Journal of Official Statistics
Vol. 7. No. 4, 1991, pp. 405–424
© Statistics Sweden

Inference with Survey Weights

Roderick J.A. Little[1]

Abstract: This article considers the analysis of disproportionate stratified samples from a model-based (Bayesian) perspective. It is argued that a key element of models for such samples is that they explicitly account for differences between strata, even when the target quantity is aggregated over strata. Two general classes of models with this property are proposed. The first class, which I call *fixed stratum-effects models*, yields as special cases standard probability-weighted inferences favored by survey statisticians. The second class, which I call *random stratum-effects models*, yields estimators that behave like fixed stratum-effects estimators when the stratum sample sizes are large. In moderate samples they are compromises between estimators from fixed stratum-effects models and estimators from models that ignore stratum effects. In simple settings these are weighted estimators where the weights have been smoothed towards one, yielding in certain cases a reduction in mean-squared error.

For inference about a finite population mean, a fixed stratum-effects model leads to posterior probability intervals identical to standard randomization inference based on the stratified mean; random stratum-effects models yield estimators with smoothed weights. Repeated sampling properties of these estimators and associated probability intervals are illustrated by a simulation study on normal and non-normal populations.

For inference about a population slope, it is shown that classical design-based inference using the sample weights approximates Bayesian inference under a fixed stratum-effects model. Thus the need to model stratum effects leads to the probability-weighted methods usually associated with design-based inference.

Key words: Bayesian methods; design consistency; James-Stein shrinkage; random effects; regression; stratified sampling; superpopulation models.

1. Introduction and Framework

The role of sampling weights in statistical analysis of survey data is the subject of controversy amongst theorists and confusion amongst practitioners. For descriptive inference about means and totals, *probability or π-weighted estimates*, where cases are weighted by the inverse of the probability of selection and response, are widely accepted. For more complex modeling exercises, there is a wide spectrum of opinions on the role of weights, from modelers who view weights as largely irrelevant to survey statisticians who incorporate weights, along with other features of the sample design, routinely into every analysis (Klein and Morgan 1951; Konijn 1962; Brewer and Mellor 1973; Kish

[1] Department of Biomathematics, U.C.L.A. School of Medicine, Los Angeles, CA 90024, U.S.A.

Acknowledgements: This research was supported by grant USPH MH 37188 from the National Institute of Mental Health. I wish to thank Wil Dixon, Danny Pfeffermann, Robert Weiss, and two referees for helpful comments on an earlier draft, and Don Rubin for his many stimulating insights.

and Frankel 1974; Särndal 1978, 1980; Holt, Smith and Winter 1980; DuMouchel and Duncan 1983; Hansen, Madow and Tepping 1983; Little 1983a,b; Rubin 1983a, Pfeffermann and Holmes 1985; Chambers 1986; Ghosh and Lahiri 1987; Pfeffermann and Lavange 1989; and Skinner, Holt and Smith 1989).

My own view is that (a) focusing on finite population quantities is a useful discipline, even for analytic inferences; (b) inference for finite population quantities should in principle be based on suitable models; (c) models need to be robust, in the sense that inferences based on them are insensitive to misspecification errors rendered important by the sample design; in the context of disproportionate stratified sampling, robust models need to reflect stratum differences, even if these differences are not detectable from diagnostic tests applied to the sample at hand; (d) simple models that reflect stratum differences often lead to π-weighted inferences similar to those derived from randomization theory, thus providing a model-based justification of at least some design-based m.thods; (e) the modeling approach provides principled modifications of π-weighted inference that can yield better inferences in small or moderate samples.

This viewpoint is developed here in the context of stratified samples, where the population is grouped into J strata defined by values of a variable Z, and units are sampled with probability π_j in stratum j, where π_j varies across the strata. To avoid additional complications such as clustering of the sample, I assume that a simple random sample of units is selected in each stratum, so that $\pi_j = n_j/N_j$ where N_j is the number of population units in stratum j and n_j is the number that are sampled; also let $\mathcal{S}_j$ denote the set of units sampled in stratum j. I focus on situations where the π_j vary across the strata, but the selection probabilities are constant within strata; important examples include disproportionate stratified sampling with sampling probabilities π_j, and post-stratification for surveys with nonresponse, where respondents are weighted up to known post-strata totals. In the latter case n_j is the number of respondents in post-stratum j and N_j represents a known total from census data.

In practice, surveys often do not involve simple random sampling within strata. However, it is still useful to think of Z as defining the strata of the population over which the sampling or nonresponse fractions vary. For example, in the U.S. Panel Study of Income Dynamics, Z involves region and Standard Metropolitan Statistical Area, and variables used to form nonresponse adjustments. In the U.S. National Health and Examination Survey, Z includes a poverty/non-poverty stratum since poverty strata were oversampled; age and household size also enters the definition of Z for some estimands, since one individual was sampled from each household, with young and old age groups being oversampled relative to intermediate groups. In the U.S. Statistics of Income Survey of tax returns, Z is a complex measure of size of return, with large returns being sampled at much higher rates than small returns.

Suppose K variables $X_1, \ldots, X_K$ are measured in the survey, and let $\mathbf{X}$ denote the $N \times K$ matrix of values of these variables in the population. I consider inference about a finite population quantity $Q = Q(\mathbf{X})$ based on the sample. For example, Q could be the mean of a particular variable, a regression coefficient in a multiple regression, or a factor score in some complex factor-analytic model.

For analytic rather than descriptive inference, the parameters θ of a superpopulation model, which I shall call the *target model*, may be of interest. In such cases I choose to regard the target quantity as not θ itself, but rather the population quantity $Q(\mathbf{X}) = \theta_{\text{pop}}(\mathbf{X})$ that would be obtained by fitting the target model to the entire population, using some specified fitting procedure such as least squares. Statisticians who build models for the data tend to focus on θ, whereas survey statisticians who base inference on the sampling distribution treating **X** fixed tend to focus on θ_{pop} (Brewer and Mellor 1973; Hansen, Madow and Tepping 1983; and DuMouchel and Duncan 1983). Although a modeler by philosophy, I like the survey sampler's focus on θ_{pop} since it is a real entity that exists irrespective of the validity of the model. The parameter θ exists only within the context of the target model, and given model misspecification has no clear definition. It is true that the target model usually must have some face validity for θ_{pop} to be a reasonable estimand; for example, estimation of the population least squares slope of X_2 on X_1 is questionable unless the regression of X_2 on X_1 is at least approximately linear. Nevertheless models are simplified descriptions that ignore fine structure, particularly in large populations. Focusing on θ_{pop} keeps the target well-defined in the presence of model misspecification. For those unimpressed by this argument who still prefer to focus on θ, I suggest that if trouble is taken to design a probability sample of the population, then any good estimator of θ must also be a good estimator of θ_{pop}, so focusing on θ_{pop} should not lead us seriously astray.

