SURVEY INFERENCE WITH WEIGHTS FOR DIFFERENTIAL SAMPLE SELECTION OR NONRESPONSE

Roderick J.A. Little, U.C.L.A.
Department of Biomathematics, U.C.L.A. School of Medicine, Los Angeles CA 90024

KEY WORDS: Bayesian inference, random effects, regression, sampling weights, 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 and Frankel 1974, Sarndal 1978, 1980, Holt, Smith and Winter 1980, DuMouchel and Duncan 1983, Hansen, Madow and Tepping 1983, Little 1983ab, Rubin 1983a, Pfefferman and Holmes 1985, Chambers 1986, Ghosh and Lahiri 1987) .

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 sensitive to misspecification errors rendered important by the sample design; in the context of disproportionate stratified sampling, 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 methods; e) the modeling approach provides principled modifications of π–weighted inference that improve precision 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 of fixed size 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. I focus on situations where the π_j vary across the strata j; important examples include disproportionate stratified sampling with sampling probabilities π_j, and poststratification 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.

Suppose K variables $X_1,...,X_K$ are measured in the survey, and let X denote the N × K matrix of values of these variables in the population. I consider inference about a finite population quantity Q = Q(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 $Q(Y) = \hat{\theta}_{pop}(Y)$ 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 Y fixed tend to focus on $\hat{\theta}_{pop}$ (Brewer and Mellor 1973; Hansen, Madow and Tepping 1983; DuMouchel and Duncan 1983). Although a modeler by philosophy, I like the survey sampler's focus on $\hat{\theta}_{pop}$ since it has one important conceptual advantage: $\hat{\theta}_{pop}$ is a real entity that exists irrespective of the validity of the model. The parameter θ is a fictitious entity that exists only within the context of the target model, and given model misspecification it is not clear what θ represents. Models are simplified descriptions that ignore fine structure, particularly in large populations. Focusing on $\hat{\theta}_{pop}$ keeps the target well–defined in the presence of model misspecification.

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

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

$$p(X^{(j)}) = \int \prod_{i=1}^{N_j} p(x_{ji}|\lambda_j,\varphi_j)\, p(\lambda_j,\varphi_j)\, d\lambda_j d\varphi_j, \quad (1)$$

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

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:

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

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

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

where λ and δ are respectively location and scale-/shape parameters, which are assigned uniform priors. The first application to surveys of random effects models of this type was in the seminal paper of Scott and Smith (1968), in the context of multi-stage cluster sampling.

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 x_{ji} $(i=1,\ldots,N_j)$ as exchangeable within strata. By De Finetti's theorem, this justifies an iid model for 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). More specifically, Little (1983b) and Pfefferman and Holmes (1985) have argued that models need to be constructed that yield *design-consistent* estimators, where design consistency means that as the sample sizes increase the estimates of $\hat{\theta}_{pop}$ converge to $\hat{\theta}_{pop}$ even when the model is misspecified. 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 not do this if the strata are not analytically meaningful. For example, $\hat{\theta}_{pop}$ 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 models 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). The latter article proves asymptotic optimality properties for empirical Bayes estimators of stratum means, and shows reductions in risk over stratum means by theory and simulation.

2. 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:

$$\bar{x}_\pi = \frac{1}{N}\left(\Sigma_j \Sigma_{i\in \mathscr{I}_j} x_{ij}/\pi_j\right) = \Sigma_j P_j \bar{x}_j, \qquad (4)$$

where $\mathscr{I}_j$ denotes the set of sampled units 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:

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

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.

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.

Fixed Stratum-Effects Model.

Since normal specifications are a natural starting point, consider the fixed stratum-effects model

$$x_{ji} \mid \lambda_j, \varphi_j \sim_{ind} G(\lambda_j, \varphi_j^2), \qquad (6)$$

$$p(\lambda_j, \ln \varphi_j) \propto \text{const.}, \qquad (7)$$

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. Standard Bayesian calculations (e.g. Ericson 1969) yields the posterior distribution of $\bar{X}_j$ as

$$X_j | x_{obs} \sim_{ind} t(\bar{x}_j,\ (1-f_j)s_j^2/n_j,\ n_j-1), \qquad (8)$$

where t(a,b,d) denotes the t distribution with mean a, scale b and degrees of freedom d, $\bar{x}_j, s_j^2$ denotes the sample mean and variance of X in stratum j, and f_j denotes the sampling fraction n_j/N_j. Note the presence of finite population corrections (fpc's) $1-f_j$ in (8), which do not appear in the posterior variance of λ_j.

