ROBUST MODEL-BASED INFERENCE FOR A FINITE POPULATION MEAN FROM UNEQUALLY WEIGHTED SAMPLES

Roderick J. A. Little, Datametrics Research, Inc.

1. It is remarkable that despite the proliferation of statistics research in numerous specialized directions, there is no agreement about how to tackle the basic problem of estimating a finite population mean from a probability sample of the population. A variety of inferential approaches have been suggested. However, it is argued elsewhere (Little and Rubin, 1981) that the main distinction lies between two approaches, the randomization approach where population items are treated as fixed and inferences are based on the known distribution of sample selection, and the model-based approach, where inferences are based on a model for the population items. Under simple random sampling and a normal model, both approaches lead to estimating the population mean by the sample mean, with estimated variance $(1-n/N)s^2/n$, where s is the sample standard deviation, n is the sample size and N is the population size. However, when the sample design deviates from simple random sampling, and in particular when variable probability sampling is adopted, the two approaches lead to different inferences.

Let y_i denote the value of an item y for unit i, $s_i=1$ or 0 according to whether unit i is selected or not, $\pi_i=E(s_i)$ denote the probability of selection for unit i, for i=1,...,N. Then the Horvitz-Thompson estimator

$$HT = N^{-1}\sum_{i=1}^{N} \pi_i^{-1} s_i y_i \qquad (1)$$

(Horvitz and Thompson, 1952) is the standard estimator of the population mean $\bar{Y}$ in the randomization theory. In particular, it is design-unbiased. However in small or moderate sized samples it can have a high mean squared error. See, for example, Basu's famous circus example (Basu, 1971).

Sarndal (1980) proposes improving the HT estimator by selecting an estimator of $\bar{Y}$ from the class of generalized difference estimators

$$Y_{GD}(\hat{\mu}) = N^{-1}\sum_{i=1}^{N} \pi_i^{-1} s_i (y_i - \hat{\mu}_i) + N^{-1}\sum_{i=1}^{N} \hat{\mu}_i, \qquad (2)$$

where $\hat{\mu}_i$ is a value chosen to predict y_i, available for all units in the population, the first term on the right hand side of (2) is the HT estimator applied to the residuals $y_i-\hat{\mu}_i$, and $\hat{\mu}=(\hat{\mu}_1,\ldots,\hat{\mu}_N)^T$. Setting $\hat{\mu}_i=0$ gives $Y_{GD}(0)=HT$. Setting $\hat{\mu}_i=\bar{y}$, the sample mean, gives the estimator

$$SA=\bar{y}+N^{-1}\sum_{i=1}^{N} \pi_i^{-1} s_i (y_i-\bar{y}). \qquad (3)$$

In his discussion of Basu (1971), Hajek suggests the ratio estimator

$$HK=\sum_{i=1}^{N} \pi_i^{-1} s_i y_i \Big/ \sum_{i=1}^{N} \pi_i^{-1} s_i, \qquad (4)$$

where the population size N in HT is replaced by the denominator of (4). HK can be obtained as a generalized difference estimator by setting $\hat{\mu}_i=HK$. Sarndal compares SA and HK, and also considers estimators where $\hat{\mu}_i$ is obtained by regression on covariates x known for all units.

In all these cases the predicted values vary in repeated sampling, and hence the estimators are not design unbiased. However, Sarndal shows that they are in a sense asymptotically design unbiased. The inclusion of modelling information tends to reduce the variance, and as a result the estimators often have lower mean squared error than the Horvitz-Thompson estimate.

Sarndal's approach is a compromise between the modelling and randomization approaches. To the pure modeller, the estimate of $\bar{Y}$ should be based entirely on an appropriate model for the population, and the Horvitz-Thompson estimator applied to the residuals in (2) is unnecessary. The latter component affords protection against mispecification error. However we suggest that this component is not necessary, provided models are chosen which yield estimates which are insensitive to mispecification error in large samples. Indeed this requirement is a key ingredient to successful modelling in the survey context.

We shall adopt the Bayesian modelling approach to survey inference, as discussed by Ericson (1969). Other modelling formulations, such as the superpopulation approach of Royall (1970), lead to similar estimators for the problem we consider.

2. A Bad Model and a Better One.

Suppose we specify that the values y_i are iid Normal with mean μ and variance σ^2. Then the standard Bayesian approach with flat priors on μ and σ^2 leads to the posterior distribution

$$Y|\text{data} \sim G(\bar{y},(1-n/N)s^2/n) \qquad (5)$$

for the population mean, where G(a,b) is the Normal (Gaussian) distribution with mean a, variance b, and $\bar{y}$ and s are the sample mean and standard deviation. This distribution yields estimates and probability intervals for $\bar{Y}$ which are numerically equivalent to randomization analogs based on simple random sampling.