Following the Bayesian formulation of finite population inference (Ericson 1969), I base inference about $Q(\mathbf{X})$ given the sampled data $\mathbf{X}_{\text{obs}}$ on its posterior predictive distribution $p(Q|\mathbf{X}_{\text{obs}})$ under a *working model* for **X**, characterized by a prior distribution $p(\mathbf{X})$ for the population values. The working models I consider have the general form

$$p(\mathbf{X}) = \prod_{j=1}^{J} p(\mathbf{X}^{(j)});$$

$$p(\mathbf{X}^{(j)}) = \prod_{i=1}^{N_j} \int p(\mathbf{x}_{ji}|\lambda_j, \phi_j)p(\lambda_j, \phi_j)d\lambda_j d\phi_j, \qquad (1)$$

where $\mathbf{X}^{(j)}$ is the $(N_j \times K)$ matrix of population values in stratum j; $\mathbf{x}_{ji}$ is the $(1 \times K)$ vector of population values for unit i in stratum j ; λ_j is a set of location parameters indexing the distribution of $\mathbf{x}_{ji}$, ϕ_j is a set of dispersion or shape parameters, and (λ_j, ϕ_j) has prior distribution $p(\lambda_j, \phi_j)$.

A crucial feature of this model is the fact that *distinct parameters* (λ_j,ϕ_j) *are specified for each stratum j*. Borrowing ANOVA terminology, I call the location parameters $\{\lambda_j\}$ *stratum effects*, and define *fixed stratum-effects models* as models with noninformative priors on λ_j:

$$p(\lambda_1,\ldots,\lambda_k) \propto \text{const.} \qquad (2)$$

Alternatively I consider *random stratum effects models* where the prior for λ_j has the form

$$p(\lambda_1,\ldots,\lambda_K) = \int \prod_{j=1}^{J} p(\lambda_j|\lambda, \delta)d\lambda d\delta \qquad (3)$$

where λ and δ are respectively location and scale/shape parameters, which are assigned uniform priors. Previous applications to surveys of random effects models of this type include Scott and Smith (1969) in the context of multistage cluster sampling, and Battese and Fuller (1981) in the context of small area estimation.

Notes:

1. In large samples, inferences are insensitive to the form of the prior, and this

Bayesian formulation is practically indistinguishable from non-Bayesian superpopulation models that avoid priors for λ_j and ϕ_j and treat these parameters as fixed; for arguments in favor of the Bayesian formulation see for example Little and Rubin (1983).

2. The simple random sampling design within strata motivates a model that treats the vectors $\mathbf{x}_{ji} (i = 1, \ldots, N_j)$ as exchangeable within strata. By De Finetti's theorem, this justifies an iid model for $\mathbf{x}_{ji}$ conditional on stratum parameters. (Ericson 1969; Rubin 1987, Section 2.5.)

3. The inclusion of distinct parameters λ_j, ϕ_j for each stratum j is important to overcome distortions in the sample introduced by the differential selection probabilities (Little 1983a; Rubin 1983a). In particular, models need to be constructed that yield *design-consistent* estimators, where design consistency means that as the sample sizes increase the estimates of θ_{pop} converge to θ_{pop} even when the model is misspecified (Brewer 1979; and Robinson and Tsui 1979). Working models that distinguish stratum parameters are more likely to be design-consistent than models that do not, as can be seen from the examples in Little (1983b) and in this article.

4. *The target quantity exists quite independently of the working model.* In particular, the working model needs to reflect differences between strata, but the target quantity may be aggregated over strata. For example, θ_{pop} might be the slope of the regression of X_2 on X_1 in the whole population, pooled across strata since the conceptual model does not treat Z as an exogenous variable.

5. In small or moderate sized samples, the form of the prior for λ_j and ϕ_j becomes more important. Priors should in principle be tailored to each specific problem; we consider the class (3) of random stratum-effects since they provide useful compromises between estimates from models that recognize stratum effects and estimates from models that ignore them. They lead to James-Stein-type estimators of location parameters (for example Efron and Morris 1973), and were previously considered for estimating survey means in Little (1983b), Ghosh and Meeden (1986), and Ghosh and Lahiri (1987). Ghosh and Lahiri (1987) proved asymptotic optimality properties for empirical Bayes estimators of stratum means, and showed reductions in risk over stratum means by theory and simulation.

Sections 2 and 3 concern the application of this modeling approach to inference about a finite population mean or total. Section 4 considers inference about regression slopes.

2. Normal Models for Means and Totals

Two kinds of weights arise in the analysis of disproportionate stratified samples: probability weights determined by the probabilities of selection, and variance weights determined by within-stratum variation of the outcome variable. We first consider the role of these weights for the basic problem of inference about the population mean $\bar{X}$ of a scalar variable X. Then $\bar{X} = \Sigma_j P_j \bar{X}_j$, where $P_j = N_j/N$ and $\bar{X}_j$ are respectively the population proportion and mean of X in stratum j. Weighting sampled units by the inverse of the selection probability π_j in stratum j yields the π-weighted (or stratified) mean

$$\bar{x}_\pi = \frac{1}{N}(\Sigma_j \Sigma_{i \in \mathcal{S}_j} x_{ij}/\pi_j) = \Sigma_j P_j \bar{x}_j \quad (4)$$

where $\mathcal{S}_j$ denotes the sample in stratum j (Horvitz and Thompson 1952). Weighting sampled units by the inverse of the sample variance s_j^2 in stratum j yields the variance-weighted mean:

$$\bar{x}_v = \Sigma_j v_j \bar{x}_j; \quad v_j = \frac{n_j/s_j^2}{\Sigma_j n_j/s_j^2}. \quad (5)$$

The π-weighted estimator aims at controlling bias, the variance-weighted estimator aims at controlling variance. Thus $\bar{x}_\pi$ is unbiased for $\bar{X}$, but it can have excessive variance if the variance of x is high in strata with low selection probabilities, as when an extreme value of x has low probability of selection; $\bar{x}_v$ is the weighted average of the stratum means with lowest variance (ignoring errors in estimating the variances), but it can be seriously biased if the variance-weights differ markedly from the design weights and stratum means are far apart.

Since π-weighting relates to the sample design and variance-weighting relates to the distribution of x in the population, it is natural to view $\bar{x}_\pi$ as a design-based estimator and $\bar{x}_v$ as a model-based estimator. However I prefer to view both of these estimators as arising from models for the population. An abstract philosophical argument between "design-based" and "model-based" inference is thereby replaced by a concrete pragmatic argument concerning the appropriate choice of model.

Table 1 displays the approximate posterior mean and variance of $\bar{X}$ under five models for the distribution of $\{x_{ji}\}$; the approximation arises from ignoring variance components due to estimating the population variances in the models. Posterior probability intervals for $\bar{X}$ take the form

$$m \pm z\sqrt{v}, \tag{6}$$

where m is the posterior mean, v is the posterior variance and z is the appropriate normal percentile (e.g., 1.96 for 95% intervals). The probability-weighted (PWT) model specifies fixed stratum effects:

$$x_{ji} | \lambda_j, \phi_j \sim {}_{\text{ind}}G(\lambda_j, \phi_j^2),$$
$$p(\lambda_j, \log \phi_j) \propto \text{const.} \tag{7}$$

where x_{ji} is the value of x for unit i in stratum j, $G(a, b)$ denotes the normal distribution with mean a, variance b. The posterior mean from the PWT model is the π-weighted estimator (4). Moreover posterior intervals (6) based on model (7) are identical to the randomization-based confidence intervals from classical stratified sampling theory (Ericson 1969).

