac{1}{K^2}\sum_{h=1}^{L}\sigma_{zh}^2 \cdot E_F\left[\frac{K_h(K_h - k_h)}{k_h}\right] \tag{2.16}$$

and

$$E[\widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{r})] \cong \frac{1}{K^2}\sum_{h=1}^{L}\sigma_{zh}^2 E_F\left[\frac{K_h^2}{k_h}\right], \tag{2.17}$$

where

$$\sigma_{zh}^2 = \mathrm{Var}_F\left(\frac{U - \rho X}{\mu_x} \mid \eta = h\right),$$

$$\Delta_{z\mathrm{betw}} = \sum_{h=1}^{L}\pi_h\left(\frac{\mu_{uh} - \rho\mu_{xh}}{\mu_x}\right)^2.$$

Here μ_{uh} and μ_{xh} are the expected values of U and X in the hth stratum with respect to F.

The quantity $\Delta_{z\mathrm{betw}}$ equals 0 if the within-stratum ratio μ_{uh}/μ_{xh} does not vary across the strata. Under this condition, the with-replacement variance estimator $\widehat{\text{var}}_{\text{wr}}(\bar{r})$ is asymptotically unbiased for $\text{Var}(\bar{r})$. Note that whether the μ_{uh}, μ_{xh} or the within-stratum correlation of U and X varies across the strata is irrelevant to this conclusion. A second interesting thing to note is that, unlike the case of estimating a simple mean, the repeated-sampling variance estimators (2.13) and (2.14) incorporate some additional variability when the target parameter varies from strata to strata. This implies that the negative bias of these variance estimators when estimating the variance of a ratio will not be as large as in the simple mean case. This is easiest to see for the case of estimating a mean of a variable (Y) over a subdomain (D) of the population. This is a special case of a ratio estimator, as seen by letting $U_i \equiv Y_i I(i \in D)$ and $X_i \equiv I(i \in D)$. As an example, suppose that (in obvious notation) $\sum \pi_h \sigma_{yh}^2 = \sum(\mu_{yh} - \mu_y)^2$, that the probability that an observation is in D is π_D in each strata and that the sampling fractions are 25% in each of the strata. Then the negative bias of $\widehat{\text{var}}_{\text{wr}}(\bar{r})$ is asymptotically $\pi_D/(8 - 3\pi_D)$, which can be notably smaller than the 20% when $\pi_D = 1$, for example, 7.7% when $\pi_D = 1/2$.

To obtain an asymptotically unbiased estimator of the variance of $\bar{r}$, one follows (2.11) and defines

$$\begin{aligned}\widehat{\text{var}}_{\text{SP}}(\bar{r}) &\equiv \widehat{\text{var}}_{\text{wo}}(\bar{r}) + \widehat{\text{var}}[\overline{R}] \\ &= \widehat{\text{var}}_{\text{wr}}(\bar{r}) + \hat{\Delta}_{z\text{betw}}/K,\end{aligned} \tag{2.18}$$

where $\overline{R} = \overline{U}/\overline{X}$,

$$\begin{aligned}\widehat{\text{var}}[\overline{R}] = {} & \frac{1}{K-1}\sum_{h=1}^{L}\frac{K_h}{K}\bar{z}_h^2 \\ & + \frac{1}{K}\sum_{h=1}^{L}\frac{K_h}{K}\left[1 - \frac{K-K_h}{(K-1)k_h}\right]s_{zh}^2\end{aligned}$$

and

$$\begin{aligned}\hat{\Delta}_{z\text{betw}} = {} & \frac{K}{K-1}\sum_{h=1}^{L}\frac{K_h}{K}\bar{z}_h^2 \\ & - \sum_{h=1}^{L}\frac{K_h(K-K_h)}{K(K-1)}\frac{s_{zh}^2}{k_h}.\end{aligned}$$

Note that $\bar{z} = 0$, so that the terms corresponding to $\bar{y}$ in (2.8) and (2.9) are not needed.

2.4 Estimating Superpopulation Linear Regression Coefficients and Other Superpopulation Parameters

The population values (Y_i, X_i, η_i), $i = 1, \dots, K$, are assumed to be independent and identically distributed, each with the same distribution as the random vector (Y, X, η), which has a multivariate distribution function F. Here the X_i and X are p-dimensional row vectors. The target parameter is the p-dimensional vector

$$\beta = \left[E_F(X'X)\right]^{-1} E_F(X'Y) \equiv \mu_{xx}^{-1}\mu_{xy},$$

where μ_{xx} is assumed to be an invertible $p \times p$ matrix. With stratified simple random sampling without replacement, the estimator of β is

$$\hat{\beta} = \left(\sum_{h=1}^{L}\frac{K_h}{K}\overline{xx}_h\right)^{-1}\sum_{h=1}^{L}\frac{K_h}{K}\overline{xy}_h,$$

where $\overline{xx}_h = \sum_{i=1}^{k_h} x_{hi}' x_{hi}/k_h$ and $\overline{xy}_h = \sum_{i=1}^{k_h} x_{hi}' \times y_{hi}/k_h$ are means over the sampled observations in stratum h.

For the ith sampled observation in stratum h, the linearized p-vector variate is defined by

$$z_{hi} = \left(\sum_{t=1}^{L}\frac{K_t}{K}\overline{xx}_t\right)^{-1}\left[x_{hi}'(y_{hi} - x_{hi}\hat{\beta})\right].$$

The repeated-sampling variance estimator of $\hat{\beta}$ is given by the right-hand side of (2.13), where now s_{zh}^2 is the $p \times p$ sample covariance of the z's in the hth stratum. The repeated-sampling variance estimator, treating the sample as if it had been a with-replacement sample, $\widehat{\text{var}}_{\text{wr}}(\hat{\beta})$ is given by the right-hand side of (2.14). The asymptotic variance of $\hat{\beta}$ and the asymptotic means of the variance estimators are given by the right-hand sides of (2.15), (2.16) and (2.17), where now

$$\sigma_{zh}^2 = \text{Var}_F\left[\mu_{xx}^{-1}(X'Y - X'X\beta) \mid \eta = h\right]$$

and

$$\begin{aligned}\Delta_{z\text{betw}} = {} & \sum_{h=1}^{L}\pi_h\left[\mu_{xx}^{-1}(\mu_{xyh} - \mu_{xxh}\beta)\right] \\ & \cdot\left[\mu_{xx}^{-1}(\mu_{xyh} - \mu_{xxh}\beta)\right]'.\end{aligned}$$

Here μ_{xxh} and μ_{xyh} are the expected values of $X'X$ and $X'Y$ in stratum h with respect to F.

The quantity $\Delta_{z\text{betw}}$ equals 0 if the within-stratum regression coefficient vector $\mu_{xxh}^{-1}\mu_{xyh}$ does not vary across the strata. Under this condition, $\widehat{\text{var}}_{\text{wr}}(\hat{\beta})$ is asymptotically unbiased for $\text{Var}(\hat{\beta})$. Note that, in general, the jth diagonal element of $\Delta_{z\text{betw}}$ corresponding to the variance of the jth estimated coefficient of X will not be 0 if the within-stratum regression coefficients of the other independent variables vary across the strata.

In particular, for a simple linear regression with an intercept and one covariate, the constancy of the within-stratum slopes across the strata is not sufficient to guarantee that the with-replacement variance estimator will be asymptotically unbiased for the variance of the estimated slope. What is required, in addition, is that either (1) the within-stratum intercepts do not vary across the strata or (2) the within-stratum means of the covariates do not vary across the strata. On the other hand, if strata effects are included completely in the linear regression model (as $L-1$ indicator variables), then $\Delta_{z\text{betw}} = 0$ so that the with-replacement variance estimators can be used for the estimated regression coefficients of the other nonindicator covariates.

In general, an asymptotically unbiased estimator of the variance of $\hat{\beta}$ can be obtained analogously to the estimator (2.18) for the variance estimator of the superpopulation ratio. The same approach works for other parameters after the linearized variates are calculated for the problem at hand. For example, for a logistic regression analysis with p covariates (including the intercept), the linearized p-vector variate is defined by

$$z_{hi} = \left(\sum_{t=1}^{L} \frac{K_t}{K} \frac{1}{k_t} \sum_{i=1}^{k_t} x'_{ti} x_{ti} \hat{p}_{ti}(1-\hat{p}_{ti}) \right)^{-1} \cdot \bigl[x'_{hi}(y_{hi} - \hat{p}_{hi})\bigr],$$

where $\hat{p}_{hi}$ is the estimated predicted probability that the outcome equals 1 for the ith sampled observation in stratum h.

2.5 Estimating a Superpopulation Mean with a Ratio Estimator

In many applications, information known about the population values of variables related to the variable of interest can be used to improve the estimation. In this section, we consider the ratio estimation of $\mu = E_F(U)$ from a sample obtaining U and X information, when certain population mean information about the auxiliary variable X is known. The stochastic model for the finite-population values is assumed to be the same as in Section 2.3. Note, in particular, that this model does not specify any particular relationship between U and X.

With stratified simple random sampling, Cochran (1977) suggests two possible ratio estimators, the combined ratio estimator, $\bar{u}_R = (\overline{X}/\bar{x})\bar{u}$, and the separate ratio estimator,

$$\bar{u}_R = \frac{1}{K} \sum_{h=1}^{L} \frac{\overline{X}_h}{\bar{x}_h} K_h \bar{u}_h.$$

Here $\bar{u}$ and $\bar{x}$ are stratified estimators of the means, and $\overline{X}$ ($\overline{X}_h$) is the mean of X over all the population units (in stratum h). For either estimator, we have the decomposition

$$\mathrm{Var}(\bar{u}_R) = E_F\bigl[\mathrm{Var}_{\mathrm{RS}}(\bar{u}_R)\bigr] + \mathrm{Var}_F\bigl[E_{\mathrm{RS}}(\bar{u}_R)\bigr]$$

so that the variance of $\bar{u}_R$ can be estimated from the sum of a standard without-replacement variance estimator of $\bar{u}_R$, $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{u}_R)$, plus an estimator of the (model) variance of $\overline{U}$, $\widehat{\mathrm{var}}[\overline{U}]$. One can use the estimator (2.8) (replacing y's with u's) as an estimator of $\widehat{\mathrm{var}}[\overline{U}]$. One could also consider a ratio-type estimator of $\widehat{\mathrm{var}}[\overline{U}]$, but ratio adjusting to the "wrong" covariate can lead to a very inefficient estimator of $\widehat{\mathrm{var}}[\overline{U}]$. (The choice of whether to use the combined ratio estimator or the separate ratio estimators typically depends on what is known about the relationship between U and X in different strata.)