Suppose that the correct model specifies that the population forms J subclasses, where the item values in subclass j are Normally distributed with mean μ_j and variance σ^2, not all the μ_j being equal. Then if the units are selected by simple random sampling, we conjecture that (5) leads to approximately valid inferences for $\bar{Y}$ even though the model is wrong. On the other hand, if items are sampled at different rates π_j in each subclass, then (5) is unsatisfactory if the μ_j are not equal. In particular, the sample distribution over the subclasses no longer approximates the distribution in the population, and hence the centre $\bar{y}$ of the posterior distribution does not weight the subclass sample means appropriately.

Let us define J subclasses such that the probability of selection is a constant π_j in subclass j. Let N_j and n_j denote the population and sample size in subclass j, and

let y_{ij} denote the value of y for unit i in subclass j. The population mean can be written

$$\bar{Y} = \sum_{j=1}^{J} p_j \bar{Y}_j, \qquad (6)$$

where p_j is the population proportion and $\bar{Y}_j$ is the population mean in subclass j. A key aspect of the model is heterogeneity of the population across subclasses. Accordingly, we consider the model

$$\begin{matrix} y_{ij} & \text{ind } G(\mu_j, \sigma^2) \\ \mu_j & \text{ind } G(\mu, \tau^2). \end{matrix} \qquad (7)$$

We assume first flat priors for μ, σ^2 and τ^2. If the population proportions p_j are known, then the posterior mean of $\bar{Y}$ given the data is

$$M1 = \sum_{j=1}^{J} p_j \, E(\bar{Y}_j | \text{data}, p_j), \qquad (8)$$

where

$$E(\bar{Y}_j | \text{data}, p_j) = \lambda_j \bar{y}_j + (1-\lambda_j)\bar{y} + (1-\lambda_j)(\bar{y}_j - \bar{y})n_j/N_j, \qquad (9)$$

and

$$\lambda_j = n_j \hat{\tau}^2 (n_j \hat{\tau}^2 + \hat{\sigma}^2)^{-1}, \qquad (10)$$

where $\hat{\tau}^2$ and $\hat{\sigma}^2$ are maximum likelihood estimates of τ^2 and σ^2, obtained from a random effects analysis of variance of the data.

Equation (9) gives an estimate of the mean for subclass j under this model. The last term on the right hand side is a finite population correction which is small if n_j/N_j is small. Aside from this term, the estimate is a weighted average of $\bar{y}_j$ and $\bar{y}$, with weights given by (10). Hence M1 is a shrinkage type estimator, where shrinkage is from the weighted to the unweighted sample mean. Note that the weight (10) given to $\bar{y}_j$ depends on the relative size of the within and between subclass variances, and is a monotone function of the sample size n_j which increases to one as n_j becomes large. If the subclass sample sizes n_j are large, we might simply set $\lambda_j=1$. The resulting estimator of $\bar{Y}$ is simply

$$PS = \sum_{j=1}^{J} p_j \bar{y}_j, \qquad (11)$$

which is a post-stratified estimator (Holt and Smith, 1979). This estimator is design unbiased. Thus the estimators based on (7) are asymptotically design unbiased in the sense discussed by Sarndal. (Note that PS is also obtained by setting $\tau^2=\infty$, which corresponds to a fixed effects analysis of variance).

Note that neither M1 nor PS uses the sample weights π_j. However the weights do enter the model-based analysis if the proportions p_j are unknown. The maximum likelihood estimate of p_j is then

$$\hat{p}_j = n_j \pi_j^{-1} \Big(\sum_{k=1}^{J} n_k \pi_k^{-1} \Big). \qquad (12)$$

Substituting this expression for p_j in equations (8) to (10) gives an estimator which we denote as M2. Substituting it in (11) gives

$$HK = \sum_{j=1}^{J} n_j \pi_j^{-1} \bar{y}_j \Big/ \sum_{j=1}^{J} n_j \pi_j^{-1}, \qquad (13)$$

which is Hajek's estimator (4) rewritten in the notation of this section.