VWT (variance-weighted) and UWT (unweighted) are null stratum-effects models which assume $\lambda_j = \lambda$ for all j. VWT allows distinct variances in each stratum

$$x_{ji} | \lambda_j, \phi_j \sim {}_{\text{ind}}G(\lambda_j, \phi_j^2),$$
$$p(\lambda, \log \phi_j) \propto \text{const.} \tag{8}$$

and UWT assumes a constant within-stratum variance

$$x_{ji} | \lambda, \phi_j \sim {}_{\text{ind}}G(\lambda, \phi^2),$$
$$p(\lambda, \log \phi) \propto \text{const.} \tag{9}$$

When the sampling fraction f is small, the posterior mean under (8) reduces to the variance-weighted estimator (5), and the posterior mean under (9) reduces to the unweighted sample mean $\bar{x}$. These models violate the notion of allowing different location parameters across strata. Although they yield better inferences than PWT when correctly specified, the assumption of equal stratum means is usually unrealistic, particularly since efficiently-designed samples are heterogeneous between strata. We show in Section 3 that inferences from these models are very vulnerable to model misspecification.

EBV, which denotes empirical Bayes shrinkage towards the variance-weighted mean, is obtained by replacing the uniform prior for λ_j in (7) by a proper prior:

$$x_{ji} | \lambda_j, \phi_j \sim {}_{\text{ind}}G(\lambda_j, \phi_j^2),$$
$$(\lambda_j | \lambda, \delta^2) \sim {}_{\text{iid}}G(\lambda, \delta^2);$$
$$p(\lambda, \log \phi_j^2, \log \delta^2) \propto \text{const.} \tag{10}$$

where the stratum means λ_j are assumed to be an iid sample from an underlying distribution (e.g., Ericson 1965). The estimate $\hat{\delta}^2$ of δ^2 in this model is obtained using the

Table 1. Approximate Bayesian inference for $\bar{X}$ under five models

Model (Eq.)	Posterior mean	Posterior variance
PWT (7)	$\bar{x}_\pi = \Sigma_j P_j \bar{x}_j$	$\Sigma_j P_j^2(1 - f_j)s_j^2/n_j)$
VWT (8)	$f\bar{x} + (1 - f)\bar{x}_v$	$\Sigma_j P_j(1 - f_j)s_j^2/N$ $+ (1 - f)^2\{\Sigma_j n_j/s_j^2\}^{-1}$
UWT (9)	$\bar{x}$	$(1 - f)s^2/n$
EBV (10)	$f\bar{x} + \Sigma_j P_j(1 - f_j)$ $\times \{w_j\bar{x}_j + (1 - w_j)\bar{x}_w\}$	$\Sigma_j P_j^2(1 - f_j)\{f_j + (1 - f_j)w_j\}s_j^2/n_j$ $+ \{\Sigma_j P_j(1 - f_j)(1 - w_j)\}^2\hat{\delta}^2/\Sigma_j w_j$
EBU (11)	$f\bar{x} + \Sigma_j P_j(1 - f_j)$ $\times \{u_j\bar{x}_j + (1 - u_j)\bar{x}_u\}$	$\Sigma_j P_j^2(1 - f_j)\{f_j + (1 - f_j)u_j\}s_j^2/n_j$ $+ \{\Sigma_j P_j(1 - f_j)(1 - u_j)\}^2\tilde{\delta}^2/\Sigma_j u_j$

Notation:
$P_j = N_j/N; f_j = n_j/N_j; f = n/N;$
$\bar{x}$, s^2 = sample mean and variance; $\bar{x}_j$, s_j^2 = sample mean and variance, stratum j;
$\bar{x}_\pi = \Sigma_j P_j\bar{x}_j$; $\bar{x}_v = \Sigma_j v_j\bar{x}_j$; $v_j = (n_j/s_j^2)/(\Sigma_k n_k/s_k^2)$;
$\bar{x}_w = \Sigma_j w_j\bar{x}_j/\Sigma_j w_j$; $w_j = n_j\hat{\delta}^2/\{n_j\hat{\delta}^2 + s_j^2\}$;
$\hat{\delta}^2$ = solution of the fixed point equation: $(J - 1)\hat{\delta}^2 = \Sigma w_j(\bar{x}_j - \bar{x}_w)^2$
$\bar{x}_u = \Sigma_j u_j\bar{x}_j/\Sigma_j u_j$; $u_j = n_j\tilde{\delta}^2/\{n_j\tilde{\delta}^2 + \bar{s}^2\}$, $\bar{s}^2 = \Sigma_j(n_j - 1)s_j^2/\Sigma_j(n_j - 1)$;
$\tilde{\delta}^2$ = solution of the fixed point equation: $(J - 1)\tilde{\delta}^2 = \Sigma u_j(\bar{x}_j - \bar{x}_u)^2$

iterative method of Carter and Rolph (1974) described below Table 1. The posterior mean and variance approximate values for PWT when n_j is large, $\hat{\delta}^2 >> s_j^2/n_j$ and $w_j \simeq 1$; this property implies design consistency of the posterior mean. On the other hand the posterior mean and variance approximate values for VWT when $\hat{\delta}^2 << s_j^2/n_j$ and $w_j \simeq n_j\hat{\delta}^2/s_j^2$. Thus the posterior mean is a Stein-type shrinkage estimate that behaves like $\bar{x}_\pi$ when the sample size is large and bias is the main issue, and moves towards $\bar{x}_v$ when the sample size is small and variance is more of a concern.

EBU, which stands for empirical Bayes shrinkage towards the unweighted mean, differs from EBV in assuming a constant variance across strata:

$$x_{ji}|\lambda_j, \phi_j \sim {}_{\text{ind}}G(\lambda_j, \phi^2),$$
$$(\lambda_j|\lambda, \delta^2) \sim {}_{\text{iid}}G(\lambda, \delta^2);$$
$$p(\lambda, \log\phi^2, \log\delta^2) \propto \text{const.} \qquad (11)$$

The posterior mean is again a Stein-type shrinkage estimate, which behaves like $\bar{x}_\pi$ when the sample size is large, and moves towards $\bar{x}$ when the sample size is small.

Refinements of EBV and EBU may be important in applications:

1. The assumption of exchangeability of the stratum means in (10) and (11) is crucial, as can be seen from simulations in Section 3. It can be refined by modifying the priors to model systematic variation. For example, if covariates are available that characterize the strata (such as measures of size), the prior mean $E(\lambda_j)$ might be modeled as a linear combination of these covariates.
2. The expressions for EBV and EBU in Table 1 effectively treat the variances ϕ_j^2 and δ^2 as if they were known. In small samples the posterior variance should be increased to allow for uncertainty in estimating these variances. See Rubin (1984) and Kackar

and Harville (1984) respectively for Bayesian and frequentist approaches to this problem.

3. A more elaborate treatment of the variances ϕ_j^2 is to specify a prior that models the ϕ_j^2 as iid from a common distribution, yielding estimates of ϕ_j^2 that smooth the sample variances $\{s_j^2\}$ towards a pooled value.

4. Although the normal is a standard baseline model, other distributions also yield design-consistent estimates of $\bar{X}$. For example, if x_{ji} is binary, the Beta-Binomial model

$$x_{ji} | \lambda_j \sim \text{Binomial}\ (\lambda_j);$$

$$\lambda_j \sim \text{Beta}\ (\lambda, \delta\lambda(1 - \lambda))$$

is more natural, where λ and $\delta\lambda(1 - \lambda)$ are the mean and variance of the Beta distribution; or if x_{ji} is a count, one might assume the Gamma-Poisson model

$$x_{ji} | \lambda_j \sim \text{Poisson}\ (\lambda_j);$$

$$\lambda_j \sim \text{Gamma}\ (\lambda, \delta\lambda)$$

where λ and $\delta\lambda$ are the mean and variance of the Gamma distribution. These models have more plausible variance structures for proportions and counts. They yield design-consistent estimates of $\bar{X}$ since in both cases the posterior mean of $\bar{X}_j$ converges to $\bar{x}_j$ in large samples.