2.6 Case–Control Sampling

In case–control sampling, risk factors are compared for a sample of cases (diseased individuals) and a sample of controls (nondiseased individuals). This type of sampling can be considered a special case of stratified simple random sampling with two strata, the case stratum and the control stratum. Consider comparing the association of a risk factor and the disease by examining the difference in sample means, $\bar{x}_{\text{case}} - \bar{x}_{\text{control}}$. The without-replacement finite-population variance estimator, which is appropriate for inference about $\overline{X}_{\text{case}} - \overline{X}_{\text{control}}$, will have finite-population correction factors. Dropping these factors yields the with-replacement variance estimator. Is this estimator appropriate for superpopulation inference for $\bar{\mu}_{\text{case}} - \bar{\mu}_{\text{control}}$, or does one need to add a between-strata component as in (2.11)? Since $\bar{x}_{\text{case}}$ and $\bar{x}_{\text{control}}$ are uncorrelated, one obtains the variance of the difference as the sum of the variances of the separate means. Since each of the means is estimating an unstratified subdomain mean, there is no between-strata component and the with-replacement variance estimator is the superpopulation variance estimator.

Frequently, case–control data will be analyzed with a logistic regression treating the case/control status as the outcome Y. For the intercept in such a regression to be estimating a meaningful quantity, the sampling fractions should be utilized in the estimation (Scott and Wild, 1986). Although not obvious, the with-replacement variance estimators can be used for superpopulation inference for the regression coefficients but not for the intercept. The superpopulation variance estimator based on the linearized variate described in Section 2.4 can be used for both the intercept and the regression coefficients. However, it would not be expected to be as stable a variance estimator for the regression coefficients as the with-replacement variance estimator.

The sampling could involve additional stratification beyond the case/control stratification. For example, within each of L "strata," one might sample cases and controls separately, yielding $2L$ sampling strata. The variability of the estimated risk difference could be estimated with a with-replacement variance estimator if one assumes there is no between-strata variability in the risk difference. However, the with-replacement variance estimator of the estimated logistic regression coefficient will generally be an underestimate of the estimated coefficient's superpopulation variability unless either (1) the coefficient and the intercept are constant across the "strata" or (2) the coefficient is constant across the "strata" and $L - 1$ indicator variables are included in the model for the "strata" effects. A different type of stratification is a stratified selection of cases and/or a stratified selection of controls. In this situation, there is no option for including "strata" effects in the model so that the with-replacement variance estimator will generally underestimate the superpopulation variance of the estimated risk-factor difference or the estimated logistic regression coefficient. The superpopulation variance estimator based on the linearized variate can be used in any of the above situations, but may be less efficient than the with-replacement variance estimator when that estimator can be used because of further model assumptions.

2.7 Model Strata Misspecification

To this point, it might appear that sampling strata need to be defined in the superpopulation (by F), raising questions concerning the results if the sampling strata and the superpopulation strata are not the same. In fact, although it was convenient for quantifying bias in Section 2.1, the following generalization shows it is unnecessary to define strata in the superpopulation for the unbiasedness of $\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y})$ given by (2.11). Let η be a (possibly multidimensional continuous) variable that is observed on all finite population units and this is used to determine the sampling strata in the finite population. Note that the strata can be determined *after* the finite population has been observed and need not be defined in the superpopulation. Using (2.4), (2.10) and

$$\begin{aligned}\mathrm{Var}(\bar{y}) &= E_F\left(\frac{1}{K-1}\sum_{h=1}^{L}\frac{K_h}{K}(\overline{Y}_h - \overline{Y})^2\right)\\ &\quad + E_F\left(\sum_{h=1}^{L}\frac{K_h^2}{K^2}\frac{S_h^2}{k_h}\right)\\ &\quad + E_F\left(\sum_{h=1}^{L}\frac{K_h - K}{K^2(K-1)}S_h^2\right),\end{aligned}$$

we still have the result that $\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y})$ is an unbiased estimator for $\mathrm{Var}(\bar{y})$. For example, 20 equallysized strata may be determined in the finite population by a clustering algorithm using η. Another example is given by case–control sampling, where Y is binary and $\eta = Y$ determines the two sampling strata.

2.8 Relationship with Two-Phase Sampling

In two-phase sampling (also known as double sampling), an initial sample of the finite population is drawn from which design variable information is obtained on each sampled unit. A subsample of the initial sample is then taken with a sample design that uses the design variable information. The variable of interest is recorded only on the subsample. For example, in two-phase sampling for stratification, the initial sample may be a simple random sample for which a stratification variable is recorded. This stratification variable is used to obtain a subsample using stratified simple random sampling. In the present context, we note that if we associate the two-phase initial sample with our finite population, and the two-phase finite population with our superpopulation (by letting its size approach ∞), then two-phase variance estimators for the "finite population" should yield our superpopulation variance estimators. This indeed can be verified for stratified simple random sampling; Cochran (1977, page 333), provides the following formula for the variance of a mean obtained from a two-phase stratified simple random sample:

$$\begin{aligned}v(\bar{y}) &= \frac{K(N^{\mathrm{FP}}-1)}{(K-1)N^{\mathrm{FP}}}\Bigg[\sum_{h=1}^{L} w_h s_h^2\left(\frac{1}{Kf_h} - \frac{1}{N^{\mathrm{FP}}}\right)\\ &\quad + \frac{g'}{K}\sum_{h=1}^{L} s_h^2\left(\frac{w_h}{N^{\mathrm{FP}}} - \frac{1}{Kf_h}\right)\\ &\quad + \frac{g'}{K}\sum_{h=1}^{L} w_h(\bar{y}_h - \bar{y})^2\Bigg],\end{aligned} \tag{2.19}$$

where N^{FP} is the size of the two-phase finite population, $w_h = K_h/K$ and $g' = (N_{\mathrm{FP}} - K)/(N_{\mathrm{FP}} - 1)$. Letting $N^{\mathrm{FP}} \to \infty$, (2.19) reduces to precisely the superpopulation variance estimator (2.11). In theory, general formulas for two-phase sampling (Särndal, Swensson and Wretman, 1992, page 351) could be similarly modified for superpopulation inference involving more complex sampling schemes.

3. MODEL WITH CLUSTERS

In this section we consider a model that accommodates multiple levels of clustering in the population, although only two levels will need to be explicitly modeled. Detailed derivations and asymptotic results are given in Korn and Graubard (1998). The population consists of K primary clusters, the ith of which consists of N_i secondary clusters. In the jth secondary cluster of the ith primary cluster, there are M_{ij} population values (Y's) with total $T_{ij} = Y_{ij1} + \cdots + Y_{ijM_{ij}}$. The ith primary cluster has associated with it a stratum variable $\eta_i \in \{1, \ldots, L\}$ and a "size" variable Z_i which will be used for probability-proportional-to-size (pps) sampling. We assume that the $\{(M_{ij}, T_{ij}) \mid j = 1, \ldots, N_i\}$ are independent and identically distributed with mean (α_i, τ_i) and variances–covariances $(\sigma_{11i},\ \sigma_{22i},\ \sigma_{12i})$ and that the $(\alpha_i, \tau_i, \sigma_{11i}, \sigma_{22i}, \sigma_{12i}, N_i, Z_i, \eta_i)$ are independent and identically distributed from an eight-dimensional random variable with distribution function G. The superpopulation mean is defined as $\mu = E_G(N\tau)/E_G(N\alpha)$. We present only formulas in this section for the superpopulation mean; extension to other parameters via linearization is immediate.

Section 3.1 describes the multistage sampling considered and the notation for the estimation of the superpopulation mean using a weighted mean. Section 3.2 presents methods of inference for the superpopulation mean. The negative bias associated with using a finite-population repeated-sampling variance estimator is described in Section 3.3. We end in Section 3.4 with a brief discussion of superpopulation model misspecification.

3.1 Estimating a Superpopulation Mean with Multistage Sampling

We assume initially that the multistage sampling is consistent with the clustering just described for the population; deviations are considered in Section 3.4. At the first stage of sampling, k_h primary clusters are sampled from the K_h primary clusters in stratum h as a pps sample without replacement. That is, the primary clusters are the primary sampling units (PSUs), and the inclusion probability of a given PSU in stratum h is proportional to its Z value. Stratified simple random sampling without replacement of PSUs is a special case of this sampling. At the second stage of sampling, we assume that n_{hi} secondary clusters are sampled with replacement from the ith sampled PSU from stratum h. The schemes for the third to final stages of sampling are left unspecified. However, we do assume that the sample weights (i.e., inverses of the inclusion probabilities) are known for all sampled observations.

Let

$$t_{hij} = \sum_{l=1}^{m_{hij}} w_{hijl} y_{hijl} \quad \text{and} \quad d_{hij} = \sum_{l=1}^{m_{hij}} w_{hijl},$$

where y_{hijl} and w_{hijl} are the population value and sample weight of the lth sampled observation in the jth sampled secondary cluster in the ith sampled PSU of stratum h, and m_{hij} is the number of sampled observations in that secondary cluster. Letting

$$t_{hi} = \sum_{j=1}^{n_{hi}} t_{hij} \quad \text{and} \quad d_{hi} = \sum_{j=1}^{n_{hi}} d_{hij},$$

we assume that under repeated sampling of the population

$$t = \sum_{h=1}^{L} \sum_{i=1}^{k_h} t_{hi} \quad \text{and} \quad d = \sum_{h=1}^{L} \sum_{i=1}^{k_h} d_{hi}$$

are approximately unbiased estimators of the total of Y and the population size, respectively. The weighted estimator of $\overline{Y}$ and μ is $\bar{y} = t/d$.

