UNEQUAL PROBABILITY SAMPLING UNDER MEASUREMENT ERROR

Promod Chandhok, Ohio University
420 Copeland Hall, Athens, OH 45701

1. Introduction

The customary textbooks in survey sampling compare sampling strategies when measurement errors are taken into account. The usual unbiased estimators in equal and unequal probability sampling are studied in the presence of measurement errors.

2. Unequal Probability Sampling

Consider the situation in which the values $X_1, X_2, \ldots, X_N$ of the x-characteristic of the units $U_1, U_2, \ldots, U_N$ are known at the time of designing the survey. A sample of n units is selected according to probability proportional to size (pps) x with replacement, and observations are made on the y-characteristic of the selected units. We denote the sampling scheme by PPSWR. The probability of selecting the j-th unit at each trial is $P_j = X_j/\sum_j^N X_j$. Let the sample be

$$y_1, y_2, \ldots, y_n$$
$$P_1, P_2, \ldots, P_n$$

The observation y_{it} made for the character y on U_i is subject to measurement error. We shall use the fairly general model

$$y_{it} = Y_i + \mu_i + e_{it} = Y_i' + e_{it}$$

where Y_i is the true value of Y and μ_i the bias associated with unit U_i. For the errors we assume

$$E_2(e_{it}/i) = 0;\ V_2(e_{it}/i) = \sigma_{ei}^2;$$
$$COV_2(e_{it}, e_{jt}/i,j) = \rho\sigma_i\sigma_j,\ j \neq i.$$

Using the usual estimator

$$\hat{\bar{Y}}_N = (nN)^{-1}\sum_j^n (y_j/P_j)$$

for estimating the population mean $\bar{Y}_N = N^{-1}\sum^N Y_i$, we have

$$E(\hat{\bar{Y}}_N) = E(nN)^{-1}\sum_j^n (y_j/P_j)$$
$$= E_1(nN)^{-1}\sum_j^n E_2(y_j/P_j)$$
$$= E_1(nN)^{-1}\sum_j^n (Y_j'/P_j)$$
$$= N^{-1}\sum^N Y_k' = \bar{Y}_N + \bar{\mu}_N$$

Thus the bias in $\hat{\bar{Y}}_N$ is the average bias of the units in the population. Let t_j be the number of times the j-th unit appears in the sample. Then the vector $(t_1, t_2, \ldots, t_N)$ follows a multinomial distribution. Thus, $E(t_j) = nP_j$, and $COV(t_j, t_{j'}) = n P_j P_{j'}$.

To obtain the variance, we have

$$E_2(\hat{\bar{Y}}_N) = (nN)^{-1}\sum_j^n E_2(y_j/P_j) = (nN)^{-1}\sum_j^n (Y_j'/P_j)$$

$$V_1E_2(\hat{\bar{Y}}_N) = n^{-1}\sum_j^N P_j[(Y_j'/NP_j) - \bar{Y}_N']^2$$

$$V_2(\hat{\bar{Y}}_N) = V_2[(nN)^{-1}\sum^n (Y_j/P_j)]$$
$$= V_2[(nN)^{-1}\sum_j^n (y_jt_j/P_j)]$$
$$= (nN)^{-2}[\sum_j^N (t_j^2/P_j^2)\sigma_j^2 + \sum_{j\neq j'}^N (t_jt_{j'}/P_jP_{j'})\rho\sigma_j\sigma_{j'}$$

$$V(\hat{\bar{Y}}_N) = V_1E_2(\hat{\bar{Y}}_N) + E_1V_2(\hat{\bar{Y}}_N)$$
$$= n^{-1}\sum_j^N P_j[(Y_j/NP_j) - \bar{Y}_N']^2 + n^{-1}N^{-2}\sum_j^N(\sigma_j^2/P_j)$$
$$+ (n-1)n^{-1}N^{-2}\sum_j^N \sigma_j^2 + (n-1)n^{-1}N^{-2}\rho\sum_{j\neq j'}\sigma_j\sigma_{j'} \tag{1}$$

The first term on the right hand side of (1) is the sampling variance. The sum of the second and third terms is called the simple response variance, and the fourth term is called the correlated response variance. The sum of the second, third and fourth terms is consequently called "response variance".