5. A tempting modification to achieve robustness in the presence of outliers is to replace the normal by longer-tailed distributions such as the *t* (for example, West 1984; Lange, Little and Taylor 1989). Interestingly, estimates under such models are not design consistent, since they rely on an assumption of symmetry, often violated by the skewed variables many surveys measure. Transformation to symmetry is not necessarily a solution when interest is in the mean on the original scale (Rubin 1983b).

3. Simulation Study

3.1. Description of the study

Repeated-sampling properties of inferences based on (7)–(11) were assessed by a simulation study, for populations generated under a variety of conditions.

3.1.1. Populations studied

Sixteen populations of $N = 3600$ values of a variable X were constructed in 10 strata. Population sizes $\{N_j\}$ in the strata were as follows:

Stratum j:	1	2	3	4	5	6	7	8	9	10
N_j	1000	750	500	400	300	200	150	120	100	80

Equal-sized samples were selected from each stratum, so π_j increased with j. The 16 populations were generated in a 2^4 factorial design with the following four factors:

CORR = Correlation between selection probability (π_j) and mean (λ_j) (Low, High)

BVAR = Variation in Stratum Means (Low, High)

DIST = Distribution of X-Values (Normal, Chi-square)

CTAM = Contamination by Outliers (0%, 10%)

Specifically, values of X in stratum j were sampled from a distribution with mean

$$\lambda_j = 100 + k\delta_j$$

where the elements of $\delta = (\delta_1, \ldots, \delta_{10})$ were essentially linear transforms of uniform draws. Two choices of δ were used:

$$\delta_L = (-2, -7, 17, -12, 21, -4, -20, 2, 11, -4)$$

(CORR = Low)

$$\delta_H = (-5, -3, -13, 5, 6, 2, 21, 8, 28, 33)$$

(CORR = High).

In both cases $\Sigma N_j \delta_j / \Sigma N_j = 0$, so the expected value of the overall population mean is 100. The between-stratum variance was controlled by k, which was set at either 1 (BVAR = Low) or 2 (BVAR = High).

Let z_{ji} denote a standard normal deviate. For the uncontaminated normal populations, the value of X for unit i in stratum j was computed as

$$x_{ji} = \lambda_j + 84 z_{ji}.$$

For the contaminated normal populations

$$x_{ji} = \lambda_j + 60.94 z_{ji}^*$$

where $z_{ji}^* = z_{ji}$ with probability 0.9, $\sqrt{10} z_{ji}$ with probability 0.1; the scale factor 60.94 is chosen so that x_{ji} has the same marginal standard deviation (84) as for the uncontaminated populations. For the chi-square populations

$$x_{ji} = 0.2\lambda_j (z_{ji} + 2)^2$$

yielding scaled noncentral chi-squared deviates with mean λ_j, coefficient of variation 0.85 in each stratum, and average within-stratum standard deviation 82.44 when CORR = Low, 85.3 when CORR = High, close to that in the normal populations (84). Since the variance depends on the mean, these populations exhibit both skewness and heteroskedasticity. For the contaminated chi-square populations

$$x_{ji} = 0.1375\lambda_j (z_{ji}^* + 2.444)^2$$

where z_{ji}^* is defined above and the constants are chosen to match the mean and variance of x_{ji} in the absence of contamination. The 16 populations were generated from the same random number seeds to reduce the variance of comparisons of methods between populations.

To give some indication of the distributions generated by these methods, Figure 1 shows samples of size 50 from the five odd-numbered strata for the four populations with CORR = High and BVAR = High; samples with CORR = Low are similar but lack the systematic increase in the means across the strata, and samples with BVAR = Low have stratum means that are closer together. Figure 1A shows well-behaved homoskedastic normal data; Figure 1B shows a symmetric distribution with outliers; Figure 1C shows right-skewed data such as might be encountered with establishment surveys measuring inherently positive variables; within-stratum variances increase with the mean in these cases, as is typically the case in practice; and Figure 1D shows data with a combination of skewness, heteroskedasticity and outliers.

3.1.2. Sampling scheme

A stratified sample of $n_j = 10$ values was chosen without replacement from each stratum, yielding a total sample size of $n = 100$. This scheme implies probabilities of selection that increase across the strata from $\pi_1 = 1/100$ to $\pi_{10} = 1/8$. This procedure was repeated 1000 times for each population (starting with the same random seed for each population), and estimates of the population mean computed for each sample. To assess the effect of increasing sample size, this procedure was repeated with samples of 20 in each stratum (and a different random number seed), yielding a total sample size of $n = 200$.

3.2. *Results*

For each population and sample size, Table 2 displays average bias of the posterior mean of $\bar{X}$ under each model over the 1000 samples, Table 3 shows average root mean squared error (RMSE), and Table 4 shows the number of samples for which the 95% interval (6) does not include $\bar{X}$ – nominally we expect 50 such cases. Bias and RMSE are expressed as a percentage of the RMSE for

PWT, which is viewed as the standard method. We first discuss results for the five models PWT, VWT, UWT, EBU, and EBV described above. Results for a filtered variance-weighted procedure (VWTF) and a filtered unweighted procedure (UWTF) are described in Section 3.3.

A. PWT

As expected, PWT has good repeated sampling properties, with low bias and noncoverage close to or a bit above the nominal value (the large sample approximation was less satisfactory for 99% intervals, where noncoverage rates ranged from 1.6% to 4%). However, PWT does not always have the lowest RMSE, reflecting lack of control of variance.

B. UWT

The parameter CORR plays a key role in the performance of UWT. When CORR = Low the stratum means are weakly correlated with the sampling rates, and the unweighted average of the stratum means (100.25 when BVAR = Low, 100.5 when BVAR = High) is close to the weighted mean (100). Thus biases from assuming no stratum effects in UWT tend to cancel out. Thus the bias of UWT is small (Table 2A, B), and UWT has consistently lower RMSE than PWT, with reductions ranging from 22–27% (Table 3A, B). Noncoverage rates of UWT are close to nominal levels (Table 4A, B).

When CORR = High the unweighted average of the stratum means (108 when BVAR = Low, 116 when BVAR = High) is larger than the weighted average (100). Thus UWT is seriously biased when BVAR = Low, and disastrously biased when BVAR = High (Table 2C, D), when 95% intervals miss the true population value most of the time (Table 4C, D).

C. VWT

In the normal populations VWT has slightly higher RMSE than UWT, presumably because in these populations the within-stratum variance is constant, so a pooled estimate of variance is optimal. In the chi-squared populations, the within-stratum variance increases systematically with the mean, smaller means get a higher variance weight, so VWT yields a smaller estimate than UWT. Thus when CORR = Low and UWT is nearly unbiased, VWT has a negative bias, which is particularly severe for cases where BVAR = High. On the other hand when CORR = High, VWT tends to do better than UWT, since variance weighting reduces the positive bias of $\bar{x}_u$ (Table 2). Its performance is still very erratic, however.