3.2 Inference for the Superpopulation Mean

Under repeated sampling of the same finite population, a variance estimator of $\bar{y}$ is given by Korn and Graubard (1998)

$$\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y}) = \frac{1}{d^2}\Biggl\{\sum_{h=1}^{L}\sum_{i=1}^{k_h}\sum_{j<i}^{k_h}\biggl[\frac{\lambda_{hi}\lambda_{hj}}{\lambda_{hij}} - 1\biggr] \cdot \bigl[(t_{hi} - \bar{y}d_{hi}) - (t_{hj} - \bar{y}d_{hj})\bigr]^2 + K s_w^2\Biggr\},$$

where

$$s_w^2 = \frac{1}{K}\sum_{h=1}^{L}\sum_{i=1}^{k_h} \lambda_{hi} n_{hi} s_{hi}^2 \tag{3.1}$$

and where λ_{hi} is the inclusion probability of the ith PSU in stratum h, λ_{hij} is the joint inclusion probability of the ith and jth PSUs in stratum h and

$$s_{hi}^2 = \frac{1}{n_{hi}-1}\sum_{j=1}^{n_{hi}}\big[(t_{hij}-\bar{y}d_{hij})-(t_{hi}-\bar{y}d_{hi})/n_{hi}\big]^2.$$

To avoid the necessity of specifying the joint inclusion probabilities for variance estimation, the sampling design is frequently approximated as a *with*-replacement stratified pps sample of PSUs (Durbin, 1953). The repeated-sampling variance estimator is given by

$$\widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y}) = \frac{1}{d^2}\sum_{h=1}^{L}\frac{k_h}{k_h-1} \cdot \sum_{i=1}^{k_h}\left[(t_{hi}-\bar{y}d_{hi}) - \frac{1}{k_h}\sum_{j=1}^{k_h}(t_{hj}-\bar{y}d_{hj})\right]^2.$$

To obtain an asymptotically unbiased estimator of the variance of $\bar{y}$ that incorporates the repeated-sampling variability as well as the model variability G, one can use $\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y}) = \widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y}) + \widehat{\mathrm{var}}[\bar{Y}]$, where

$$\widehat{\mathrm{var}}[\bar{Y}] = \frac{1}{d^2}\left[\frac{K}{K-1} \cdot \sum_{h=1}^{L}\sum_{i=1}^{k_h}\lambda_{hi}(t_{hi}-\bar{y}d_{hi})^2 - Ks_w^2\right] \tag{3.2}$$

and s_w^2 is given by (3.1). The conditions for this asymptotic result include the case of increasing numbers of sampled PSUs in a fixed number of strata and the case of increasing numbers of strata with a small number of PSUs (e.g., 2) sampled from each stratum.

An approximately unbiased variance estimator that does not require specifying the joint inclusion probabilities is given by $\widehat{\mathrm{var}}_{\mathrm{SP}-a}(\bar{y}) = \widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y}) + \widehat{\mathrm{var}}_b - \widehat{\mathrm{var}}_w$, where

$$\widehat{\mathrm{var}}_b = \frac{1}{d^2}\sum_{h=1}^{L}\frac{1}{K_h}\left[\sum_{i=1}^{k_h}(t_{hi}-\bar{y}d_{hi})\right]^2$$

and

$$\widehat{\mathrm{var}}_w = \frac{1}{d^2}\sum_{h=1}^{L}\frac{k_h}{K_h(k_h-1)} \cdot \sum_{i=1}^{k_h}\left[(t_{hi}-\bar{y}d_{hi}) - \frac{1}{k_h}\sum_{j=1}^{k_h}(t_{hj}-\bar{y}d_{hj})\right]^2.$$

In theory, both $\widehat{\mathrm{var}}[\bar{Y}]$ in (3.2) and $\widehat{\mathrm{var}}_b - \widehat{\mathrm{var}}_w$ are estimating nonnegative quantities and could be truncated to 0 if negative. Although this is an area for further study, we do not recommend truncation and prefer to retain approximate unbiasedness. The properties of $\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y})$ and $\widehat{\mathrm{var}}_{\mathrm{SP}-a}(\bar{y})$ are discussed further in Korn and Graubard (1998). Note that these estimators, like $\widehat{\mathrm{var}}_{\mathrm{wr}}(\bar{y})$ but unlike $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y})$, do not require estimation of the within-PSU variance component s_w^2. [This term cancels in the calculation of $\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y})$.] This is advantageous since the second-stage identifiers needed to calculate s_w^2 are frequently not publicly available because of confidentiality concerns.

An approximate 95% confidence interval for the superpopulation mean is given by

$$\bar{y} \pm t_d\big[\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y})\big]^{1/2}$$

or

$$\bar{y} \pm t_d\big[\widehat{\mathrm{var}}_{\mathrm{SP}-a}(\bar{y})\big]^{1/2},$$

where t_d is the 0.975 quantile of a t distribution with d degrees of freedom. An open question is what is the appropriate value for d? With with-replacement sampling of PSUs and a set of strong assumptions, d should be set equal to the number of sampled PSUs minus the number of sampling strata (Korn and Graubard, 1990). When the assumptions are weakened, this nominal degrees of freedom is potentially too large (Korn and Graubard, 1999, pages 197–199). With without-replacement sampling, it is not clear what degrees of freedom to associate with $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y})$ for finite-population inference; see Eltinge and Jang (1996) for some recent work. Similarly, it is not clear the degrees of freedom to associate with superpopulation inference. We use the following ad hoc approach: PSUs sampled with certainty are assumed to all come from the same single "certainty" stratum. The degrees of freedom d are then taken to be the number of sampled PSUs minus the number of strata. This is an area for further research.

3.3 The Negative Bias of Repeated-Sampling Estimators

To examine the effects of using a repeated-sampling variance estimator of the variability of $\bar{y}$ instead of an estimator of $\mathrm{Var}(\bar{y})$ [like $\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y})$ or $\widehat{\mathrm{var}}_{\mathrm{SP}-a}(\bar{y})$], we consider a special case of stratified two-stage cluster sampling with simple random sampling without replacement at the first stage and simple random sampling with replacement at the second stage. Let $M_{ij} \equiv 1$, $Y_{ij} \equiv T_{ij}$ and the population sizes and sample sizes of the PSUs be constant within each stratum, denoted by N_{h0} and n_{h0}, respectively. The proportional negative bias in using a repeated-sampling variance estimator is

$$\text{Negative bias} = \frac{\mathrm{Var}(\bar{y}) - E_G[\mathrm{Var}_{\mathrm{RS}}(\bar{y})]}{\mathrm{Var}(\bar{y})}$$
$$\cong \frac{\mathrm{Var}_G(\overline{Y})}{\mathrm{Var}_G(\overline{Y}) + E_G[\mathrm{Var}_{\mathrm{RS}}(\bar{y})]},$$

where $\mathrm{Var}(\bar{y}) \cong \mathrm{Var}_G(\overline{Y}) + E_G[\mathrm{Var}_{\mathrm{RS}}(\bar{y})]$ since $E_{\mathrm{RS}}(\bar{y}) \cong \overline{Y}$. We have

$$\mathrm{Var}_G(\overline{Y}) = \mathrm{Var}_G\, E_G(\overline{Y}|K_h) + E_G\, \mathrm{Var}_G(\overline{Y}|K_h) \cong B + W,$$

with the between- and within-strata components given by (using linearization)

$$B = \frac{1}{KE(N)^2} \sum_{h=1}^{L} \pi_h N_{h0}^2 [E(\tau|h) - \mu]^2$$

and

$$W = \frac{1}{KE(N)^2} \sum_{h=1}^{L} \pi_h N_{h0}^2 \biggl[\frac{E(\sigma_{22}|h)}{N_{h0}} + \mathrm{Var}(\tau|h)\biggr],$$

where $E(N) = \sum_{h=1}^{L} \pi_h N_{h0}$ and $\pi_h = P(\eta = h)$, the proportion of PSUs in stratum h. Using

$$E_G\, \mathrm{Var}_{\mathrm{RS}}(\bar{y}) \cong \frac{1}{KE(N)^2} \sum_{h=1}^{L} \pi_h N_{h0}^2 \cdot \biggl\{\biggl(\frac{1}{f_h} - 1\biggr)\biggl[\frac{E(\sigma_{22}|h)}{N_{h0}} + \mathrm{Var}(\tau|h)\biggr] + \frac{(N_{h0} - 1)E(\sigma_{22}|h)}{N_{h0} f_h n_{h0}}\biggr\},$$

where f_h is the sampling fraction of PSUs in stratum h, we have

$$\text{Negative bias} \cong (B + W)\biggl\{B + \frac{1}{KE(N)^2} \cdot \sum_{h=1}^{L} \frac{\pi_h N_{h0}^2}{f_h}\biggl[\frac{E(\sigma_{22}|h)}{N_{h0}} + \mathrm{Var}(\tau|h)\biggr] + \frac{1}{KE(N)^2} \sum_{h=1}^{L} \frac{\pi_h N_{h0}(N_{h0} - 1)}{f_h n_{h0}} E(\sigma_{22}|h)\biggr\}^{-1}. \tag{3.3}$$

Note that if $N_{h0} \equiv 1$ then (3.3) reduces to (2.7), appropriate for stratified simple random sampling [associating the $E(\tau|h)$ and $E(\sigma_{22}|h) + \mathrm{Var}(\tau|h)$ here with the μ_h and σ_h^2 there]. If $N_{h0} \equiv N_0$ and $B = W$, then (3.3) equals (2.7) provided that $E(\sigma_{22}|h) \equiv 0$. Otherwise, the negative bias (3.3) is less than the stratified simple-random-sampling case and is approximately 0 if $N_{h0} \gg n_{h0}$. When the N_{h0} vary with h, the negative bias depends on the particular parameter values. For example, suppose 4% of the PSUs have $N_{h0} = 10^6$ and are in strata with $f_h = 1$ and the remaining 96% of the PSUs have $N_{h0} = 5000$ and are in strata with $f_h = 1/16$. We continue to assume that $B = W$ and now also assume that $E(\sigma_{22}|h) \equiv c_1$ and $\mathrm{Var}(\tau|h) \equiv c_2$ for two constants, c_1 and c_2. If $c_1 = 0$ then the negative bias = 99.6%. If $c_2 = 0$ and $n_{h0} < 10{,}000$ then the negative bias is close to 0. If $c_2 = c_1/100$, and $n_{h0} \equiv 100$, then the negative bias is 66%. This example is important because it shows that, even with a relatively small overall sampling fraction of PSUs (10%), the negative bias can be substantial when a small proportion of large PSUs are sampled with high sampling fractions. This is the situation in the surveys used in the examples given in Sections 5.1 and 5.2 and is not uncommon.