The posterior distribution of $\bar{X}$ is a weighted combination of t distributions, which given large samples can be approximated by the asymptotic normal distribution:

$$\bar{X} | x_{obs} \sim G(\bar{x}_\pi,\ \Sigma_j P_j^2(1-f_j)s_j^2/n_j). \qquad (9)$$

Note that the posterior mean from this model is the π-weighted estimator (4). Moreover posterior probability intervals based on (9) are identical to the randomization-based confidence intervals from classical stratified sampling theory (Ericson 1969).

Null Stratum-Effects Model.

Now suppose model (6) is modified by assuming $\lambda_j=\lambda$ for all j (null stratum effects). The asymptotic posterior distribution of $\bar{X}$ is then easily shown to be normal with mean and variance:

$$E(\bar{X}|x_{obs}) = f\bar{x}_u + (1-f)\bar{x}_v; \quad Var(\bar{X}|x_{obs}) = \Sigma_j P_j(1-f_j)s_j^2/N + (1-f)^2\{\Sigma_j n_j/s_j^2\}^{-1} \qquad (10)$$

where f = n/N is the overall sampling fraction and $\bar{x}_u$ is the unweighted sample mean. Thus if f is small the posterior mean is the variance-weighted estimator (5). It is a better estimator than $\bar{x}_\pi$ when the stratum means are equal, but it is in general design inconsistent, and for dispropoortionate stratified sampling is not robust to departures from the assumption of equality in the stratum means. Since efficient sample designs are homogeneous within strata and heterogeneous between strata, this assumption is usually unrealistic, so inference for $\bar{X}$ based on (10) is not in general recommended.

Random Stratum-Effects Model

Now consider the random stratum-effects model (6) with prior

$$(\lambda_j | \lambda, \delta^2) \sim_{iid} G(\lambda,\ \delta^2); \qquad p(\lambda,\ \ln \varphi_j^2,\ \ln \delta^2) \propto \text{const.}, \qquad (11)$$

where the stratum means are assumed to be an iid sample from an underlying distribution. The posterior distribution of $\bar{X}$ is then asymptotically normal with mean and variance

$$E(\bar{X}|x_{obs}) = f\bar{x}_u + \Sigma_j P_j(1-f_j)\{w_j\bar{x}_j + (1-w_j)\bar{x}_w\},$$
$$Var(\bar{X}|x_{obs}) = \Sigma_j P_j^2(1-f_j)\{f_j+(1-f_j)w_j\}s_j^2/n_j + \{\Sigma_j P_j(1-f_j)(1-w_j)\}^2\hat{\delta}^2/\Sigma_j w_j, \qquad (12)$$

where $w_j = n_j\hat{\delta}^2/\{n_j\hat{\delta}^2 + s_j^2\}$, $\bar{x}_w = \Sigma_j w_j\bar{x}_j / \Sigma_j w_j$, and $\hat{\delta}^2$ is a consistent estimate of the between-stratum variance δ^2, computed for example by solving the fixed point equation:

$$(J-1)\hat{\delta}^2 = \Sigma w_j(\bar{x}_j - \bar{x}_w)^2$$

for $\hat{\delta}$ (Carter and Rolph 1974). Note that when n_j is large so that $\hat{\delta}^2 >> s_j^2/n_j$, $w_j \simeq 1$ and (12) approximates the standard answer (9); this property implies design consistency of the posterior mean. On the other hand if the between-stratum variance $\hat{\delta}^2 << s_j^2/n_j$ then $w_j \simeq n_j\hat{\delta}^2/s_j^2$ and (12) approximates (10). The posterior mean is a Stein-type shrinkage estimate that behaves like $\bar{x}_\pi$ when sample sizes are 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. The following refinements of (12) may be important in applications:

1. The assumption of exchangeability of the stratum means in (9) is crucial, as can be seen from simulations in Section 3. It can be refined to model systematic variation. For example, the prior mean λ might be modeled as a linear combination of stratum covariates.