Next we consider simple random sampling with replacement (SRSWR). This is a particular case of PPSWR when $P_j = N^{-1}$ for $j = 1,2, \ldots, N$. Thus the usual unbiased estimator of $\bar{Y}_N$ is

$$\bar{y}_n = n^{-1}\sum_{i=1}^n y_i$$

with mean

$$E(\bar{y}_n) = \bar{Y}_N + \mu_N$$

and variance

$$V(\bar{y}_n) = (nN)^{-1}\sum(Y_j - \bar{Y}_N')^2 + (nN)^{-1}(N+n-1)_{,}$$
$$N^{-1}\sum\sigma_j^2 + (n-1)n^{-1}N^{-2}\rho\sum_{j\neq j'}\sigma_j\sigma_{j'} \tag{2}$$

It can be easily shown that under simple random sampling without replacement (SRSWOR) the usual unbiased estimator

$$\bar{y}_n = n^{-1}\sum_j^n y_j$$

has

$$E(\bar{y}_n) = \bar{Y}_N + \bar{\mu}_N$$

and

$$V(\bar{y}_n) = (N-n)(nN)^{-1}(N-1)^{-1}\sum^N(Y_j'-\bar{Y}')^2$$
$$+ (nN)^{-1}\sum^N\sigma_j^2 + (n-1)(nN)^{-1}(N-1)^{-1}\sum_{j\neq j'}\rho\sigma_j\sigma_{j'} \tag{3}$$

3. Empirical Study

Various statisticians, including Cochran (1977), have called for empirical work in the area of response errors. It is in this spirit that one population has been selected and the effect of measurement errors studied under the model considered in this paper.

The population considered is taken from Kish (1965 , Appendix E). The population relates to the 270 blocks in Ward I of Fall River, Massachussets, and is taken from the column of Block Statistics of the 1950 U.S., Census. The total number of dwellings (X_i) and the number of dwellings occupied by renters (Y_i) are known for each block. The purpose is to estimate, from the sample, the average number of rented dwellings per block. We will assume that the number of dwellings occupied by renters in the i-th block, i.e., Y_i as given in Kish (1965, Appendix E), is the true value of y. We note that the correlation between X and Y is 0.96, which is typical of the populations considered for the study of the ratio estimator (Royall and Cumberland, 1981).

To study different strategies under measurement errors, response errors will be introduced in the data in the following directions:

1. the bias associated with unit U_j, i.e., μ_j, will be assumed to be at levels A_1Y_j with $A_1 = \pm0.05, \pm0.01, 0.00$;
2. the within-trial variance σ^2 will be taken as $A_2\sigma_Y^2$ with $A_2 = 0.00, 0.05, 0.10, 0.3, 1.0$;
3. the correlation coefficient ρ will be taken as $\rho = 0.00, 0.01, 0.05$.

The sample sizes to be considered are 30, 45 and 60.

The values of A_1, A_2 and ρ are chosen in view of the studies undertaken by Gray (1955) and Kish (1962).

Let

$$S_1 = (\hat{\bar{Y}}_N, \text{PPSWR}),$$

$$S_2 = (\bar{y}_n, \text{SRSWR}),$$

$$S_3 = (\bar{y}_n, \text{SRSWOR}),$$

where

$$\hat{\bar{Y}}_N = (Nn)^{-1}\sum^{n}(Y_j/P_j),$$

$$\bar{y}_n = n^{-1}\sum^{n} Y_j,$$

$$P_j = X_j/X$$

In the absence of errors, the variances of the three strategies are as given in Table 1.

Sample Size	PPS	SRSWR	SRSWOR
30	0.829	14.269	12.730
45	0.553	9.512	7.956
60	0.415	7.134	5.569

We observe that the variance of S_1 is considerably smaller than the variance of S_2 and S_3. This is to be expected, as the Y_i's are highly correlated with the X_i's. We also note that doubling the sample size halves the variance. Since the three strategies are unbiased, the mean square error (MSE) is the same as the variance.