3.4 Superpopulation Model Misspecification

For stratification, the model-misspecification issue is the same as previously discussed for models without clusters; there is no problem with strata misspecification since one need not actually define the strata in the superpopulation (although it is convenient for examining the properties of the estimators); see Section 2.7. For clustering, if the actual primary clusters in the superpopulation do not correspond to the PSUs (which are the modeled ones), then there can be bias in the proposed variance estimators. For example, suppose the PSUs are census tracts and they are modeled as the primary clusters in the superpopulation model. If there are larger clusters (e.g., counties) that are inducing an intraclass correlation of the data from different census tracts in the same county, then the proposed variance estimators will be underestimates of true variability. On the other hand, clustering that occurs within the PSUs does not lead to bias even when unaccounted for in the superpopulation model. For example, the proposed variance estimators will be approximately unbiased even if there is intraclass correlation related to blocks (within census tracts) that is not explicitly modeled. The major concern with model misspecification, therefore, is that there are larger clusters than the PSUs that are inducing an intraclass correlation between PSUs. However, it can be shown that the magnitude of the bias is a function of the population sizes of these larger clusters times the intraclass correlation.

In our experience, this product tends to be small, minimizing the potential for bias of our proposed variance estimators; see Korn and Graubard (1998) for further details.

4. RELATED WORK CONCERNING VARIANCE ESTIMATORS FOR SUPERPOPULATION PARAMETERS

Classical parametric statistics is concerned with stochastic models for observed data as representative of a population, and even classical nonparametric statistics can be viewed in this context (Lehmann, 1975, pages 55–65). However, we consider here only (published) work where the targets of inference are parameters associated with stochastic models for the units in a finite population, which is then sampled to obtain the observed data [case 3 of Hartley and Sielken (1975)]. We are especially interested in noting if authors omit finite-population correction factors (like us) and, if they consider stratification, whether they have a between-strata component of their variance estimators (as we do). We consider in turn previous work involving models without clusters and with clusters.

4.1 Models without Clusters

4.1.1 *Means.* Konijn (1962) considers the finite population to be sampled as a proportionate stratified sample from a superpopulation stratified into "classes." His target parameter is the weighted mean of the superpopulation slopes from each class, where the weights are the finite-population sizes of the classes. One of the situations he considers is stratified simple random sampling with the classes being treated as sampling strata. For estimating a mean, where Konijn's parameter estimator and ours coincide, his variance estimators differ from ours because he uses the finite-population sizes to define his target parameter rather than the superpopulation probabilities (the π_h of Section 2.1). In particular, his variance estimator does not have a between-strata component.

The target parameter of Potthoff, Woodbury and Manton (1992) is the sample-weighted mean of the superpopulation means of the sampled units; see also Bouza (1995) and Longford (1996). Like the model of Konijn (1962), this parameter will vary depending on the realized finite population. In addition, it will vary depending on the particular sample obtained (Kott, 1993), and thus it is very different from the superpopulation parameters considered in this paper.

Porter (1973) considers each unit in the population to have a vector of independent variables, a vector of linear regression coefficients and a dependent variable generated stochastically from a linear regression model. His target parameter is the (vector) mean of the regression coefficients over the units in the population. Restricting attention to a linear regression model with only an intercept, his target parameter can be considered a finite-population mean with allowance for measurement error. Because of this, his variance estimator of the sample mean for simple random sampling without replacement differs from ours in that it contains a finite-population correction factor for the component of variance not associated with the measurement error.

Särndal (1980) considers the N finite population values to be independent random variables $Y_1, \dots, Y_N$ with means $\mu_1, \dots, \mu_N$ determined by a linear regression model. He discusses the differences between estimators of $\overline{Y}$ and $(\mu_1 + \cdots + \mu_N)/N$, but does not discuss variance estimation. For a more general model, Haslett (1985) discusses the variance estimation of $(\mu_1 + \cdots + \mu_N)/N$, but his putative variance estimators appear to include superpopulation parameters. Koop (1986) discusses inference for the superpopulation mean, but assumes a normal distribution for the superpopulation values when considering variance estimation. Arnab (1992) allows for the superpopulation values to be correlated, but does not consider variance estimation.

4.1.2 *Linear regression coefficients.* Klein and Morgan (1951) consider, associated with each unit in the finite population, a vector of independent variables and a dependent variable generated stochastically from a linear regression model with independent and homoscedastic errors. Their target parameter is the common vector of (superpopulation) linear regression coefficients. Assuming stratified sampling, they consider estimating the regression coefficients with standard weighted estimators. Implicitly assuming noninformative sampling, they construct sandwich-type variance estimators. Their variance estimators do not have finite-population correction factors and do incorporate a between-strata component, but differ from our variance estimators because they utilize the noninformative-sampling and homoscedasticity assumptions.

Fuller (1975) shows that, in the nonstratified unclustered sampling situation, it is asymptotically correct to used standard survey variance formulas without the finite-population correction factors for inference about superpopulation regression coefficients. This is consistent with our results. (With stratified sampling, he presents results only for finite-population regression coefficients.)

Thomsen (1978) considers each unit in the finite population to have a dependent variable and independent variables generated stochastically from a linear regression model with homoscedastic errors. His target parameter is the (superpopulation) slope and intercept. Restricting attention to noninformative sampling, he utilizes ordinary-least-squares estimators and variance estimators.

DeMets and Halperin (1977) consider 3-dimensional vectors (X_1, X_2, X_3) in the finite population to have a trivariate normal distribution. The target parameter is the superpopulation coefficient of X_2 from the linear regression of X_1 on X_2. Assuming that the observed data are sampled from the finite population with probabilities that depend only on X_3 and that the values of X_3 are known for all units in the finite population, they derive the maximum likelihood estimator of the superpopulation regression coefficient; see also Holt, Smith and Winter (1980) and Chambers (1986). Nathan and Holt (1980) and Quesenberry and Jewell (1986) use the same framework except they use weaker assumptions than trivariate normality: Nathan and Holt (1980) assume that the conditional expectations of X_1 and X_2, given X_3, are linear in X_3 and show that the normal-theory estimator of DeMets and Halperin (1977) is asymptotically unbiased even with this weaker assumption; see also Pfeffermann and Holmes (1985). Quesenberry and Jewell (1986) assume that the conditional expectations of X_1 given X_2 and X_3 given X_2 are linear in X_2. They derive an estimator of the target parameter that depends on the estimation of a (superpopulation) residual distribution.

The variance estimators of the maximum likelihood estimators of Holt, Smith and Winter (1980) do not contain finite-population correction factors. DeMets and Halperin (1977), Nathan and Holt (1980) and Pfeffermann and Holmes (1985) do not explicitly consider variance estimators, although they could be constructed from their variance calculations. Presumably, such variance estimators would also not have finite-population correction factors. Chambers (1986), although providing no explicit variance calculations, suggests one might use a replication-based technique like the jackknife. His approach would lead to variance estimators for stratified sampling that have no between-strata component. Quesenberry and Jewell (1986) do not appear to suggest a variance estimator for their estimator of the key target parameter.

Besides considering a maximum likelihood approach, Holt, Smith and Winter (1980) consider the (superpopulation) variance of the usual weighted regression coefficient under stratified sampling. However, they assume the sampling fraction is small, so that the usual design-based variance estimator (with finite-population correction factors and no between-strata component) is reasonable.

In the context of stratified sampling, DuMouchel and Duncan (1983) consider various possible target parameters associated with linear regression models, one of which is the superpopulation regression coefficient vector. Since they assume that the sampling fractions in the strata are negligible, the issues concerning finite-population correction factors and between-strata variance components do not arise.

Ten Cate (1986) considers each unit in the finite population to have a vector of independent variables and a dependent variable generated stochastically from a linear regression model with independent and identically distributed errors. His target parameter is the vector of (superpopulation) linear regression coefficients. Assuming the observed data are sampled from the finite population with probabilities that can depend on the dependent variable, he estimates the linear regression coefficients with standard weighted estimators. His variance estimators appropriately account for sampling and model variability, but rely on the homoscedasticity of the errors, and also apparently require the specification of a parametric distribution for them. Hausman and Wise (1981) and Jewell (1985) consider a similar framework but restrict attention to stratified sampling. Hausman and Wise (1981) assume that the residual distribution is normal. Jewell (1985) estimates the (superpopulation) residual distribution more generally but provides no variance calculations.

4.1.3 *Other regression coefficients.* Scott and Wild (1989, 1991) consider stratified case–control data analyzed with logistic regressions. Their variance estimators do not have finite-population correction factors or a between-strata component. This is consistent with our results because they assume that the sampling fractions are negligible or that the strata effects are modeled correctly.

Nordberg (1989) considers superpopulation parameters in the setting of generalized linear models, but only considers variance formulas for unweighted estimators. The variance estimators he considers (with noninformative sampling) do not have finite-population correction factors.

Skinner (1994) considers superpopulation parameters in a general regression setting. His estimation approach involves modeling the sample weights as a function of the independent variables. He mentions replication methods as a possibility for estimating the variances of his parameter estimators. With stratified sampling, these methods would not include a between-strata component.

4.1.4 *Other parameters.* Patil and Rao (1978) consider the finite-population values of the variable of interest, as well as possibly a secondary variable, to be realizations from random vectors with a parametric distribution function. Their target parameters are (a subset of) the parameters associated with this distribution. The finite population is sampled with probability proportional to size, with the size variable being a known function of the variable of interest or the secondary variable. This framework induces a parametric distribution on the sampled observations, which can be used for inference; see also Rao (1965), Krieger and Pfeffermann (1992), Breckling et al. (1994) and Chambers, Dorfman and Wang (1998). As the parametric distribution of the sampled observations is known (up to the unknown parameters) with this "size-biased sampling," there is no need to consider finite-population correction factors.

Cosslett (1981) discusses estimation of general parameters with choice-based sampling. He assumes that the sampling fractions are negligible, so the issues discussed here do not arise.

Godambe and Thompson (1986) consider superpopulation parameters in a general modeling setting by the use of estimating equations; they do not discuss variance estimators.

4.2 Models with Clusters

With the "classes" of Konijn (1962) being clusters, he considers two-stage cluster sampling with simple random sampling at each stage. His target parameter is the weighted mean of the superpopulation slopes from each cluster, where the weights are the finite-population sizes of the clusters. For estimating a mean, where Konijn's parameter estimator and ours coincide, his variance estimator for the weighted mean has a finite-population correction factor associated with the sampling of the clusters (because his weights are the finite-population sizes of the classes). Pfeffermann and Nathan (1981) consider the target parameter to be an (arbitrarily) weighted linear combination of the superpopulation slopes from each cluster. The weight of each cluster in the population is assumed known, even if not sampled. They do not discuss variance estimators.