2. The distribution (12) effectively treats 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 for example Rubin (1981).

3. The model includes a separate variance φ_j^2 for each stratum, which is poorly estimated by the sample variances in strata with small sample sizes. Thus some smoothing of the within-stratum variances may be useful. Checks of homoskedasticity might support treating these variances as equal, as in Little (1983b) and Ghosh and Lahiri (1987); the sample variances $\{s_j^2\}$ are replaced by a single pooled variance. A more elaborate approach 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 is more natural, or if x_{ji} is a count, one might assume the Gamma-Poisson model. These models have more plausible variance structures for proportions and counts, and also yield design-consistent estimates of $\bar{X}$.

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 since many surveys measure skewed variables. 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 Study

A simulation study was performed to illustrate the properties of the methods of Section 2.

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:

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

The 16 populations were points in a 2^4 factorial design, consisting of combinations of the following 4 factors:

CORR = Correlation between stratum (j) and stratum mean (μ_j) (Low, High)
BVAR = Variation in Stratum Means (Low, High)
DIST = Distribution of X-Values (Normal, Chisquare)
CONTAM = Contamination by Outliers (0%, 10%)

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

$$\mu_j = 100 + k\ \delta_j,$$

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

$$\delta_L = (-2,-7,17,-12,21,-4,-20,2,11,-4)\ \text{(CORR=Low)}$$
$$\delta_H = (-5,-3,-13,5,6,2,21,8,28,33)\ \text{(CORR=High)}.$$

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, set at either 1 (BVAR=low) or 4 (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

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

For the contaminated normal populations:

$$x_{ji} = \mu_j + 60.94z^*_{ji},$$

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:

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

yielding scaled noncentral chi-squared deviates with mean μ_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:

$$x_{ji} = 0.1375\mu_j(z^*_{ji}+2.444)^2,$$

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. Figure 1 shows samples of size 30 from the 5 odd-numbered strata for 4 of the 8 populations with CORR=High; plots for samples with CORR=Low are similar but lack the systematic increase in the means across the strata. Note that the chi-square samples are skewed and do not have constant variance across strata.

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 $\pi_j = 1/100$ to $\pi_j = 1/8$. This procedure was repeated 1000 times for each population (with the same random number 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 30 in each strata (and a different random number seed), yielding a total sample size of n=300.

Choice of Estimators and Standard Errors.

Tables 1-3 summarizes the results of applying the following procedures:
PWT: Normal inference based on the stratified mean and associated standard error given in Eq. (9).
VWT: Normal inference based on the variance-weighted estimator $\bar{x}_v$ and associated standard error given in Eq. (10).
UWT: Normal inference based on the unweighted mean $\bar{x}_u$ and associated variance $(1-f)s^2/(\Sigma_j n_j)$ where s^2 is the overall sample variance ignoring strata.
EBV: Normal inference based on the distribution (12) under the random effects model (6,11).
EBU: Normal inference under the random effects model (6,11), assuming constant within-stratum variance φ_j^2. The estimator under this model shrinks towards the unweighted mean rather than the variance-weighted mean.

For each population and sample size, Table 1 displays average bias of each estimator of $\bar{X}$ over the 1000 samples, Table 2 shows average root mean squared error (RMSE), and Table 3 shows the number of samples for which the 95% interval [est ± 1.96 (se)] does not include $\bar{X}$ – nominally we expect 50 such cases. RMSE for methods other than PWT are expressed as a percentage of values for PWT, which can be viewed as the standard method.

3.2 Results

A) PWT

As expected, PWT had 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 did not always have the lowest RMSE, reflecting lack of control of variance.