It is instructive to compare these estimators with generalized difference estimators given by (2). In the notation of this section the generalized difference estimator takes the form

$$\bar{Y}_{GD}(\hat{\mu}) = N^{-1} \sum_{j=1}^{J} n_j \pi_j^{-1} (\bar{y}_j - \hat{\mu}_j) + N^{-1} \sum_{j=1}^{J} N_j \hat{\mu}_j,$$

where $\hat{\mu}_j$ is the predicted mean for subclass j. This estimator can be rewritten in the form

$$\bar{Y}_{GD}(\hat{\mu}) = \sum_{j=1}^{J} p_j \bar{y}_j,$$

where

$$\bar{y}_j = w_j \bar{y}_j + (1-w_j)\hat{\mu}_j,$$

and

$$w_j = \pi_j^{-1} n_j / N_j.$$

Hence estimators in this class, like the model-based estimator (M1), combine a weighted average of the observed mean ($\bar{y}_j$) and a fitted mean ($\hat{\mu}_j$) in subclass j. However the weight w_j is not intuitively appealing. It has the sensible property of tending to one as the sample size n_j increases, but it can take values greater than one. Also, it is not clear why the relative weight given to $\bar{y}_j$ and $\hat{\mu}_j$ should depend on the realization of the sample design.

In particular cases $\bar{Y}_{GD}$ does yield model-based estimators. We have noted that HK is obtained by setting $\hat{\mu}_j$ = HK. The post-stratified estimator PS is obtained by setting $\hat{\mu}_j=\bar{y}_j$. However, both these estimators are extreme cases where the sample mean $\bar{y}_j$ is given weight 1. We contend that in small or moderate sample sizes where the subclass means are poorly determined, the modelling weights λ_j provide a more appealing combination of smoothed and observed data than the weights w_j in the generalized difference estimator.

3. Simulation Study.

A limited simulation study on 12 artificially generated populations was carried out to provide some numerical comparisons of the estimators described above.

3.1 Population Structure and Sample Design.

Each population had values of a variable y recorded for N=2000 units, arranged in 10 strata as in the first row of Table 1. One thousand samples were selected from each population by stratified random sampling, with probabilities of selection in each stratum given in the second row of Table 1, yielding expected sample sizes in the last row of the table. The overall expected sample size is 79.

Table 1 About Here

3.2 Generation of Population Values.

The 12 populations are labelled

1N1, 1N2, 4N1, 4N2, 1L1, 1L2, 4L1, 4L2, 1S1, 1S2, 4S1, 4S2.

The first numeral indicates the ratio of the within stratum variance (σ^2) to the between

stratum variance (τ^2) and takes the value 1 or 4. The letter indicates the distribution used to sample values (N=Normal, L=Long-tailed, S=Skewed). The second numeral indicates replicate (1 or 2). Detailed formulae for generating y_{ij}, the value of y for unit i in stratum j, are given below.

i) Normal populations (1N1, 1N2, 4N1, 4N2).

$$y_{ij} = \tau\mu_j + \sigma z_{ij},$$
$$\mu_j \sim \text{iid } G(0,1),$$
$$z_{ij} \text{ iid } G(0,1),$$

where G(0,1) is the standard Normal distribution, $\sigma^2=1$, and $\tau^2=1$ (1N1, 1N2) or 0.25 (4N1, 4N2). These populations are generated from the random effects model discussed in the previous section, and hence the model based estimators M1 and M2 of $\bar{Y}$ should perform well.

ii) Long-tailed populations (1L1, 1L2, 4L1, 4L2).

$$y_{ij} = \sqrt{3/5}(\tau\mu_j + \sigma z_{ij}),$$
$$\mu_j \sim \text{iid } t_5,$$
$$z_{ij} \sim \text{iid } t_5,$$

where t_5 denotes the t distribution on 5 degrees of freedom, the factor $\sqrt{3/5}$ is chosen so that the within stratum variance is unity, and $\tau^2=1$ (1L1, 1L2) or 0.25 (4L1, 4L2).

iii) Skewed populations (1S1, 1S2, 4S1, 4S2).

$$y_{ij} = k(\mu_j + \lambda z_{ij})^2,$$
$$\mu_j \sim \text{iid } U(4,6),$$
$$z_{ij} \sim \text{iid } G(0,1),$$

where U(4,6) denotes the uniform distribution between 4 and 6. The exponent of 2 in the expression for y_{ij} produces a skewed distribution of y-values in each stratum. The within stratum variance is not constant, but varies with the mean. Values of k and λ are chosen so that the average within stratum variance is unity, that is

$$\sigma^2 = E(\text{var}(y_{ij}|\mu_j)) = 1,$$

and the between stratum variance

$\tau^2 = \text{var}(E(y_{ij}|\mu_j)) = 1$ (1S1, 1S2) or 0.25 (4S1, 4S2).