Sedransk (1965), Campbell (1977), Holt and Scott (1981), Scott and Holt (1982), Christensen (1984) and Goldstein (1986) consider linear regression models for the *sampled* observations that account for clustering by using a random effect (Sedransk, 1965; Goldstein, 1986), a fixed effect (Christensen, 1984) or a (common) intracluster correlation of the residuals (Campbell, 1977; Holt and Scott, 1981; Scott and Holt, 1982). With noninformative sampling of the clusters and noninformative sampling of the units within cluster, these models for the sampled data can also be considered models for the finite-population values. Their variance formulas do not include finite-population correction factors. [Christensen (1987) includes in his models independent variables that are cluster sample means of unit-level variables. Even with noninformative sampling, his models cannot be considered models for the finite-population values.]

Graubard and Korn (1996a) and Pfeffermann et al. (1998) consider models for the *population* values that account for the clustering with a random effect. Their variance estimators do not contain finite-population correction factors. (They do not present variance estimators for stratified sampling.)

Kott (1991) accommodates stratified multistage sampling and shows that when the sample weights are noninformative the usual design-based with-replacement variance estimators are appropriate. This is consistent with our results, since he assumes that the superpopulation regression coefficient vector (including the intercept) is constant across the strata.

Magee (1998, Appendix B) accommodates "subsamples" (strata or clusters) with modified weighted estimators. His weighted-least-squares-type variance calculations assume that the regression coefficient vector (including the intercept) is constant across the strata, and are therefore not comparable to our calculations.

5. EXAMPLES

We first consider examples of inference for superpopulation means using three national health surveys of the United States: the 1987 National Health Interview Survey (NHIS), a household interview survey; the third National Health and Nutrition Examination Survey (NHANES III), a medical examination survey; and the 1986 National Hospital Discharge Survey (NHDS), an institutional survey involving hospitals. This is followed by examples of inference involving linear and logistic regression using the 1987 NHIS. These examples are meant to demonstrate the types of effects that may be seen and some complicating issues and are not meant to be substantive analyses. Each of the surveys used in these examples has stratified cluster sample designs. Therefore, we use the superpopulation model described in Section 3. There is no need to assume that the sample strata are the same as the superpopulation strata (see Section 3.4). In these examples, we assume that the primary clusters in the superpopulation correspond to the PSUs. Section 3.4 shows that the superpopulation variance estimators are robust with respect to this assumption. The finite-population variances in the examples (i.e., $\widehat{\text{var}}_{\text{wo}}$ and $\widehat{\text{var}}_{\text{wr}}$) were calculated using SUDAAN (Shah, Barnwell and Bieler, 1997).

5.1 Means of Some Characteristics Measured in the 1987 NHIS

The design of this survey was multistage with the first stage involving the selection of PSUs consisting of individual counties, two or more contiguous counties or metropolitan statistical areas (Massey, Moore, Parsons and Tadros, 1989). Out of the 1894 PSUs comprising the population, 52 of the largest were sampled with certainty. The remaining 1842 PSUs were classified into 73 strata, from each of which two PSUs were sampled without replacement with probability proportional to size. Within sampled PSUs, secondary sampling units consisted of census enumeration districts, which were further subsampled, leading eventually to all eligible individuals in a sampled housing unit being interviewed. As part of the subsampling of some of the sampled PSUs, geographic areas with the highest concentrations of black persons were oversampled. For the analyses here, we use data from the Cancer Control Supplement which was given to individuals aged 18 years or older for half the sampled housing units (Schoenborn and Marano, 1988). We consider estimating the mean family income, mean height, proportion of women reporting having mammograms in the past year and the proportion of females in the population. The analysis was restricted to women aged 40 years or older for the mammogram variable.

Table 1 displays the weighted means for the selected variables as well as their standard errors estimated in different ways. It can be seen in lines 3 and 4 of the table that the underestimation of the standard errors based on the without-replacement variance estimators as compared to the superpopulation variance estimators can range from large (e.g., income) to negligible (e.g., sex). The fact that there can be large differences in these estimators may appear surprising because the sampling fraction of PSUs is a relatively small 10% (= 198/1894). However, over 80% of the population is in the 52 PSUs that were sampled with certainty, so large differences are possible; see the earlier discussion in section 3.3. One suspects that the large difference seen in variance estimators for income as compared to the other variables is due to large between-strata differences in income level as compared to the other variables; income was one of the variables used to form the

TABLE 1
Weighted means and standard errors for four variables using data from the 1987 *National Health Interview Survey*

	Income (dollars)	**Height (inches)**	**Mammograms (proportion within last year)**	**Sex proportion female)**
Sample size	19,561	21,944	6,522	22,043
Weighted mean	28,064	66.9	0.16	0.53
Standard error estimator				
Original data				
$[\widehat{\text{var}}_{\text{wo}}(\bar{y})]^{1/2}$	193	0.0328	0.00518	0.00402
$[\widehat{\text{var}}_{\text{SP}}(\bar{y})]^{1/2}$	515	0.0496	0.00611	0.00437
Certainty PSUs paired				
$[\widehat{\text{var}}_{\text{wr}}(\bar{y})]^{1/2}$	385	0.0461	0.00560	0.00417
$[\widehat{\text{var}}_{\text{SP}-a}(\bar{y})]^{1/2}$	516	0.0499	0.00611	0.00440
Certainty PSUs divided				
$[\widehat{\text{var}}_{\text{wr}}(\bar{y})]^{1/2}$	197	0.0335	0.00518	0.00404

strata in the 1987 NHIS. This suspicion was partially confirmed with ad hoc components of variance models for the variables (not shown). The small difference in the variance estimators for sex and larger differences for height and mammogram are because the sex distribution is very similar across geographical areas in the US, but height is related to race (NHANES III http://www.cdc.gov/nchs/about/major/nhanes/hgtfem.pdf) and mammogram rates are related to education (Breen and Kessler, 1994); race (% Hispanic) and economic variables (income and % below poverty), which correlate with education, were used to form the sampling strata.

We take 124 [$= 73 + (52 - 1)$] to be the degrees of freedom associated with the superpopulation variance estimators. Therefore, for example, a 95% confidence interval for the superpopulation mean income is given by $28,064 \pm 1.98 \times 515$. On the other hand, a 95% confidence interval for the finite-population mean income is given by $28,064 \pm 1.98 \times 193$ [if we also associate 124 degrees of freedom with $\widehat{\mathrm{var}}_{\mathrm{wo}}(\bar{y})$].

The last three lines of Table 1 are based on with-replacement variance estimators that would typically be used when the PSU joint inclusion probabilities were not available; the certainty PSUs were paired or divided so that there were two or more pseudo-PSUs in each pseudo-stratum. The with-replacement variance estimators that use the pairing of the certainty PSUs into pseudo-strata incorporate some between-strata variability; they therefore are intermediate to the without-replacement and superpopulation estimators. The approximate superpopulation variance estimators in line 6, which are based on the with-replacement estimators, are almost exactly equal to the superpopulation estimators in line 4. The with-replacement variance estimators on line 7 treat each census enumeration district in a certainty PSU as a pseudo-PSU.

5.2 Means of Some Characteristics Measured in the NHANES III

This multistage survey was conducted between 1988 and 1994 with the first stage involving the selection of PSUs consisting of counties or two or more adjacent counties (Ezzati et al., 1992; National Center for Health Statistics, 1994). Out of the 2962 PSUs comprising the population, 13 large PSUs were sampled with certainty. The remaining 2949 PSUs were classified into 34 strata, from each of which two PSUs were sampled without replacement with probability proportional to size. Within sampled PSUs, secondary sampling units consisted of area segments consisting of city or suburban blocks or other contiguous geographic areas. Households within these segments were subsampled, as well as individuals within the selected households. The sampling was done in a manner so that black American and Mexican–American groups as well as the very young and very old were sampled at higher rates. Sampled persons completing a household interview were invited to have a medical examination at a mobile examination center. For the analyses here, we use data from individuals aged 18 years or older. We consider estimating the mean family income, mean (self-reported) height, mean bone density (“bone mineral density of total region”), and the proportion of females in the population.

Table 2 displays the weighted means for the selected variables as well as their standard errors estimated in different ways. The only variable with a substantial underestimation of its standard error based on the without-replacement variance estimator as compared to the superpopulation variance estimator is income. Even with income, the underestimation is considerably smaller than seen for income measured in the 1987 NHIS (Table 1). One possible cause of this is that the between-strata variability of income could be larger in the 1987 NHIS than in the NHANES III. However, an ad hoc variance components model suggests that this is not the case (not shown). Instead, we believe that the smaller underestimation is because the estimated percentage of the relevant population in certainty PSUs is only 12% for the NHANES III (as opposed to 80% for the 1987 NHIS), and the overall sampling fraction of PSUs is also smaller in the NHANES III, 3% ($= 81/2962$).