B) UWT

The parameter CORR played a key role in the performance of UWT. When CORR=Low the stratum means were weakly correlated with the sampling rates, and the unweighted average of the stratum means (100.25 when BVAR=Low, 101.0 when BVAR=High) was close to the weighted mean (100). Thus biases from assuming no stratum effects in UWT tended to cancel out. Thus the bias of UWT was small (Table 1A,B), and UWT had consistently lower RMSE than PWT, with reductions ranging from 18–28% (Table 2A,B). Noncoverage rates of UWT were close to nominal levels (Table 3A,B).

When CORR=High the unweighted average of the stratum means (108 when BVAR=Low, 132 when BVAR=High) was larger than the weighted average (100). Thus UWT was seriously biased when BVAR=Low, and disastrously biased when BVAR=High (Table 1C,D), when 95% confidence intervals missed the true population value most of the time (Table 3C,D).

VWT

In the normal populations VWT had slightly higher RMSE than UWT, presumably because in these populations the within–stratum variance was constant, so a pooled estimate of variance was optimal. In the chi–squared populations, the within–stratum variance increased systematically with the mean, smaller means got a higher variance weight, so VWT yielded a smaller estimate than UWT. Thus when CORR=Low and UWT was nearly unbiased, VWT had a negative bias, which was particularly severe for cases where BVAR=High. On the other hand when CORR=High, VWT tended to do better than UWT, since variance–weighting reduced the positive bias of $\bar{x}_u$ (Table 1). The erratic behavior of UWT and VWT emphasizes their sensitivity to the structure of the population.

EBU

EBU had RMSE values between those for PWT and UWT, reflecting the fact that it was a compromise between these estimators. When CORR=Low EBU was shrinking towards a good value, and the exchangeability assumption of the stratum means was justified. EBU then achieved good reductions in RMSE over PWT when the between variance was low and modest reductions when the between variance was high. Noncoverage rates were also close to nominal levels. When CORR=High, EBU was shrinking towards a biased value, and was generally inferior to PWT. However it performed much better than UWT in this unfavorable situation, and actually had slightly lower RMSE than PWT when n=100 and BVAR=Low.

EBV

In the normal populations EBV had similar RMSE values to EBU (Table 2). Its noncoverage rates were generally a bit higher, perhaps reflecting failure to allow for estimating the variances (Table 3). In the chi–squared populations it had higher RMSE than EBU when CORR=Low (and EBV was shrinking towards an inferior estimate), and lower RMSE than EBU when CORR=High (and EBV was shrinking towards a superior estimate). The disasters of VWT were largely mitigated: The RMSE of EBV ranged from 20% below PWT to 25% above PWT, whereas the RMSE of VWT ranged from 23% below PWT to 810% above PWT (Table 2). However noncoverage rates of this method (and EBU) deteriorated when the assumptions of the model were violated (Table 3).

4 . INFERENCE ABOUT A SLOPE.

We now consider the role of weights when interest concerns the linear regression of one survey variable (say X_2) on another (say X_1). The choice of target quantity is a key issue. Let x_{1ji} and x_{2ji} denote values of X_1 and X_2 for unit i in stratum j, and write $x_{3ji}=x_{1ji}^2$, $x_{4ji}=x_{1ji}x_{2ji}$. Also define the slope function

$$B(a_1,a_2,a_3,a_4) \equiv (a_4 - a_1a_2)/(a_3 - a_1^2).$$

The population least squares slope of X_2 on X_1 in stratum j is then

$$B_j = B(\bar{X}_{1j},\bar{X}_{2j},\bar{X}_{3j},\bar{X}_{4j})$$

where $\bar{X}_{kj} = \Sigma_{i=1}^{N_j} x_{kji}/N_j$, the population mean of X_k in stratum j. The least squares regression slope in the entire population is