D. EBU

EBU has RMSE values between those for PWT and UWT, reflecting the fact that it is a compromise between these estimators. When CORR = Low, the exchangeability assumption of the stratum means is reasonable, and EBU shrinks towards a good estimate. EBU then achieves good reductions in RMSE over PWT, particularly when the between variance is low. Noncoverage rates are also close to nominal levels. When CORR = High, exchangeability is violated, EBU shrinks towards a biased value, and is generally inferior to PWT. Nevertheless, it performs much better than UWT in this unfavorable situation, and actually has slightly lower RMSE than PWT when n = 100 and BVAR = Low. Coverage of intervals is poor when exchangeability is violated (Table 4).

E. EBV

In the normal populations EBV has similar RMSE values to EBU (Table 3). Its noncoverage rates are generally a bit higher,

Figure 1. Histograms and summary statistics for samples of size 50 from five strata in four populations
A) Distribution = Normal, 0% Contamination

	STRATUM				
MIDPOINTS	1	3	5	7	9
320					*
300	*			*	
280				**	*
260	*	*	*	**	***
240	*		*	***	**
220	*	*	**	****	*****
200	*		*	***	*
180	*	***	*	***	*********
160	****	**	*****	****	***
140	*****	****	*****	**	M**
120	**	**	*******	M***	***
100	*	******	M***	***	******
80	M***	******	*****	***	******
60	****	M******	*****	***	***
40	***	***	*******	*******	*
20	********	****	*		
0	******	***	**	***	*
– 20	**	*****		**	**
– 40	*	*		*	
– 60	*				
– 80	**		**		
– 100	*		*		
– 120		**			
	GROUP MEANS ARE DENOTED BY M'S				
MEAN	70.2	69.8	95.9	127.2	145.2
STD. DEV.	91.3	77.2	75.8	91.7	79.7

perhaps reflecting failure to allow for estimating the variances (Table 4). In the chi-squared populations EBV has higher RMSE than EBU when CORR = Low (and EBV is shrinking towards an inferior estimate), and lower RMSE than EBU when CORR = High (and EBV is shrinking towards a superior estimate). The disasters of VWT are largely mitigated: The RMSE of EBV ranges from 22% below PWT to 15% above PWT, whereas the RMSE of VWT ranges from 23% below PWT to 167% above PWT (Table 3). Noncoverage rates of intervals deteriorate when the assumptions of the model are violated (Table 4).

3.3. Modified inferences based on a filter for bias

The above results show that null-stratum effects models (VWT, UWT) provide sharper inferences than PWT under favorable conditions, but are very vulnerable to model misspecification. The random effects models (EBV, EBU) can also improve on the small-sample performance of PWT, and avoid the major disasters of the null-stratum models under misspecification. These methods, however, remain inferior to PWT when the stratum means are systematically related to the selection probabilities. Thus these

Figure 1 ctd. Histograms and summary statistics for samples of size 50 from five strata in four populations

B) Distribution = Normal, 10% Contamination

```
                                         STRATUM
MIDPOINTS 1              3               5               7         9

   455                                   *               **
   420
   385
   350
   315                                   *
   280                                                   *
   245    *                                              *         ****
   210    ***            *               *               ********  *******
   175    **             ***             *****           *******   ***********
   140    ********       *******         **********      M*******  M*******
   105    *****          ********        M************   ********  ************
    70    M*********     M*************  ************    ********  ***
    35    *************  ******          ****            ***       ****
     0    ***            *******                         *
   -35    ****                           ***
   -70                   **
  -105                                                   *
  -140    *              *
  -175                                                   *
  -210
  -245                                                   *
  -280
  -315
  -350                                   *
                         GROUP MEANS ARE DENOTED BY M'S
MEAN      74.4           70.2            96.2            129.6     146.4
STD. DEV. 74.6           64.0            97.9            121.4      60.2
```

methods cannot be recommended without a preliminary check of this violation of exchangeability.

The last two methods in Tables 2–4 incorporate such a check. Specifically, a filtered version of the variance-weighted procedure (VWTF) computes the quantity

$$T_v^2 = \frac{(\bar{x}_v - \bar{x}_\pi)^2}{\mathrm{Var}(\bar{x}_v - \bar{x}_\pi)}$$

$$\mathrm{Var}(\bar{x}_v - \bar{x}_\pi) = \Sigma_j (P_j - v_j)^2 s_j^2 / n_j$$

and bases inference on VWT if $T_v^2 < 3.92$ and on PWT if $T_v^2 > 3.92$. Similarly UWTF is a filtered version of UWT, with T_v^2 replaced by a statistic that compares the unweighted and π-weighted means

$$T_u^2 = \frac{(\bar{x} - \bar{x}_\pi)^2}{\mathrm{Var}(\bar{x} - \bar{x}_\pi)},$$

$$\mathrm{Var}(\bar{x} - \bar{x}_\pi) = \Sigma_j (P_j - p_j)^2 s_j^2 / n_j .$$

These filters are similar to the comparison of weighted and unweighted estimates proposed by Dumouchel and Duncan (1983) in the regression setting. They correspond to a preliminary 5% level significance test for bias of the null effects model estimator. As a

Figure 1 ctd. Histograms and summary statistics for samples of size 50 from five strata in four populations

C) Distribution = Chi-Squared, No Contamination

	STRATUM				
MIDPOINTS	1	3	5	7	9
500					*
475					
450				*	
425					
400				**	
375	*				*
350					**
325			*	**	*
300	*		*		*
275	*	*		***	
250			**	**	*****
225	*			**	*
200	*	*	**	***	*
175		*		***	******
150	*****	***	****	*	***
125	****	***	*****	M**	M**
100	***	****	M*******	***	***
75	M***	M********	*****	***	****
50	******	*************	********	******	*******
25	****************	*********	*********	**********	********
0	*******	******	*****	******	***
	GROUP MEANS ARE DENOTED BY M'S				
MEAN	76.9	68.2	91.2	127.7	136.7
STD. DEV.	82.6	55.3	73.3	119.5	116.2

general rule inferences based on a preliminary test can be unreliable (e.g., Hurvich and Tsai 1990). Also our filters are crude in that the critical value (3.92) ignores t corrections for estimating variance, and more seriously does not change according to sample size. However, the filters provide some protection against model misspecification.

For cases where UWT and VWT estimates are biased, applying the filter reduces the bias of UWT and VWT estimates significantly (Table 2). The effect of filtering on RMSE and noncoverage is summarized in Table 5, which compares average RMSE and noncoverage rate classified by whether the model (UWT or VWT) yields biased or unbiased estimates. Filtering considerably improves RMSE and noncoverage of UWT and VWT in problems where the models are biased, at the expense of some deterioration in problems where the models are unbiased. Interestingly UWTF and VWTF are still outperformed by their Empirical Bayes counterparts, EBU and EBV, and indeed remain inferior to PWT. Thus a preliminary check for bias helps but does not save the models that ignore stratum effects.