5.3 Percentage of Hospital Discharges in the United States with Medicaid as the Principal Source of Payment Measured in the 1986 NHDS

The NHDS samples nonfederal short-stay hospitals and discharges from the sampled hospitals. It has been in continuous operation since 1965. For 1986 and before, the design can be approximated by a stratified simple random sample of hospitals, followed by simple random samples of discharges from these hospitals every year (Graves, 1988). The sampling strata are given in Table 3 and were based on region of the country and bed-size categories for the initial sample of hospitals in 1965. For hospitals added to the universe and sampled in later years, only bed-size strata were used. For 1986, the sample consisted of 558 hospitals (out of a population of 8178 hospitals)

TABLE 2

Weighted means and standard errors for four variables using data from the third National Health and Nutrition Examination Survey

	Income (dollars)	**Height (inches)**	**Bone density ($\mathrm{gm/cm}^2$)**	**Sex (proportion female)**
Sample size	17,513	18,724	14,646	19,618
Weighted mean	31,479	66.7	0.94	0.52
Standard error estimator				
Original data				
$[\widehat{\mathrm{var}}_{\mathrm{WO}}(\bar{y})]^{1/2}$	473	0.0501	0.00299	0.00435
$[\widehat{\mathrm{var}}_{\mathrm{SP}}(\bar{y})]^{1/2}$	533	0.0521	0.00309	0.00435
Certainty PSUs paired				
$[\widehat{\mathrm{var}}_{\mathrm{WR}}(\bar{y})]^{1/2}$	557	0.0507	0.00306	0.00409
$[\widehat{\mathrm{var}}_{\mathrm{SP}-a}(\bar{y})]^{1/2}$	557	0.0523	0.00314	0.00430
Certainty PSUs divided				
$[\widehat{\mathrm{var}}_{\mathrm{WR}}(\bar{y})]^{1/2}$	534	0.0507	0.00306	0.00425

TABLE 3

Number of hospitals in the 1986 *National Hospital Discharge universe and number of sampled responding in-scope hospitals by sampling strata*

	Hospitals sampled in 1965				**Hospitals sampled later**
Bedsize	**Northeast**	**Midwest**	**South**	**West**	
6–49 beds					
Universe	199	830	1438	646	449
Sample	5	10	12	4	10
50–99 beds					
Universe	288	442	587	306	326
Sample	8	13	16	8	16
100–199 beds					
Universe	277	378	332	157	314
Sample	20	23	20	9	23
200–299 beds					
Universe	182	151	134	85	90
Sample	23	19	12	9	13
300–499 beds					
Universe	110	129	96	51	28
Sample	17	25	21	8	8
500–999 beds					
Universe	42	46	28	13	6
Sample	13	16	10	7	3
≥1000 beds					
Universe	9	3	5	1	0
Sample	8	3	5	1	0

of which 75 refused to participate and 65 were out of scope. The sampling strata of the remaining 418 hospitals are given in Table 3; these hospitals provided information on approximately 193,000 discharges (out of a population of 34.3 million discharges).

The public-use data files for the 1986 NHDS contain a sample weight for each discharge. These weights incorporate various ratio and nonresponse adjustments as well as representing the sampling rates (Simmons and Schnack, 1970). We modify these weights to

TABLE 4

Percentage of hospital discharges for which Medicaid was the principal expected source of payment, based on data from the 1986 *National Hospital Discharge Survey*

	Estimates using:		**Standard errors estimated using:**			
	Original weights	**Modified weights**	$\widehat{\mathbf{var}}_{\mathbf{wo}}(\bar{\mathbf{y}})$	$\widehat{\mathbf{var}}_{\mathbf{SP}}(\bar{\mathbf{y}})$	$\widehat{\mathbf{var}}_{\mathbf{SP-a}}(\bar{\mathbf{y}})$	$\widehat{\mathbf{var}}_{\mathbf{wr}}(\bar{\mathbf{y}})$
All hospitals	10.400%	10.398%	0.429%	0.463%	0.463%	0.456%
Hospitals with bed size ≥500	11.993%	11.994%	0.917%	1.094%	1.094%	1.064%

be consistent with two-stage stratified simple random sampling by setting all the weights for the discharges in a given hospital to be equal to the mean sample weight of these discharges. For the purposes of variance estimation, the inclusion probabilities for the hospitals are taken to be the inverses of the sampling fractions given in Table 3, and the second-stage conditional inclusion probabilities for the discharges are taken as the inverses of the hospital inclusion probabilities divided by the modified discharge weight. Note that, for the calculation of the inclusion probabilities of the hospitals, sampled nonresponding hospitals are treated as if they were in the universe but unsampled. Thus, with the usual missing data assumptions, the resulting inferences are for all hospitals and not just for the universe of hospitals that would respond if sampled. In addition, the universe figures in Table 3 include some hospitals that were out of scope in 1986. If information had been available about which hospitals they were, we would have reduced the universe figures accordingly. Although we present point estimates using both the original sample weights and the modified weights, the variance estimators use the modified weights.

The first row of Table 4 contains estimates of the percentage of discharges for which Medicaid was the principal expected source of payment. The point estimates using the original sample weights and the modified weights are almost identical. The finite-population repeated-sampling standard error ($[\widehat{\text{var}}_{\text{wo}}(\bar{y})]^{1/2}$) is 7% smaller than the superpopulation standard error ($[\widehat{\text{var}}_{\text{SP}}(\bar{y})]^{1/2}$). To calculate the standard error based on a with-replacement variance estimator $[\widehat{\text{var}}_{\text{wr}}(\bar{y})]$ or based on the approximate superpopulation variance $[\widehat{\text{var}}_{\text{SP}-a}(\bar{y})]$ requires at least two sampled hospitals in each stratum. Since the largest bed-size stratum for the western region has only one sampled hospital, we pooled this stratum with largest bed-size stratum for the southern region for calculation of these two standard errors. The standard error based on the approximate superpopulation variance ($\widehat{\text{var}}_{\text{SP}-a}(\bar{y})$) is almost identical to the one based on the superpopulation variance $[\widehat{\text{var}}_{\text{SP}}(\bar{y})]$, although there is no need for using the approximation in this application since with stratified simple random sampling the joint inclusion probabilities are self-evident.

The second row of Table 4 displays results restricting consideration to hospitals with 500 or more beds in 1986. (This set of 91 sampled hospitals is not exactly the same as the set of 66 sampled hospitals with 500 or more beds given in Table 3 because that table refers to bed size at the time of sampling of the hospitals.) Since the sampling fraction of the larger hospitals is larger, the difference between the standard errors is larger—the finite-population repeated-sampling standard error is 16% smaller than the superpopulation variance estimator.

5.4 Some Linear and Logistic Regressions Using the 1987 NHIS

Table 5 presents the results from four simple linear regressions. The underestimation of the without-replacement variance estimators is much smaller for the regression coefficients in this table than for the simple means of income and height given in Table 1. The inherent variability of all the variance estimators makes any additional definitive statements suspect. However, it does appear that regressing on sex eliminates the underestimation more than regressing on race. This is understandable if one thinks that income (and height) levels for men and women may be higher and lower together in primary sampling units, reducing between-PSU differences in sex differences in income (and height). Table 6 presents the results of the multiple linear regression of income on race and sex. The results look similar to the results in Table 5.

Table 7 presents the results of the multiple linear regression of income on race, sex and the race-by-sex interaction. The interesting finding here is that the superpopulation variance estimator appears considerably smaller than the without-replacement estimator.

TABLE 5
Weighted regression coefficients and standard errors for four simple linear regressions using data from the 1987 *National Health Interview Survey*

Dependent variable	**Income**	**Income**	**Height**	**Height**
Independent variable	**Race (white vs. nonwhite)**	**Sex (men vs. women)**	**Race (white vs. nonwhite)**	**Sex (men vs. women)**
Sample size	19,561	19,561	21,944	21,944
Weighted regression coefficient	7,102	3,138	0.67	5.71
Standard error estimator				
Original data				
$[\widehat{\text{var}}_{\text{WO}}(\hat{\beta})]^{1/2}$	488	290	0.0953	0.0435
$[\widehat{\text{var}}_{\text{SP}}(\hat{\beta})]^{1/2}$	585	268	0.1107	0.0444
Certainty PSUs paired				
$[\widehat{\text{var}}_{\text{WR}}(\hat{\beta})]^{1/2}$	494	253	0.1153	0.0431
$[\widehat{\text{var}}_{\text{SP}-a}(\hat{\beta})]^{1/2}$	516	271	0.1112	0.0454
Certainty PSUs divided				
$[\widehat{\text{var}}_{\text{WR}}(\hat{\beta})]^{1/2}$	585	293	0.0964	0.0435

TABLE 6
Weighted regression coefficients and standard errors for the multiple linear regression of income on race and sex using data from the 1987 *National Health Interview Survey* (*sample size* = 19.561)

Independent variable	**Race (white vs. nonwhite)**	**Sex (men vs. women)**
Weighted regression coefficient	7018	3047
Standard error estimator		
Original data		
$[\widehat{\text{var}}_{\text{WO}}(\hat{\beta})]^{1/2}$	486	286
$[\widehat{\text{var}}_{\text{SP}}(\hat{\beta})]^{1/2}$	580	259
Certainty PSUs paired		
$[\widehat{\text{var}}_{\text{WR}}(\hat{\beta})]^{1/2}$	500	250
$[\widehat{\text{var}}_{\text{SP}-a}(\hat{\beta})]^{1/2}$	581	261
Certainty PSUs divided		
$[\widehat{\text{var}}_{\text{WR}}(\hat{\beta})]^{1/2}$	493	288

The with-replacement estimator with certainty PSUs paired is also considerably smaller than the without-replacement estimator, another counterintuitive finding. Both findings are consistent with the variability of a certain set of residuals between second-stage sampling units within certainty PSUs being larger than would be expected from the variability of these residuals between PSUs. Extensive simulations may be required to understand which differences in variance estimators seen in Table 7 are real and which are just due to the variability of the variance estimators. In any case, note that the regression coefficient for the interaction is small compared to its standard error, regardless of the manner in which the standard error is estimated.

Table 8 presents the results of logistic regressions of the probability of mammogram in the last year on race and income. For simple and multiple logistic regressions, it appears that the without-replacement variance estimator underestimates the variability of

TABLE 7

Weighted regression coefficients and standard errors for the multiple linear regression of income on race, sex and race × sex using data from the 1987 *National Health Interview Survey* (*sample size* = 19.561)

Independent variable	**Race (white vs. nonwhite)**	**Sex (men vs. women)**	**Race × sex**
Weighted regression coefficient	7099	3067	149
Standard error estimator			
Original data			
$[\widehat{\mathrm{var}}_{\mathrm{WO}}(\hat{\beta})]^{1/2}$	654	308	810
$[\widehat{\mathrm{var}}_{\mathrm{SP}}(\hat{\beta})]^{1/2}$	755	283	665
Certainty PSUs paired			
$[\widehat{\mathrm{var}}_{\mathrm{WR}}(\hat{\beta})]^{1/2}$	738	271	700
$[\widehat{\mathrm{var}}_{\mathrm{SP}-a}(\hat{\beta})]^{1/2}$	758	288	680
Certainty PSUs divided			
$[\widehat{\mathrm{var}}_{\mathrm{WR}}(\hat{\beta})]^{1/2}$	662	314	830

TABLE 8

Weighted logistic regression coefficients and standard errors for two simple logistic regressions and one multiple logistic regression of mammogram (*probability within the last year*) *on race and income using data from the* 1987 *National Health Interview Survey*

	Simple logistic regressions		**Multiple logistic regression**	
Independent variable	**Race (white vs. nonwhite)**	**Income ($\times 10^{-6}$)**	**Race (white vs. nonwhite)**	**Income ($\times 10^{-6}$)**
Sample size	6522	5620	5620	
Weighted regression coefficient	−0.290	24.4	−0.119	24.2
Standard error estimator				
Original data				
$[\widehat{\mathrm{var}}_{\mathrm{WO}}(\hat{\beta})]^{1/2}$	0.127	2.48	0.136	2.48
$[\widehat{\mathrm{var}}_{\mathrm{SP}}(\hat{\beta})]^{1/2}$	0.178	2.57	0.192	2.53
Certainty PSUs paired				
$[\widehat{\mathrm{var}}_{\mathrm{WR}}(\hat{\beta})]^{1/2}$	0.141	2.82	0.143	2.78
$[\widehat{\mathrm{var}}_{\mathrm{SP}-a}(\hat{\beta})]^{1/2}$	0.178	2.65	0.191	2.62
Certainty PSUs divided				
$[\widehat{\mathrm{var}}_{\mathrm{WR}}(\hat{\beta})]^{1/2}$	0.126	2.60	0.135	2.60

the race coefficient. This is also seen in the multiple logistic regression that contains the race-by-income interaction (Table 9).