$$B_u = B(\bar{X}_1,\bar{X}_2,\bar{X}_3,\bar{X}_4)$$

where $\bar{X}_k = \Sigma_j P_j \bar{X}_{kj}$ is the overall population mean of X_k. Two target quantities (or superpopulation analogs) are considered in the literature, B_u and $B_a = \Sigma_j P_j B_j$. These quantities are not in general equal unless $\bar{X}_1=\bar{X}_{1j}$ and $\bar{X}_3=\bar{X}_{3j}$ for all j, that is, the mean and variance of X_1 are the same in all the strata. Survey samplers tend to consider B_u and modelers B_a (cf. DuMouchel and Duncan 1983), but in my view the choice is a substantive issue: whether the effect of X_2 on X_1 is adjusted or not adjusted for the stratifying variable Z. B_u measures the unadjusted effect of X_2 on X_1, and B_a measures the effect of X_2 on X_1 *adjusted for Z*, the overall slope in a hypothetical population with the same values of $\{P_j\}$ and $\{B_j\}$ as in the actual population, but where the distribution of X_1 is the same in all the strata. Note that B_a is not a slope in the actual population; if the effects of X_1 and Z are additive it equals the common within–stratum slope; if X_1 and Z interact (that is the slopes vary across the strata), then the definition of adjusted effect is sensitive to the weights $\{P_j\}$, other choices such as $\{P_j^*=1/J\}$ being equally plausible.

Classical randomization–based inference, weighting sampled units by the inverse of their selection probabilities, yields the estimators

$$\hat{b}_u = B(\bar{x}_{1\pi}, \bar{x}_{2\pi}, \bar{x}_{3\pi}\ \bar{x}_{4\pi}) , \qquad (13)$$

for B_u, and

$$\hat{b}_a = \Sigma_j P_j b_j,\quad b_j = B(\bar{x}_{1j}, \bar{x}_{2j}, \bar{x}_{3j}, \bar{x}_{4j}) \qquad (14)$$

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

$$\mathrm{Var}(\hat{b}_u) = \Sigma_j P_j^2 (1-f_j) s_{dj}^2/n_j, \qquad (15)$$

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

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

The sampling variance of $\hat{b}_a$ is simply

$$\mathrm{Var}(\hat{b}_a) = \Sigma_j P_j^2 (1-f_j) s_{bj}^2, \qquad (16)$$

where s_{bj}^2 is the usual least squares estimate of $\mathrm{Var}(b_j)$. I now provide a model–based justification for inferences based on (13–16).

<u>Bivariate Normal Model with Fixed Stratum Effects.</u>

Suppose x_{1ji} and x_{2ji} have distinct bivariate normal distributions in each stratum:

$$(x_{1ji}, x_{2ji}) \sim_{ind} G(\lambda_j, \Phi_j);$$
$$p(\lambda_1, \dots, \lambda_J) = \text{const}, \qquad (17)$$

so $\lambda_j = (\lambda_{1j}, \lambda_{2j})$ and Φ_j are the mean and 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 the target model that yields B_u as the target quantity implies a linear regression of X_2 on X_1 *in the whole population.*

<u>Lemma</u>. The posterior mean and variance of B_u under model (17) are approximated by (13) and (15), respectively. The posterior mean and variance of B_a are approximated by (14) and (16), respectively.

<u>Proof.</u> Standard results on Bayesian regression with flat priors applied within stratum j yield $E(B_j|\text{data}) = b_j$ and $\mathrm{Var}(B_j|\text{data}) = (1-f_j)s_{bj}^2$. Hence the posterior mean and variance of B_a are given by (14) and (16). Also, it is easily shown that under (17), $E(\bar{X}_{jk}|\text{data}) = \bar{x}_{jk}$ 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 $E\{B_u|\text{data}\} \simeq B(E\{\bar{X}|\text{data}\}) = B(\bar{x}_\pi) = \hat{b}_u$. The same expansion yields

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

where $\bar{D}_j$ is the population mean of d_{ji} (defined below Eq. 15) in stratum j. The approximation (15) for $\mathrm{Var}(B_u|\text{data})$ follows by substituting

$$\mathrm{Var}(\bar{D}_j|\text{data}) \simeq (1-f_j)s_{dj}^2/n_j. \qquad (18)$$

Note that (18) 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 (18) 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 (1983) provided a non–Bayesian, superpopulation–model based justification for regression with sample weights. The Bayesian approach given here seems to me more straightforward and direct.