3.3 Estimators Compared.

For each sample seven estimates of the population mean $\bar{Y}$ were calculated: PS given by equation (11), M1 given by equations (8) to (10), HK given by equation (13), M2 given by equations (8) to (10) with p_j replaced by the estimate in equation (12), SA given by equation (3), HT given by equation (1), and the unweighted sample mean YM. The mean bias and the root mean squared error for each estimator are displayed in Tables 2 and 3.

Tables 2 and 3 About Here

3.4 Summary of Simulation Results.

A fuller discussion of the results of the simulations is given in Little (1981). Like any such study, the one we have described is limited in scope. Only point estimation of the population mean is considered. Only one population structure and sampling scheme is adopted. A variety of distributions are chosen to generate the values, but exchangeability of the values in each stratum and the means across strata is assumed. Despite these limitations, a number of interesting points emerge:

1) In terms of overall mean squared error, the methods ranked as follows:

$$M1 \lesssim PS \lesssim M2 \simeq HK \lesssim SA \lesssim YM \lesssim HT.$$

2) The Horvitz-Thompson estimator is unbiased, but is not robust, with particularly disastrous mean squared error in the skewed populations. Thus design unbiasedness can incur an extreme penalty in terms of mean squared error.

3) The unweighted sample mean has low variance, but poor mean squared error when the weighted and unweighted averages of the population stratum means are divergent, leading to a bias which persists as the sample size increases. This is a typical example of a bad estimator for sample surveys.

4) Methods which use the population stratum sample sizes (M1 and PS) are markedly superior to methods which do not. Thus this information should be used if known.

5) The model-based shrinkage estimators M1 and M2 display bias, particularly for the non-normal populations. However the bias of these methods declines as the sample size increases. In terms of mean squared error these methods compare favorably with methods motivated by design unbiasedness, in both normal and non-normal populations. The value of shrinkage to the unweighted mean appears greater when the stratum proportions p_j are known (M1$\lesssim$PS) than when they are estimated (M2$\simeq$HK).

6) The two generalized difference estimators, SA and HK, perform similarly in simulations, with HK marginally superior.

4. Extensions to Include Covariates.

A fuller version of this paper (Little, 1981) outlines extensions of the modelling approach to include covariate information. Again the key aspect is to model heterogeneity of the population across subclasses indexed by the probability of selection. The chosen model for a single covariate is

$$y_{ij} \sim_{\text{ind}} G(\alpha_j + \beta_j x_{ij}, \sigma^2),$$
$$\alpha_j \sim_{\text{ind}} G(\bar{\alpha}, \tau_\alpha^2),$$
$$\beta_j \sim_{\text{ind}} G(\bar{\beta}, \tau_\beta^2),$$

where x_{ij} is the value of the covariate for unit i in subclass j, preferably centered so that the population mean of x in stratum j is zero. This model also leads to estimates of Y which are asymptotically design unbiased as the sample size increases.

5. Conclusion.

In theory the role of the sample design in the model-based approach to survey inference appears limited. The sampling distribution does not form the basis for inferences, and provided the sampling design does not lead to selection bias (is ignorable, in the sense discussed by Rubin, 1976), the sampling distribution drops out of the likelihood function. Nevertheless the sample design is extremely important to the applied modeller, who never has a true model on which to base inferences. Firstly, probability sampling is needed to ensure that the model-based analysis is not sensitive to unknown biases that cannot be detected in the observed sample. Secondly, the chosen model needs to be sensitive to aspects of the design which will lead to bias if the model is incorrectly specified. In our particular example, the fact that unequal selection probabilities are adopted in different subclasses of the population means that heterogeneity of the population between the subclasses must be adequately modelled, to avoid serious misspecification error. We believe that the study and adoption of such models will eliminate the need for compromise estimators such as those in the generalized difference class. These procedures, although ingenious, are not appealing to the strict modeller since they include estimators (such as SA) which are not efficient under a well defined model of the population.

References

Basu, D. (1971). An essay on the logical foundations of survey sampling, Part One. In Foundations of Statistical Inference, Eds. V.P. Godambe and D.A. Sprott, pp. 202-233. Toronto: Holt, Rinehart and Winsten.

Ericson, W.A. (1969). Subjective Bayesian models in sampling finite populations, 1. J. Roy. Statist. Soc., Ser. B 31, 195-234.

Holt, D. and Smith, T.M.F. (1979). Poststratification. J. Roy. Statist. Soc. Ser. A 142, 33-46.

Horvitz, D.G. and Thompson, D.J. (1952). A generalization of sampling without replacement from a finite universe. J. Am. Statist. Assoc. 47, 663-685.