To test or form a confidence interval for any single regression coefficient in Tables 5–9, one uses a t distribution with 124 degrees of freedom with one of the superpopulation variance estimators. For example, in the simple logistic regression in Table 8, the association of race with the probability of mammogram is tested by comparing -1.629 $(= -0.290/0.178)$ to a t distribution with 124 degrees of freedom. Note that it does not make sense to test this association in the finite population. To simultaneously test two regression coefficients in a multiple regression, for example, race and the race-by-income interaction both being 0 in Table 9, one uses an F test with 2 and 123 degrees of freedom

TABLE 9

Weighted regression coefficients and standard errors for the multiple logistic regression of mammogram (probability within the last year) on race, income and race-by-income using data from the 1987 *National Health Interview Survey* (*sample size* = 5620)

Independent variable	**Race (white vs. nonwhite)**	**Income** ($\times 10^{-6}$)	**Race-by-Income** ($\times 10^{-6}$)
Weighted regression coefficient	0.309	26.0	−17.2
Standard error estimator			
Original data			
$[\widehat{\mathrm{var}}_{\mathrm{WO}}(\hat{\beta})]^{1/2}$	0.210	2.66	7.72
$[\widehat{\mathrm{var}}_{\mathrm{SP}}(\hat{\beta})]^{1/2}$	0.279	2.53	8.72
Certainty PSUs paired			
$[\widehat{\mathrm{var}}_{\mathrm{WR}}(\hat{\beta})]^{1/2}$	0.244	2.90	8.12
$[\widehat{\mathrm{var}}_{\mathrm{SP}-a}(\hat{\beta})]^{1/2}$	0.282	2.65	8.84
Certainty PSUs divided			
$[\widehat{\mathrm{var}}_{\mathrm{WR}}(\hat{\beta})]^{1/2}$	0.212	2.78	7.92

with the 2×2 superpopulation covariance matrix of the estimated coefficients (not shown).

6. DISCUSSION

The merits of model-based versus design-based (randomization-based) inference for survey data have been thoroughly discussed in the literature, for example, Hansen, Madow and Tepping (1983). While not intending to enter into the controversy, we realize the suggestions in this paper may have some relevance to these more general issues. In particular, we have suggested for inference that standard (design-based) weighted estimators should be used but with variance estimators that incorporate both randomization and superpopulation model variability. We briefly examine the implications of this approach for some of the more general issues: the choice of target parameter, the choice of estimator (especially the use of sample-weighted estimators), the relevance of the randomization distribution induced by the sampling and the choice of conditioning statistics.

For the scientific applications on which we are focusing, the target parameters are associated with stochastic models for the finite-population values rather than with functions of the finite-population values themselves. The stochastic models we utilize are relatively simple marginal models that involve a few dependent and independent variables, and can be implemented with easy modifications of existing software. They are not models that incorporate all the variables that describe the sampling or the population, that is, all the variables involved in the sampling design, poststratification and nonresponse adjustments, as well as variables defining the levels of clustering in the population. We believe that the parameters associated with these simple marginal models tend to be the ones of most scientific interest. One could, in theory, develop a comprehensive hierarchical model that incorporates all the known sampling and population variables and then define a marginal parameter based on that model. This type of complex superpopulation modeling would likely require extensive resources for each analysis. In addition, this type of approach would not appear to offer much statistical benefit (Pfeffermann and LaVange, 1989).

Since we are using a model for the population to define the target parameter, why incorporate the weights derived from the sampling into the parameter estimator? We use sample-weighted estimators because they estimate the target parameter with a minimum of assumptions. For example, to estimate the mean of Y from a probability-proportional-to-size sample, one can use the weighted mean, whereas the unweighted mean will not estimate the superpopulation mean without a further assumption such as the size variable being independent of Y. We do not rigidly recommend weighted estimators; there are situations in which the weights are so variable that some modeling is advisable (Korn and Graubard, 1995; Magee, 1998). In general, however, we recommend weighted estimators because we believe their model-free aspects outweigh their potential inefficiency.

The third issue we address concerns the relevance of the sampling distribution to superpopulation inference. In particular, why are our superpopulation variance estimators based on corrections to repeated-sampling variance estimators? We used this approach mainly for convenience; repeated-sampling variance estimators for weighted parameter estimators are readily available. However, note that in some cases less information is needed for the superpopulation variance estimator than for the repeated-sampling variance estimator (Section 3.2). Of course, if one is willing to make enough model assumptions, then the sampling distribution will become irrelevant in the sense that it can be ignored when doing a model-based analysis. One might presume that these additional model assumptions would lead to variance estimators with reduced variability (when the assumptions hold); this has not been our experience (Graubard and Korn, 1996a).

Finally, we consider the choice of conditioning statistics. Even with finite-population inference, there is some question as to which, if any, auxiliary statistics should be conditioned on for inference (Royall and Cumberland, 1981; Rao, 1985). For example, when estimating a mean with a ratio estimator using stratified simple random sampling, should one calculate variances conditional on the sample strata means of the auxiliary variable, $\{\bar{x}_1, \dots, \bar{x}_L\}$? For superpopulation inference (Section 2.5), we have suggested not only not conditioning on $\{\bar{x}_1, \dots, \bar{x}_L\}$, but also not conditioning on the finite-population means $\{\overline{X}_1, \dots, \overline{X}_L\}$! We are unapologetic about this, as we believe that this is reasonable unless one is willing to rely on a stronger model for the (U_i, X_i, η_i) than is given in Section 2.3. Chambers (1986, page 170) suggests that an argument for a completely unconditional analysis is "attractive provided some sort of joint distribution for Y [our U], X and Z [our η] can be postulated." He further states "Unfortunately, this is rarely possible." We note that, for the estimators presented in this paper, it is unnecessary to specify this joint distribution.

The conditioning issue also arises in the context of poststratification; should one condition on the strata sample sizes when estimating the variance of a poststratified mean (Holt and Smith, 1979)? For estimating the variance of the poststratified mean around the finite-population mean, the practical importance of this question would seem to be limited to whether or not to include the second term of (2.12), which is of small order. [Conditioning on strata-specific *subdomain* sample sizes can yield very wrong inferences for subdomain means (Graubard and Korn, 1996b).] For superpopulation inference, conditioning on strata sample sizes would be a bad idea; with large sampling fractions this conditioning approaches conditioning on the population strata sizes which we have shown to underestimate the superpopulation variability.

In summary, we believe design-based inference to be an efficient and reasonably model-free approach for inference about finite-population parameters. The simple modifications of design-based variance estimators suggested in this paper allow one to make inferences with few model assumptions for superpopulation parameters, which are frequently the ones of primary scientific interest.

ACKNOWLEDGMENTS

We thank Iris Shimizu, Meena Khare and Jill Montaquila for providing information about the sample designs for the surveys analyzed in this paper, and two referees for their helpful comments.

REFERENCES

ARNAB, R. (1992). Estimation of a finite population mean under superpopulation models. *Comm. Statist. Theory Methods* **21** 1717–1724.

BELLHOUSE, D. R., THOMPSON, M. E. and GODAMBE, V. P. (1977). Two-stage sampling with exchangeable prior distributions. *Biometrika* **64** 97–103.

BOUZA, C. N. (1995). Linear rank tests derived from a superpopulation model. *Biometrical J.* **37** 497–506.

BRECKLING, J. U., CHAMBERS, R. L., DORFMAN, A. H., TAM, S. M. and WELSH, A. H. (1994). Maximum likelihood inference from sample survey data. *Internat. Statist. Rev.* **62** 349–363.

BREEN, N. and KESSLER, L. (1994). Changes in the use of screening mammography: evidence from the 1987 and 1990 National Health Interview Surveys. *Amer. J. Pub. Health* **84** 62–7.

CAMPBELL, C. (1977). Properties of ordinary and weighted least square estimators of regression coefficients for two-stage samples. In *Proceedings of the Section on Social Statistics* 800–805. Amer. Statist. Assoc., Alexandria, VA.

CASSEL, C., SÄRNDAL, C. and WRETMAN, H. H. (1977). *Foundations of Inference in Survey Sampling*. Wiley, New York.

CHAMBERS, R. L. (1986). Design-adjusted parameter estimation. *J. Roy. Statist. Soc. Ser. A* **149** 161–173.

CHAMBERS, R. L., DORFMAN, A. H. and WANG, S. (1998). Limited information likelihood analysis of survey data. *J. Roy. Statist. Soc. Ser. B* **60** 397–411.

CHRISTENSEN, R. (1984). A note on ordinary least squares methods for two-stage sampling. *J. Amer. Statist. Assoc.* **79** 720–721.

CHRISTENSEN, R. (1987). The analysis of two-stage sampling data by ordinary least squares. *J. Amer. Statist. Assoc.* **82** 492–498.

COCHRAN, W. G. (1939). The use of the analysis of variance in enumeration by sampling. *J. Amer. Statist. Assoc.* **34** 492–510.

COCHRAN, W. G. (1946). Relative accuracy of systematic and stratified random samples for a certain class of populations. *Ann. Math. Statist.* **17** 164–177.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

COSSLETT, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica* **49** 1289–1316.

DEMETS, D. and HALPERIN, M. (1977). Estimation of a simple regression coefficient in samples arising from a sub-sampling procedure. *Biometrics* **33** 47–56.