<u>Regression Models with Null Stratum Effects.</u>

Model (17) implies distinct regressions lines of X_2 on X_1 in each strata. Assuming a common slope and residual variance across the strata yields the additive model $[X_1+Z]$:

$$x_{ji} \sim_{ind} G(\lambda_j, \varphi_j^2);$$
$$y_{ji}|x_{ji} \sim_{ind} G(\alpha_j + \beta x_{ji}, \gamma^2), \qquad (19)$$
$$p(\lambda_j, \alpha_j, \beta, \ln \varphi_j, \ln \gamma) = \text{const}.$$

Assuming further a constant intercept across the strata yields the model $[X_1]$:

$$x_{ji} \sim_{ind} G(\lambda_j, \varphi_j^2);$$
$$y_{ji}|x_{ji} \sim_{ind} G(\alpha + \beta x_{ji}, \gamma^2), \qquad (20)$$
$$p(\lambda_j, \alpha, \beta, \ln \varphi_j, \ln \gamma) = \text{const}.$$

Inferences based on (19) or (20) for disproportionate stratified samples are not recommended, since they do not yield design–consistent estimators of B_u or B_a, and are sensitive to model misspecification. In particular it is easily seen that the posterior mean of B under (20) is the (unweighted) least squares estimate

$$\hat{b}_{ls} = B(\bar{x}), \qquad (21)$$

where $\bar{x} = (\bar{x}_1, \bar{x}_2, \bar{x}_3, \bar{x}_4)$, and $\bar{x}_k$ is the unweighted sample mean of $\{x_{kji}\}$, k=1,4. Although $\hat{b}_{ls}$ is more precise than $\hat{b}_u$ if the model is true, it is not design–consistent for B_u and can be badly biased when the regression lines vary across strata. In particular $\hat{b}_{ls}$ performs poorly in the simulations of Pfefferman and Holmes (1985), for the case of a continuous stratifying variable.

Regression Models with Random Slopes and Intercepts.

A better alternative to (17) retains distinct regression lines across the strata, but replaces the flat prior in (17) by an informative prior for the slopes and intercepts:

$$(\alpha_j, \beta_j \mid \alpha, \beta, \lambda_j, \varphi_j, \gamma_j) \sim_{iid} G\{(\alpha,\beta), \Delta\};$$
$$p(\alpha, \beta, \ln \Delta, \lambda_j, \ln \varphi_j^2, \ln \delta^2) \propto \text{const.} \quad (22)$$

Estimates for this model shrink between the posterior means for the fixed effects model (17) and estimates for the null stratum effects model (20). In particular the resulting estimate of B_u is a compromise between $\hat{b}_{ls}$ and $\hat{b}_u$. Alternatively, letting Δ_{11} and Δ_{12} tend to infinity but keeping Δ_{22} finite, estimates shrink towards the posterior mean for model (19) with common slope but distinct intercepts.

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 yield 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 Frankel and Kish (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"

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

Despite the practical utility of much design–based inference à la Frankel and Kish, 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. Design–based methods fail to exploit specific features of the populations being sampled, have difficulties in the area of ancillary statistics, and appear to me to have no adequate machinery 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.

ACKNOWLEDGEMENTS

This research was supported by grant USPH MH 37188 from the National Institute of Mental Health. I wish to thank Wil Dixon and Robert Weiss for helpful comments, and Don Rubin for many stimulating discussions.

REFERENCES

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, 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. (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).

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.

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," To appear in Journal of the American Statistical Association.

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. (1986), "Survey nonresponse adjustments", International Statistical Review 54, 139–157.

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

Little, R. J. A., and Rubin, D.B. (1987), Statistical Analysis with Missing Data, New York: John Wiley.

Pfefferman 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

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 Scientific Inference, Data Analysis and Robustness, New York: Academic Press, 213–244.

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

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

Sarndal, 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.