The filters were also applied to EBV and EBU, reverting to the PWT model when bias in the null stratum-effects model

Figure 1 ctd. Histograms and summary statistics for samples of size 50 from five strata in four populations

D) Distribution = Chi-Squared, 10% Contamination

	STRATUM				
MIDPOINTS	1	3	5	7	9
660				**	
630					
600					
570			*		
540					
510					
480					
450					
420					*
390				*	
360				*	
330	*				*
300				*	***
270				*	*
240	**		*	****	***
210		*	**	**	***
180	***	*	**	******	***
150	***	*	***	M	********
120	******	*******	********	*****	M**
90	M**	********	M**********	****	*****
60	*********	M***************	*********	********	********
30	*****************	************	**********	*********	********
0	******	****	***	*****	***
	GROUP MEANS ARE DENOTED BY M'S				
MEAN	75.4	67.3	97.1	139.8	133.0
STD. DEV.	69.9	46.9	87.9	144.8	98.1

was apparent. The resulting procedures had reduced bias, but showed no noticeable improvement in RMSE or noncoverage rates, suggesting that reduced bias was balanced by an increase in variance.

4. Normal Models for Slopes

The role of weights is more controversial when interest concerns the linear regression of one survey variable (say X_2) on another (say X_1). In this section I show that by focusing on a particular target quantity and modeling stratum effects as distinct parameters, the modeler is led to the probability-weighted inferences of randomization theory! Classic design-based inferences for regression thus have a model-based justification.

Let x_{1ji} and x_{2ji} denote values of X_1 and X_2 for unit i in stratum j, and write $x_{3ji} = x_{1ji}^2$, $x_{4ji} = x_{1ji}x_{2ji}$. I consider inference about the least squares regression slope of X_2 on X_1 in the entire population

$$B = \frac{\bar{X}_4 - \bar{X}_1\bar{X}_2}{\bar{X}_3 - \bar{X}_1^2},$$

Table 2. Average bias (× 100) of seven methods for estimating the mean, expressed as percentage of RMSE for PWT method

	Population parameters								
BVAR	Low	Low	Low	Low	High	High	High	High	
DIST	Normal	Normal	Chisq	Chisq	Normal	Normal	Chisq	Chisq	
CTAM	0%	10%	0%	10%	0%	10%	0%	10%	
A) CORR = Low, n = 100									MEAN
PWT	0	-1	1	0	0	-1	1	0	0
VWT	-4	4	-107	-108	0	14	-161	-160	-65
UWT	-8	-11	-14	-11	-6	-9	-13	-11	-10
EBV	-3	2	-74	-77	0	6	-87	-86	-40
EBU	-6	-8	-11	-9	-3	-5	-8	-7	-7
VWTF	-1	2	-54	-56	2	5	-49	-50	-25
UWTF	-6	-9	-11	-9	-4	-7	-10	-8	-8
B) CORR = Low, n = 200									MEAN
PWT	1	1	2	1	1	1	2	1	1
VWT	-4	7	-97	-99	3	21	-184	-181	-67
UWT	-9	-13	-21	-15	-7	-10	-18	-13	-13
EBV	-2	4	-66	-69	1	6	-70	-70	-33
EBU	-6	-9	-15	-11	-2	-3	-7	-6	-8
VWTF	-2	6	-54	-55	2	13	-33	-35	-20
UWTF	-7	-10	-18	-13	-6	-7	-17	-12	-11
C) CORR = High, n = 100									MEAN
PWT	0	-1	1	1	0	-1	1	1	0
VWT	72	85	-53	-53	153	174	-52	-53	34
UWT	65	66	58	68	140	144	136	152	104
EBV	47	52	-28	-32	64	64	-12	-16	18
EBU	47	46	44	50	64	64	79	84	60
VWTF	37	42	-30	-33	36	34	-25	-28	4
UWTF	37	36	33	38	32	32	42	45	37
D) CORR = High, n = 200									MEAN
PWT	1	1	1	0	1	1	0	-1	1
VWT	106	122	-8	-13	223	251	-5	-12	83
UWT	95	97	82	99	202	209	195	222	150
EBV	62	64	1	-4	63	60	12	6	33
EBU	58	58	55	65	58	60	78	85	65
VWTF	47	47	-6	-12	16	13	-1	-6	12
UWTF	44	44	42	46	12	11	28	20	31

where $\bar{X}_k = \Sigma_j P_j \bar{X}_{kj}$ is the overall population mean of X_k. Classical randomization-based inference, weighting sampled units by the inverse of their selection probabilities, yields the estimator

$$\hat{b}_\pi = \frac{\bar{x}_{4\pi} - \bar{x}_{1\pi}\bar{x}_{2\pi}}{\bar{x}_{3\pi} - \bar{x}^2_{1\pi}} \quad (12)$$

where $\bar{x}_{k\pi} = \Sigma_j P_j \bar{x}_{kj}$, and $\bar{x}_{kj}$ denotes the sample mean of $\{x_{kji}\}$ in statum j. A standard Taylor series approximation (for example Procedure 3 in Holt, Smith and Winter 1980) yields:

$$\mathrm{Var}(\hat{b}_\pi) \simeq \Sigma_j P_j^2 (1 - f_j) s^2_{dj}/n_j \quad (13)$$

Table 3. Average RMSE of seven methods for estimating the mean, expressed as percentage of RMSE for PWT

	Population parameters								
BVAR	Low	Low	Low	Low	High	High	High	High	
DIST	Normal	Normal	Chisq	Chisq	Normal	Normal	Chisq	Chisq	
CTAM	0%	10%	0%	10%	0%	10%	0%	10%	
A) CORR = Low, n = 100									MEAN
PWT	100	100	100	100	100	100	100	100	100
VWT	82	79	137	137	86	90	181	182	122
UWT	73	77	73	75	73	77	74	77	75
EBV	82	78	108	110	84	83	120	119	98
EBU	77	80	78	79	81	82	83	84	80
VWTF	90	85	120	119	92	92	132	130	108
UWTF	82	85	81	83	82	84	82	84	83
B) CORR = Low, n = 200									MEAN
PWT	100	100	100	100	100	100	100	100	100
VWT	77	78	125	128	81	92	200	201	123
UWT	73	78	75	75	73	78	75	76	75
EBV	81	79	103	105	87	87	112	111	96
EBU	79	81	79	79	86	86	87	86	83
VWTF	85	83	118	119	87	93	129	127	105
UWTF	82	85	83	83	82	84	83	83	83
C) CORR = High, n = 100									MEAN
PWT	100	100	100	100	100	100	100	100	100
VWT	112	117	113	110	179	198	126	123	135
UWT	97	101	97	106	157	163	162	178	133
EBV	98	97	91	91	115	112	94	94	99
EBU	93	95	93	99	113	114	122	130	107
VWTF	108	112	104	102	126	130	109	107	112
UWTF	104	104	100	108	127	128	131	139	118
D) CORR = High, n = 200									MEAN
PWT	100	100	100	100	100	100	100	100	100
VWT	132	144	92	92	238	267	108	110	148
UWT	119	123	113	128	214	223	213	240	172
EBV	108	109	87	88	116	113	97	97	102
EBU	104	106	102	110	114	115	125	132	113
VWTF	123	127	93	94	119	117	102	104	110
UWTF	117	119	113	122	116	115	131	128	120

where $s^2_{dj} = \Sigma_i(d_{ji} - \bar{d}_j)^2/n_j$, the sample variance of the d-values in stratum j, and

$$d_{ji} = \frac{\{x_{1ji} - \bar{x}_{1\pi}\}\{x_{2ji} - \bar{x}_{2\pi} - \hat{b}_\pi(x_{1ji} - \bar{x}_{1\pi})\}}{\bar{x}_{3\pi} - \bar{x}^2_{1\pi}}. \tag{14}$$