We now retain $A_2 = 0$, $\rho = 0$ but introduce bias in reporting of rented dwelling in the block. Table 2 presents the sampling variance and MSE of the three strategies. The response variance of the three strategies is zero. We observe that the sampling variance decreases when the bias is negative and increases when the bias is positive (as compared with the situation in which there is no bias). Since the sampling variance is low in the case of PPS sampling, the percentage increase in MSE is much higher in this case, as compared with sampling with equal probabilities. We also observe that the MSE for the measurement error case may be smaller than the MSE for the no-measurement-error case. This would happen when the measurement bias is large and negative and thus the decrease in sampling variance is enough to make the MSE for the measurement error case smaller than the MSE for the no-measurement-error case.

Let us consider the case, when both A_1 and A_2 are not zero. Table 3 gives the sampling variance and MSE of the three strategies for different sample sizes, A_1, $A_2 = 0.3$ and $\rho = 0$. In this case the response variances of S_1, S_2 and S_3 are 19.379, 4.758 and 4.296 respectively. By examining Table 3, we find that strategy S_3 is more efficient than S_1 or S_2. We also studied the case $A_2 = 0.05$ and 0.01. The tables, not shown here indicate that with A_2 this small, the within-trial variance is not large enough to make the response variance of S_1 large. Hence, the behavior of MSE is similar to the case $A_2 = 0$, and PPS sampling is better than equal probability sampling.

We next consider the case when ρ is not zero. Table 4 gives the sampling variance and MSE of the three strategies for different sample sizes, A_1, $A_2 = 0.3$ and $\rho = 0.01$. The response variances of S_1, S_2 and S_3 increase to 21.625, 6.004 and 5.542 respectively. After examining Table 4, we conclude that S_3 is more efficient than S_1 or S_2, precisely what we inferred from Table 3. This is to be expected since the correlated response variance of the three strategies is the same.

4. Conclusion

The results of our study indicate that if measurement errors are absent then S_1 is more efficient than S_2 or S_3. But, if measurement errors are present, then S_3 may be more efficient than S_1 or S_2. Also, we observed that the larger the within-trial variance, the better the strategies S_2 and S_3 perform in relation to S_1.

Acknowledgements

The author wishes to thank Barbara Collins for her patient typing.

5. References

Bailar, B.A., and T. Dalenius. 1969. Estimating Response Variance Components of the U.S. Bureau of the Census Survey Model. Sankhya, Series B, 31:341-360.

Cochran, W.G. 1977. Sampling Techniques. Wiley, New York.

Dalenius, T. 1977a. Bibliography of Non-Sampling Errors in Surveys, I(A-G). International Stat. Review, 45:71-89.

Dalenius, T. 1977b. Bibliography of Non-Sampling Errors in Surveys, I(H-Q). International Stat. Review, 45:181-197.

Dalenius, T. 1977c. Bibliography of Non-Sampling Errors in Surveys, III(R-Z). International Stat. Review, 45:303-317.

Fellegi, I.P. 1964. Response Variance and its Estimation. J. Amer. Stat. Assoc. 59:1016-1041.

Gray, P.G. 1965. The Memory Factor in Social Surveys. J. Amer. Stat. Assoc., 50:344-363.

Hansen, M.H., and B.J. Tepping. 1969. Progress Problems in Survey Methods and Theory Illustrated by the Work of the United States Bureau of the Census. In N.L. Johnson and H. Smith, Jr. (eds.). New Developments in Survey Sampling. Wiley Interscience, New York.

Hansen, M.H., and J. Waksberg. 1970. Research on Non-Sampling Errors in Censuses and Surveys. Rev. Int. Stat. Inst. 38:318-32.

Hansen, M.H., W.N. Hurwitz, and M.A. Bershad. 1960. Measurement Errors in Censuses and Surveys. Bull. de Institut. International de Statistique. 38(2):359-374.

Koch, G. 1973. An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to Estimators Involving Subclass Means. J. Amer. Stat. Assoc. 58:906-913.