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

Table 1. Average Bias of Five Methods for Estimating the Mean

	Between Var = Low				Between Var = High			
	Normal		Chi–squared		Normal		Chi–squared	
	0%Con	10%Con	0%Con	10%Con	0%Con	10%Con	0%Con	10%Con
A) CORR=Low n=100								
PWT	0.05	−0.09	0.15	0.04	0.05	−0.09	0.08	−0.06
UWT	−0.84	−1.10	−1.58	−1.15	−0.21	−0.48	−1.22	−1.01
VWT	−0.47	0.45	−11.94	−11.12	1.06	3.36	−55.92	−51.95
EBU	−0.61	−0.82	−1.19	−0.88	−0.03	−0.17	−0.35	−0.30
EBV	−0.35	0.26	−8.22	−7.91	0.18	0.42	−10.15	−9.08
B) CORR=Low n=300								
PWT	−0.02	−0.04	−0.12	−0.04	−0.02	−0.04	−0.15	−0.04
UWT	−0.34	−1.13	−0.68	−1.18	0.28	−0.51	0.21	−0.32
VWT	−0.58	0.04	−7.26	−7.03	−0.48	2.59	−58.71	−54.83
EBU	−0.18	−0.66	−0.47	−0.76	0.01	−0.08	−0.12	−0.08
EBV	−0.27	−0.16	−3.67	−3.84	−0.07	0.05	−3.84	−3.51
C) CORR=High n=100								
PWT	0.05	−0.09	0.16	0.06	0.05	−0.09	0.10	0.03
UWT	7.15	6.89	6.26	6.80	31.75	31.49	30.13	30.82
VWT	7.90	8.83	−5.69	−5.32	34.57	36.91	−17.89	−17.86
EBU	5.12	4.79	4.76	5.02	5.29	5.16	9.53	9.27
EBV	5.14	5.48	−3.00	−3.18	5.60	5.08	−1.77	−2.18
D) CORR=High n=300								
PWT	−0.02	−0.04	−0.11	−0.03	−0.02	−0.04	−0.08	−0.02
UWT	7.65	6.86	7.22	6.64	32.25	31.46	31.85	30.99
VWT	8.06	7.96	1.07	0.53	34.09	34.26	−10.57	−11.29
EBU	4.08	3.64	4.54	4.10	1.99	1.90	4.30	3.87
EBV	4.29	3.84	1.03	0.42	2.12	2.01	−0.11	−0.09

Table 2. Average RMSE of Five Methods for Estimating the Mean; RMSE for Methods Other than PWT Expressed as Percentage of RMSE for PWT.

	Between Var = Low				Between Var = High			
	Normal		Chi–squared		Normal		Chi–squared	
	0%Con	10%Con	0%Con	10%Con	0%Con	10%Con	0%Con	10%Con
A) CORR=Low, n=100								
PWT	10.99	10.45	11.13	10.32	10.99	10.45	11.31	10.42
UWT	73	77	73	75	72	77	78	82
VWT	82	79	137	137	103	125	499	510
EBU	77	80	78	79	91	91	92	93
EBV	82	78	108	110	92	93	126	125
B) CORR=Low, n=300								
PWT	6.07	5.98	6.24	6.23	6.07	5.98	6.48	6.29
UWT	72	74	74	72	72	72	80	77
VWT	77	78	142	137	97	141	910	883
EBU	82	81	82	80	97	97	96	96
EBV	83	80	101	98	97	97	111	108
C) CORR=High, n=100								
PWT	10.99	10.45	10.80	10.08	10.99	10.45	9.79	9.26
UWT	97	101	97	106	298	311	326	352
VWT	112	117	113	110	337	380	232	240
EBU	93	95	93	99	110	111	142	148
EBV	98	97	91	91	113	109	101	101
D) CORR=High, n=300								
PWT	6.07	5.98	6.02	6.07	6.07	5.98	5.46	5.51
UWT	145	135	144	133	536	531	594	572
VWT	153	153	92	83	570	587	241	246
EBU	113	108	117	111	105	105	129	126
EBV	116	108	91	84	106	106	101	101

Table 3. Noncoverage Rate of 95% Confidence Intervals from Five Methods, Out of 1000 Samples; Target = 50