Consider the fixed stratum-effects model where (x_{1ji}, x_{2ji}) have distinct bivariate normal distributions in each stratum:

$$(x_{1ji},\ x_{2ji}) \sim {}_{\text{ind}}G(\lambda_j,\ \Phi_j);$$

$$p(\lambda_j,\ \Phi_j) = \text{const.}\ |\Phi_j|^{-1} \tag{15}$$

so $\lambda_j = (\lambda_{1j}, \lambda_{2j})$ and Φ_j are the mean and

Table 4. Noncoverage rate of 95% confidence intervals from seven methods, out of 1000 samples; target = 50

	Population parameters								
BVAR DIST CTAM	Low Normal 0%	Low Normal 10%	Low Chisq 0%	Low Chisq 10%	High Normal 0%	High Normal 10%	High Chisq 0%	High Chisq 10%	
A) CORR = Low, n = 100									MEAN
PWT	55	59	67	61	55	59	67	66	61
VWT	106	120	445	461	127	168	701	699	353
UWT	51	49	73	69	48	50	71	63	59
EBV	81	78	254	286	71	71	297	293	179
EBU	50	44	68	66	45	38	63	58	54
VWTF	114	131	302	308	129	154	340	321	225
UWTF	67	64	86	82	64	65	87	75	74
B) CORR = Low, n = 200									MEAN
PWT	50	50	60	62	50	50	60	56	55
VWT	77	85	348	355	95	147	784	720	326
UWT	51	45	63	51	46	42	64	52	52
EBV	62	61	194	187	47	46	191	172	120
EBU	49	37	52	46	45	34	47	40	44
VWTF	89	87	256	257	102	128	267	251	180
UWTF	70	56	77	64	64	54	77	61	65
C) CORR = High, n = 100									MEAN
PWT	55	59	69	57	55	59	70	59	60
VWT	256	331	251	244	589	710	268	263	364
UWT	153	128	92	100	478	470	290	309	253
EBV	145	172	127	123	171	160	99	101	137
EBU	112	84	66	63	165	117	100	92	100
VWTF	207	253	193	189	243	265	184	186	215
UWTF	156	118	109	105	235	221	188	187	165
D) CORR = High, n = 200									MEAN
PWT	50	50	57	59	50	50	54	59	54
VWT	348	438	132	123	859	899	167	173	392
UWT	252	236	156	180	792	749	574	639	447
EBV	157	152	79	79	131	117	69	72	107
EBU	138	119	97	88	134	104	94	78	107
VWTF	243	258	127	124	142	126	133	134	161
UWTF	196	172	151	168	119	108	174	157	156

covariance matrix of X_1 and X_2 in stratum j. Note that this working model implies distinct linear regressions of X_2 on X_1 *within strata*, whereas B is the least squares slope of X_2 on X_1 *in the whole population.* Strictly speaking, if (15) holds then the overall regression of X_2 on X_1 may not be linear; the assumption is that linearity is a good enough approximation for B to be a reasonable summary measure.

Lemma. The posterior mean and variance of B under model (15) are approximated by (12) and (13), respectively.

Proof. It is easily shown that under (15),

Table 5. Aggregate summaries of RMSE and noncoverage rates of unfiltered, filtered and Empirical Bayes methods, classified by whether or not model is biased

	Model Biased RMSE	Model Biased noncov	Model Unbiased RMSE	Model Unbiased noncov
VWT	148	440	83	116
VWTF	115	221	88	117
EBV	104	159	83	65
UWT	152	350	75	56
UWTF	119	163	83	70
EBU	110	104	81	49

$E(\bar{X}_{jk}|\text{data}) = \bar{x}_{jk}$ for all j, k, and hence $E(\bar{X}_k|\text{data}) = \bar{x}_{k\pi}$. Hence the first term of a Taylor series expansion yields $E\{B|\text{data}\} \simeq \hat{b}_\pi$. The same expansion yields

$$\begin{aligned}
&\text{Var}\{B|\text{data}\} \simeq \\
&\text{Var}\left(\sum_{k=1}^{4} \bar{X}_k \frac{\partial B}{\partial \bar{X}_k}(\bar{x}_\pi)|\text{data}\right) \\
&= \text{Var}\,\{\Sigma_j P_j \bar{D}_j|\text{data}\} \\
&= \Sigma_j P_j^2\, \text{Var}\,(\bar{D}_j|\text{data}),
\end{aligned}$$

where $\bar{D}_j$ is the population mean of d_{ji} (defined in (14)) in stratum j. Substituting

$$\text{Var}\,(\bar{D}_j|\text{data}) \simeq (1 - f_j)s_{dj}^2/n_j \qquad (16)$$

in this expression yields the right side of (13) as an approximation for the posterior variance of B. Note that (16) is itself an approximation since the exact posterior variance of $\bar{D}_j$ takes into account the special forms of skewness and kurtosis for the normal distribution; however (16) seems useful given that the Taylor series method is approximate, and the normality assumption of the model might not be trustworthy.

The lemma extends in an obvious way to multiple regression. Thus the use of probability weights in multiple regression can be justified from a modeling perspective, with this choice of target quantity and model. Chambers (1986) also provided a non-Bayesian, superpopulation model-based justification for regression with sample weights.

The ordinary unweighted least squares estimator of B, which ignores the design weights, is the posterior mean of B for the null stratum effects model that assumes the same regression line of X_2 on X_1 in each stratum. However as in the case of inference about $\bar{Y}$ in Section 2, inference under this model is vulnerable to model misspecification. This is demonstrated empirically in the simulations of Holt, Smith and Winter (1980) and Pfeffermann and Holmes (1985) for the case where Z is continuous.

One might argue that if the regressions of X_2 on X_1 vary across the strata as in (15), B is not an appropriate target for inference, and attention should focus on the within-stratum slopes, or linear combinations of these slopes. This argument seems to me valid if Z is considered a truly exogenous variable. However there are situations where Z is not exogenous and the unadjusted slope of X_2 on X_1 is of primary interest. The clearest case is when stratification is based directly on the outcome variable X_2 (e.g., Hausman and Wise 1977). More indirectly, the stratification could be based on a variable associated with X_2 that is considered an outcome rather than a cause of X_1 (Skinner, Holt and Smith 1989, Sec. 1.3). As a practical matter, regression analyses of survey data such as the Panel Study of Income Dynamics (PSID) often do not fully condition on the Z variable, which in the PSID case includes extensive geographic detail. Even when Z is truly exogenous, it is often of interest to compare the effect of X_1 on X_2 when Z is included and excluded from the regression, to assess the extent to which Z affects the relationship. Thus an analysis with Z not conditioned is often of descriptive interest (e.g., Little and Perera 1981).

Alternatives to (15) retain distinct regression lines across the strata, but model the slopes as random effects; see, for example, Pfeffermann and Lavange (1989) for a random coefficient model that regresses the slopes on stratum covariates.

5. Conclusion

This article emphasizes that in the setting of disproportionate stratified sampling, models need to be sensitive to differences between strata, by allowing distinct parameters across strata. Fixed effects models with this property for means and slopes yields π-weighted inferences similar to those arising in design-based theory. Such results bring design-based and model-based survey inferences closer together. I suspect that formal links between design-based and model-based inferences can also be found for the case of cluster sampling, leading me to echo a remark by Kish and Frankel (1974) in the discussion of their article on methods for design-based variance calculations.