Pritzker, L., and R. Hanson. 1962. Measurement Errors in the 1960 Census of Population. Proc. Soc. Stat. Section Amer. Stat. Assoc. 80-89.

Raj, D. 1968. Sampling Theory. McGraw Hill, New York.

Royall, R.M., and W.G. Cumberland. 1981. An Empirical Study of the Ratio Estimator and Estimators of its Variance. J. Amer. Stat. Assoc. 76:66-77.

Sukhatme, P.W., and B.V. Sukhatme. 1970. Sampling Theory of Surveys with Applications. Iowa State University Press, Ames, Iowa.

Table 2. Variance and MSE for different sample sizes, A_1, A_2 = and ρ= 0

Sample Size	A_1	PPSWR				SRSWR				SRSWOR			
		-0.05	-0.01	+0.01	+0.05	-0.05	-0.01	+0.01	+0.05	+0.05	-0.05	-0.01	+0.01
30	Sampling variance	0.748	0.813	0.846	0.914	12.877	13.985	14.555	15.731	11.489	12.477	12.986	14.035
		(-9.77)[a]	(-1.93)	(2.05)	(10.25)	(-9.76)	(-1.99)	(2.00)	(10.25)	(-9.75)	(01.99)	(2.01)	(10.25)
	MSE	1.461	0.841	0.875	1.627	13.590	14.013	14.584	16.444	12.201	12.506	13.015	14.748
		(76.24)	(1.44)	(5.54)	(96.26)	(-4.76)	(01.79)	(2.21)	(15.24)	(04.16)	(-1.76)	(2.23)	(15.85)
60	Sampling Variance	0.374	0.406	0.423	0.457	6,439	6.992	7.278	7.866	5.026	5.459	5.681	6.140
		(-9.88)	(-2.17)	(1.93)	(10.12)	(-9.74)	(-1.99)	(2.02)	(10.26)	(-9.75)	(-1.98)	(2.01)	(10.25)
	MSE	1.087	0.435	0.452	1.170	7.151	7.021	7.306	8.578	5.739	5.487	5.710	6.853
		(161.93)	(4.82)	(8.92)	(181.93)	(0.24)	(1.58)	(2.41)	(20.24)	(3.05)	(1.47)	(2.53)	(23.06)

[a]The figures in parentheses denote the percentage increase over the case when measurement errors are absent.

Table 3. Variance and MSE for different sample sizes, A_1, A_2 = 0.3 and ρ= 0

Sample Size	A_1	PPSWR				SRSWR				SRSWOR			
		-0.05	-0.01	+0.01	+0.05	-0.05	-0.01	+0.01	+0.05	-0.05	-0.01	+0.01	+0.05
30	Sampling variance	0.748	0.813	0.846	0.914	12.877	13.985	14.555	15.731	11.489	12.477	12.986	14.035
	MSE	20.840	20.220	20.253	21.006	18.348	18.771	19.342	21.202	16.498	16.802	17.311	19.044
60	Sampling variance	0.374	0.406	0.423	0.457	6.439	6.992	7.278	7.866	5.026	5.459	5.681	6.140
	MSE	17.352	16.700	16.717	17.435	16.106	15.976	16.261	17.533	14.225	13.973	14.196	15.339

Table 4. Variance and MSE for different sample sizes, A_1, A_2 = 0.3 and ρ= 0.01

Sampling Size	A_1	PPSWR				SRSWR				SRSWOR			
		-0.05	-0.01	+0.01	+0.05	-0.05	-0.01	+0.01	+0.05	-0.05	-0.01	+0.01	+0.05
30	Sampling variance	0.748	0.813	0.846	0.914	12.877	13.985	14.555	15.731	11.489	12.477	12.986	14.035
	MSE	22.086	21.466	21.499	22.252	19.594	20.017	20.588	22.448	17.744	18.048	18.577	20.290
60	Sampling variance	0.374	0.406	0.423	0.457	6.439	6.992	7.278	7.866	5.026	5.459	5.681	6.140
	MSE	12.282	11.630	11.647	12.365	11.037	10.906	11.191	12.463	9.155	8.903	9.126	10.269