Little, R.J.A. (1981). Robust model-based inference for a finite population mean from unequally weighted samples. Submitted to J. Am. Statist. Assoc. .

-------- and Rubin, D.B. (1981). Discussion of "Six Approaches to Enumerative Survey Sampling" by K.R.W. Brewer and C.E. Sarndal. To be published in the Proceedings of the Symposium on Incomplete Data. National Academy of Sciences, Washington, D.C.

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

Rubin, D.B. (1976). Inference and Missing Data. Biometrika 63, 58-92.

Sarndal, C.E. (1980). On π-inverse weighting versus best linear unbiased weighting in probability sampling. Biometrika 67, 3,639-650.

TABLE 1
STRATIFIED PROBABILITY SAMPLING SCHEME

	Stratum										
	1	2	3	4	5	6	7	8	9	10	All
Population Size	20	40	40	100	200	200	200	300	300	600	2000
= Pr (selection)	0.40	0.24	0.16	0.08	0.05	0.04	0.03	0.03	0.02	0.133	.04
expected sample size	8	9.6	6.4	8	10	8	6	9	6	8	79

TABLE 2
MEAN BIAS (x1000) FROM 1000 SAMPLES

Population	Estimator PS	M1	HK	M2	SA	HT	YM
1N1	-0.56	-7.46	-0.95	-5.03	-3.22	-4.02	-28.19(*)
1N2	-3.24	-1.11	-0.09	-3.43	0.09	1.76	-99.89(*)
4N1	4.80	8.03	-1.88	2.42	0.84	0.64	4.45
4N2	1.44	77.73(*)	8.85	75.84(*)	6.77	4.06	204.44(*)
1L1	0.51	36.24(*)	6.74	38.37(*)	4.89	0.71	205.28(4)
1L2	-3.09	-5.97	-5.57	-10.38	8.58	-6.48	-139.75(*)
4L1	9.05	41.14	3.26	37.08	2.70	0.97	124.13
4L2	-0.04	56.56(*)	0.60	53.01(*)	2.77	1.58	151.30(*)
1S1	0.69	55.61(*)	2.03	50.44(*)	4.45	-32.09	322.84(*)
1S2	1.72	-64.26(*)	-8.95	-63.43(*)	-2.75	27.96	-408.01(*)
4S1	4.72	-51.64(*)	4.59	-46.25(*)	3.81	-5.44	-136.27(*)
4S2	-2.46	23.38(*)	1.61	22.57(*)	1.42	-8.98	38.06(*)
Mean absolute bias	2.69	35.76	3.76	34.02	3.52	7.89	155.22

*Starred values were significantly different from zero at the 0.0001 level. All other other values were not significant at the .05 level.

TABLE 3
ROOT MEAN SQUARED ERRORS x(1000)
FROM 1000 SAMPLES

Population	Estimator PS	M1	HK	M2	SA	HT	YM	Row mean RMSE
1N1	-13.5	-17.0	10.8	11.2	10.3	11.6	-13.4	181.1
1N2	-22.4	-25.3	14.6	14.2	14.4	15.8	-11.3	204.3
4N1	7.3	-7.1	5.5	-3.6	5.2	4.8	-13.1	143.0
4N2	-9.9	-8.8	-4.9	1.2	-1.5	-7.6	31.5	179.7
1L1	-13.9	-17.7	-2.5	-1.1	2.4	-0.4	33.3	187.4
1L2	-16.9	-21.7	4.4	4.2	4.2	13.5	12.2	178.3
4L1	3.4	-10.1	0.8	-5.4	2.3	2.9	6.1	160.4
4L2	-2.1	-8.2	-2.1	-3.7	-1.7	4.4	13.4	174.4
1S1	-44.2	-41.1	-29.2	-25.4	-24.3	123.7	39.5	253.5
1S2	-49.6	-45.2	-32.0	-26.5	-25.0	137.4	39.8	307.3
4S1	-13.0	-22.1	-15.2	-19.7	-14.8	94.8	-11.0	204.4
4S2	-10.2	-24.9	-11.9	-18.8	-11.9	108.3	-30.5	174.9
Column Mean RMSE	159.4	150.5	181.1	180.8	185.4	295.8	216.9	
(s.d.)	(9.2)	(11.8)	(24.0)	(29.8)	(26.5)	(185.4)	(91.5)	

*Root mean squared errors (RMSEs) appear in the row and column margins multiplied by 1000. The body of the table presents percentage deviations of the RMSEs from the row mean