> "We are not at odds with the Bayesian viewpoint . . . while a unified set of Bayesian foundations is far from complete, (we) conjecture that (1) the variance estimation techniques discussed in Section 5 will prove useful in the evaluation of posterior variance, and (2) under a Bayesian framework for inference (diffuse priors), the effects of clustering and stratification will be much the same as those we have observed."

Kish and Frankel's paper appears to me more concerned with practical inferences than in subtleties of statistical philosphy, and I think modelers as well as samplers need to take seriously their strictures on the need to analyze data in a way that takes into account features of the sample design.

Despite the practical utility of much design-based inference, I remain convinced that the model-based approach is preferable. For me, design-based methods are basically crude and asymptotic, good for large surveys where practical expediency requires simple estimation procedures, but inadequate for handling small samples. Indeed I feel (contrary to Kish and Frankel) that Bayesian foundations are much more complete and unified than design-based foundations for survey inference. What is currently lacking in the Bayesian approach is guidance about the choice of models for applications that are robust to features of the data created by the sample design.

The random effects models discussed in this article indicate one avenue of refinement for achieving better inferences from small stratified samples. However these gains are not achieved without some modeling effort; the simulations suggest that attention to the assumptions of the models, such as exchangeability of the stratum effects, may be needed to realize these gains, particularly if probability intervals for target quantities are required.

6. References

Battese, G.E. and Fuller, W.A. (1981). The Prediction of County Crop Areas Using Survey and Satellite Data. Proceedings of the Survey Research Methods Section, American Statistical Association 1981, 500–505.

Brewer, K.R.W. (1979). A Class of Robust Sampling Designs for Large Scale Surveys. Journal of the American Statistical Association, 74, 911–915.

Brewer, K.R.W. and Mellor, R.W. (1973). The Effect of Sample Structure on Analytical Surveys. Australian Journal of Statistics, 15, 145–152.

Carter, G.M. and Rolph, J.E. (1974). Empirical Bayes Methods Applied to Estimating Fire Alarm Probabilities. Journal of the American Statistical Association, 69, 880–885.

Chambers, R.L. (1986). Design-Adjusted

Parameter Estimation. Journal of the Royal Statistical Society, Series A, 149, 161–173.

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

Efron, B. and Morris, C. (1973). Data Analysis Using Stein's Estimator and Its Generalizations. Journal of the American Statistical Association, 70, 311–319.

Ericson, W.A. (1965). Optimum Sampling Strategies Using Prior Information. Journal of the American Statistical Association, 60, 750–771.

Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations, I. Journal of the Royal Statistical Society B, 31, 195–234.

Ghosh, M. and Lahiri, P. (1987). Robust Empirical Bayes Estimation of Means from Stratified Samples. Journal of the American Statistical Association, 82, 1153–1162.

Ghosh, M. and Meeden, G. (1986). Empirical Bayes Estimation in Finite Population Sampling. Journal of the American Statistical Association, 81, 1058–1062.

Hansen, M.H., Madow, W.G., and Tepping, B.J. (1983). An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys. Journal of the American Statistical Association, 78, 776–807 (with discussion).

Hausman, J.A. and Wise, D.A. (1977). Social Experimentation, Truncated Distributions and Efficient Estimation. Econometrica, 45, 919–938.

Holt, D., Smith, T.M.F., and Winter, P.D. (1980). Regression Analysis of Data From Complex Surveys. Journal of the Royal Statistical Society A, 143, 474–487.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Hurvich, C.M. and Tsai, C.-L. (1990). The Impact of Model Selection on Inference in Linear Regression. The American Statistician, 44, 214–217.

Kackar, R.N. and Harville, D.A. (1984). Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models. Journal of the American Statistical Association, 79, 853–862.

Kish, L. and Frankel, M.R. (1974). Inference From Complex Samples. Journal of the Royal Statistical Society B, 36, 1–37.

Klein, L.R. and Morgan, J.N. (1951). Results of Alternative Statistical Treatments of Sample Survey Data. Journal of the American Statistical Association, 46, 442–460.

Konijn, H.S. (1962). Regression Analysis in Sample Surveys. Journal of the American Statistical Association, 57, 590–606.

Lange, K., Little, R.J.A., and Taylor, J.M.G. (1989). Robust Statistical Inference Using the T Distribution. Journal of the American Statistical Association, 74, 881–896.

Little, R.J.A. (1983a). Comment on 'An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys', by M.H. Hansen, W.G. Madow, and B.J. Tepping. Journal of the American Statistical Association, 78, 797–799.

Little, R.J.A. (1983b), Estimating a Finite Population Mean from Unequal Probability Samples. Journal of the American Statistical Association, 78, 596–604.

Little, R.J.A. and Perera, S. (1981). Illustrative Analysis: Socioeconomic Determinants of Cumulative Fertility in Sri Lanka: A Marriage Cohort Approach. WFS Scientific Reports No. 12, International Statistical Institute, Voorburg, The Netherlands.

Little, R.J.A. and Rubin, D.B. (1983), Discussion of 'Six Approaches to Enumerative Sampling', by K.R.W. Brewer and C.E. Särndal. In W.G. Madow and I. Olkin (eds.), Incomplete Data in Sample Surveys, Vol. 3: Proceedings of the Symposium, New York: Academic Press.

Pfeffermann, D. and Holmes, D.J. (1985). Robustness Considerations in the Choice of a Method of Inference for Regression Analysis of Survey Data. Journal of the Royal Statistical Society A, 148, 268–278.

Pfeffermann, D. and Lavange, L. (1989). Regression Models for Stratified Multistage Samples. In C. Skinner, D. Holt, and T.M.F. Smith (eds.), The Analysis of Complex Surveys, Chichester: Wiley.

Robinson, P.M. and Tsui, K.W. (1979). On Brewer's Asymptotic Analysis in Robust Sampling Designs for Large Scale Surveys. Technical Report No. 79–43. Institute of Applied Mathematics and Statistics, University of British Columbia, Vancouver, Canada.

Rubin, D.B. (1981). Estimation in Parallel Randomized Experiments. Journal of Educational Statistics, 6, 377–400.

Rubin, D.B. (1983a), Comment on 'An Evaluation of Model-Dependent and Probability-Sampling Inferences in Sample Surveys', by M.H. Hansen, W.G. Madow, and B.J. Tepping. Journal of the American Statistical Association, 78, 803–805.

Rubin, D.B. (1983b). A Case Study of the Robustness of Bayesian Methods of Inference: Estimating the Total in a Finite Population Using Transformations to Normality. In G.E.P. Box, T. Leonard, and C.F. Wu (eds.), Scientific Inference, Data Analysis and Robustness, New York: Academic Press, 213–244.

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

Särndal, C.E. (1978). Design-Based and Model-Based Inference in Survey Sampling. Scandinavian Journal of Statistics, 5, 25–52.

Särndal, C.E. (1980). On π-Inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling. Biometrika, 67, 639–650.

Scott, A. and Smith, T.M.F. (1969). Estimation in Multistage Sampling. Journal of the American Statistical Association, 64, 830–840.

Skinner, C.J., Holt, D., and Smith, T.M.F. (1989). Analysis of Complex Surveys. New York: Wiley.

West, M. (1984). Outlier Models and Prior Distributions in Bayesian Linear Regression. Journal of the Royal Statistical Society B, 46, 431–439.

Received December 1989
Revised September 1991