	Between Var = Low				Between Var = High			
	Normal		Chi–squared		Normal		Chi–squared	
	0%Con	10%Con	0%Con	10%Con	0%Con	10%Con	0%Con	10%Con
A) CORR=Low, n=100								
PWT	55	59	67	61	55	59	72	73
UWT	51	49	73	69	47	48	78	67
VWT	106	120	445	461	203	312	1000	1000
EBU	50	44	68	66	44	33	45	41
EBV	81	78	254	286	57	58	264	245
B) CORR=Low, n=300								
PWT	50	55	58	60	50	55	57	54
UWT	41	33	53	58	38	35	51	52
VWT	62	80	412	439	135	322	1000	1000
EBU	42	29	44	50	46	50	35	44
EBV	50	46	135	133	47	58	124	117
C) CORR=High, n=100								
PWT	55	59	69	57	55	59	68	60
UWT	153	128	92	100	973	953	809	854
VWT	256	331	251	244	939	958	614	647
EBU	112	84	66	63	97	70	40	34
EBV	145	172	127	123	105	87	96	101
D) CORR=High, n=300								
PWT	50	55	58	60	50	55	59	62
UWT	364	314	238	227	1000	1000	1000	1000
VWT	430	477	92	93	1000	1000	590	666
EBU	153	130	116	102	78	67	17	21
EBV	166	143	65	67	84	69	63	66

Description of Methods:

PWT= Probability–weighted (stratified mean)
VWT= Weighted by sample variance in each stratum
UWT= Unweighted (as 2, but with constant variance across strata)
EBV= Empirical Bayes, shrinking between PWT and VWT
EBU= Empirical Bayes, shrinking between PWT and UWT

Figure 1. Histograms for Samples of Size 30 from 5 Strata.

A) Normal, 0% Contamination, Low Between Variance

MIDPOINTS	STRATUM 1	3	5	7	9
350			*		
325					*
300					*
275				*	*
250	*		*	***	***
225	*	*	**		
200	*	**		**	****
175	*	*	******	****	**
150	***	*****	***	***	M
125	**	**	M***	M*	***
100	***	*	***	****	******
75	M******	M*****	**	*	**
50	*****	*****	**	******	**
25	**	****	***	**	***
0	*		*	*	
–25	*	**	*	*	
–50	*	*	*		*
–75	*				
MEAN	85.61	86.02	117.51	122.26	138.63
STD.DEV.	75.87	72.29	84.29	78.72	91.68

B) Normal, 10% Contamination, High Between Variance

MIDPOINTS	STRATUM 1	3	5	7	9
390				*	*
360					*
330					*
300			*	*	****
270				***	***
240				**	***
210	*		***	******	M**
180	**		******	M****	*******
150	**	***	*****	***	****
120	***	*	M****	******	*
90	****	*****	****	**	**
60	M*********	***	***		
30	****	M********	***	*	
0	**	*****			
–30	**	*			
–60		*			
–90					
–120		*			
–150					
–180					
–210		*			
MEAN	74.77	36.94	132.35	185.25	220.24
STD.DEV.	59.86	73.01	61.15	71.25	76.69

C) Chisquare, 0% Contamination, High Between Variance

MIDPOINTS	STRATUM 1	3	5	7	9
800					*
750					
700					
650					*
600			*	*	
550					*
500					***
450				***	
400					
350				*	***
300			**	*	**
250	*		*	****	M*
200	*		******	M**	
150	**	*	M***	***	****
100	*****	*****	******	***	*******
50	M**************	M************	******	*********	****
0	******	***********	****	**	**
MEAN	69.66	44.83	137.39	180.65	240.56
STD.DEV.	58.42	35.37	118.63	150.25	210.81

D) Chisquare, 10% Contamination, Low Between Variance

MIDPOINTS	STRATUM 1	3	5	7	9
400			*		*
375					
350					
325				**	**
300					*
275				*	***
250	*		*	**	
225					
200	*	*	**		**
175	**	**		**	**
150	*	*	******	***	M
125	***	****	****	M***	*
100	**	***	M**	**	**
75	M*****	M*	****	***	********
50	********	***********	****	*****	**
25	***	***	***	*****	****
0	***	***	**	*	*
MEAN	82.72	77.55	111.28	117.55	139.15
STD.DEV.	60.21	54.82	82.18	90.60	110.10