DEMING, W. E. (1953). On the distinction between enumerative and analytic surveys. *J. Amer. Statist. Assoc.* **48** 244–255.

DEMING, W. E. and STEPHAN, F. F. (1941). On the interpretation of censuses as samples. *J. Amer. Statist. Assoc.* **36** 45–49.

DUMOUCHEL, W. H. and DUNCAN, G. J. (1983). Using sample survey weights in multiple regression analyses of stratified samples. *J. Amer. Statist. Assoc.* **78** 535–543.

DURBIN, J. (1953). Some results in sampling theory when the units are selected with unequal probabilities. *J. Roy. Statist. Soc. Ser. B* **15** 262–269.

ELTINGE, J. L. and JANG, D. S. (1996). Stability measures of variance component estimators under a stratified multistage design. *Survey Methodology* **22** 157–165.

ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations (with discussion). *J. Roy. Statist. Soc. Ser. B* **31** 195–233.

EZZATI, T. M., MASSEY, J. T., WAKSBERG, J., CHU, A. and MAURER, K. R. (1992). Sample design: third National Health and Nutrition Examination Survey. *Vital Health Statist.* **2**.

FULLER, W. A. (1975). Regression analysis for sample survey. *Sankhyā Ser. C* **37** 117–132.

GODAMBE, V. P. and THOMPSON, M. E. (1986). Parameters of superpopulation and survey population: their relationships and estimation. *Internat. Statist. Rev.* **54** 127–138.

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73** 43–56.

GRAUBARD, B. I. and KORN, E. L. (1996a). Modelling the sampling design in the analysis of health surveys. *Statist. Methods Medical Res.* **5** 263–281.

GRAUBARD, B. I. and KORN, E. L. (1996b). Survey inference for subpopulations. *Amer. J. Epidemiol.* **144** 102–106.

GRAVES, E. J. (1988). Utilization of short-stay hospitals, United States, 1986, annual summary. *Vital Health Statist.* **13**.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory* **1**. Wiley, New York.

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). *J. Amer. Statist. Assoc.* **78** 776–807.

HARTLEY, H. O. and SIELKEN, R. L., Jr. (1975). A "superpopulation viewpoint" for finite population sampling. *Biometrics* **31** 411–422.

HASLETT, S. (1985). The linear non-homogeneous estimator in sample surveys. *Sankhyā Ser. B* **47** 101–117.

HAUSMAN, J. A. and WISE, D. A. (1981). Stratification of endogenous variables and estimation: the Gary income maintenance experiment. In *Structural Analysis of Discrete Data with Econometric Applications* (C. F. Manski and S. McFadden, eds.) 365–391. MIT Press, Cambridge, MA.

HOLT, D. and SCOTT, A. J. (1981). Regression analysis using survey data. *The Statistician* **30** 169–178.

HOLT, D. and SMITH, T. M. F. (1979). Post stratification. *J. Roy. Statist. Soc. Ser. A* **142** 33–46.

HOLT, D., SMITH, T. M. F. and WINTER, P. D. (1980). Regression analysis of data from complex surveys. *J. Roy. Statist. Soc. Ser. A* **143** 474–487.

ISAKI, C. T. and FULLER, W. A. (1982). Survey design under the regression superpopulation model. *J. Amer. Statist. Assoc.* **77** 89–96.

JEWELL, N. P. (1985). Least squares regression with data arising from stratified samples of the dependent variable. *Biometrika* **72** 11–21.

KLEIN, L. R. and MORGAN, J. N. (1951). Results of alternative statistical treatments of sample survey data. *J. Amer. Statist. Assoc.* **46** 442–460.

KONIJN, H. S. (1962). Regression analysis in sample surveys. *J. Amer. Statist. Assoc.* **57** 590–606.

KOOP, J. C. (1986). Some problems of statistical inference from sample survey data for analytic studies. *Statistics* **17** 237–247. [Correction (1992) *Statistics* **23** 187.]

KORN, E. L. and GRAUBARD, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *Amer. Statist.* **44** 270–276.

KORN, E. L. and GRAUBARD, B. I. (1995). Analysis of large health surveys: accounting for the sampling design. *J. Roy. Statist. Soc. Ser. A* **158** 263–295.

KORN, E. L. and GRAUBARD, B. I. (1998). Variance estimation for superpopulation parameters. *Statist. Sinica* **8** 1131–1151.

KORN, E. L. and GRAUBARD, B. I. (1999). *Analysis of Health Surveys*. Wiley, New York.

KOTT, P. S. (1991). A model-based look at linear regression with survey data. *Amer. Statist.* **45** 107–112.

KOTT, P. S. (1993). Comment on Potthoff, Woodbury, and Manton. Letter to the Editor. *J. Amer. Statist. Assoc.* **88** 716.

KRIEGER, A. M. and PFEFFERMANN, D. (1992). Maximum likelihood estimation from complex sample surveys. *Survey Methodology* **18** 225–239.

LEHMANN, E. L. (1975). *Nonparametrics*. Holden-Day, San Francisco.

LONGFORD, N. T. (1996). Model-based variance estimation in surveys with stratified clustered design. *Austral. J. Statist.* **38** 333–352.

MAGEE, L. (1998). Improving survey-weighted least squares regression. *J. Roy. Statist. Soc. Ser. B* **60** 115–126.

MASSEY, J. T., MOORE, T. F., PARSONS, V. L. and TADROS, W. (1989). Design and estimation for the National Health Interview Survey, 1985–1994. *Vital Health Statist.* **2**.

NATHAN, G. and HOLT, D. (1980). The effect of survey design on regression analysis. *J. Roy. Statist. Soc. Ser. B* **42** 377–386.

NATIONAL CENTER FOR HEALTH STATISTICS. (1994). Plan and operation of the Third National Health and Nutrition Examination Survey, 1988–1994. *Vital Health Statist.* **1**.

NORDBERG, L. (1989). Generalized linear modeling of sample survey data. *J. Official Statist.* **5** 223–239.

PATIL, G. P. and RAO, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics* **34** 179–189.

PFEFFERMANN, D. and HOLMES, D. J. (1985). Robustness considerations in the choice of a method of inference for regression analysis of survey data. *J. Roy. Statist. Soc. Ser. A* **148** 268–278. [Correction (1985) *J. Roy. Statist. Soc. Ser. A* **148** 357.]

PFEFFERMANN, D. and LAVANGE, L. (1989). Regression models for stratified multi-stage cluster samples. In *Analysis of Complex Surveys* (C. J. Skinner, D. Holt and T. M. F. Smith, eds.) 237–260. Wiley, New York.

PFEFFERMANN, D. and NATHAN, G. (1981) Regression analysis of data from a cluster sample. *J. Amer. Statist. Assoc.* **76** 681–689.

PFEFFERMANN, D., SKINNER, C. J., HOLMES, D. J., GOLDSTEIN, H. and RASBASH, J. (1998). Weighting for unequal selection probabilities in multilevel models (with discussion). *J. Roy. Statist. Soc. Ser. B* **60** 23–56.

PORTER, R. D. (1973). On the use of survey sample weights in the linear model. *Ann. Econom. Social Measurement* **2** 141–158.

POTTHOFF, R. F., WOODBURY, M. A. and MANTON, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *J. Amer. Statist. Assoc.* **87** 383–396.

QUESENBERRY, C. P., Jr. and JEWELL, N. P. (1986). Regression analysis based on stratified samples. *Biometrika* **73** 605–614.

RAO, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In *Classical and Contagious Discrete Distributions* (G. P. Patil, ed.) 320–332. Statistical Publishing Society, Calcutta.

RAO, J. N. K. (1985). Conditional inference in survey sampling. *Survey Methodology* **11** 15–31.

ROYALL, R. M. and CUMBERLAND, W. G. (1981). An empirical study of the ratio estimator and its variance (with discussion). *J. Amer. Statist. Assoc.* **76** 66–88.

SÄRNDAL, C. E. (1980). Two model-based inference arguments in survey sampling. *Austral. J. Statist.* **22** 341–348.

SÄRNDAL, C. E., SWENSSON, B. and WRETMAN, J. (1992). *Model Assisted Survey Sampling*. Springer, New York.

SCHOENBORN, C. A. and MARANO, M. (1988). Current estimates from the National Health Interview Survey, United States, 1987. *Vital Health Statist.* **10**.

SCOTT, A. J. and HOLT, D. (1982). The effect of two-stage sampling on ordinary least squares methods. *J. Amer. Statist. Assoc.* **77** 848–854.

SCOTT, A. J. and SMITH, T. M. F. (1969). Estimation in multi-stage surveys. *J. Amer. Statist. Assoc.* **64** 830–840.

SCOTT, A. J. and WILD, C. J. (1986). Fitting logistic models under case–control or choice based sampling. *J. Roy. Statist. Soc. Ser. B* **48** 170–182.

SCOTT, A. J. and WILD, C. J. (1989). Selection based on the response variable in logistic regression. In *Analysis of Complex Surveys* (C. J. Skinner, D. Holt and T. M. F. Smith, eds.) 191–205. Wiley, New York.

SCOTT, A. J. and WILD, C. J. (1991). Fitting logistic regression models in stratified case–control studies. *Biometrics* **47** 497–510.

SEDRANSK, J. (1965). Analytical surveys with cluster sampling. *J. Roy. Statist. Soc. Ser. B* **27** 264–278.

SHAH, B. V., BARNWELL, B. G. and BIELER, G. S. (1997). *SUDAAN User's Manual, Release 7.5*. Research Triangle Institute, Research Triangle Park, NC.

SIMMONS, W. R. and SCHNACK, G. A. (1970). Development of the design of the NCHS Hospital Discharge Survey. *Vital Health Statist.* **2**.

SKINNER, C. J. (1994). Sample models and weights. In *Proceedings of the Section on Survey Research Methods* 133–142. Amer. Statist. Assoc., Alexandria, VA.

SKINNER, C. J., HOLT, D. and SMITH, T. M. F., eds. (1989). *Analysis of Complex Surveys*. Wiley, New York.

TEN CATE, A. (1986). Regression analysis using survey data with endogenous design. *Survey Methodology* **12** 121–138.

THOMSEN, I. (1978). Design and estimation problems when estimating a regression coefficient from survey data. *Metrika* **25** 27–35.

YATES, F. (1981). *Sampling Methods for Censuses and Surveys*, 4th ed. Oxford Univ. Press.