THE DOUBLE SAMPLING SCHEME AND ITS E(MSE)

Camilla A. Brooks and William D. Kalsbeek
University of North Carolina

I. Introduction

Most, if not all surveys of any scope are plagued with the problem of nonresponse and the survey statistician must decide on the method of handling it. In this paper, the Hansen and Hurwitz double sampling treatment of nonresponse is revisited. However, the probabilistic aspect of a 1949 Politz-Simmons Model and the more recent models of Platek, et. al. (1977, 1980) and Lessler (1979) is incorporated. In addition to this difference, the mean square error (MSE) of the estimate, which in this case is the sample mean, is developed in terms of "total error," that is sampling error and measurement error is incorporated, but assuming that the correlation of response deviations and sampling error is zero.

We look at the expected mean square error of the model, assuming the mean square error of the sample mean is in fact random and compare it with the E(MSE) of two other models for fixed cost--a model in which no adjustment is made for nonresponse, and a model in which a Politz-Simmons type estimator is used.

II. Model Assumptions and MSE's

The error model is assumed developed for one interviewer assignment area; specifically, the survey involves the following steps:

(1) selection of a simple random sample without replacement,

(2) solicitation of response from each sampling unit,

(3) selection of a subsample, and

(4) trial to trial repeatability of the process.

Then the estimate of the mean under the Double Sampling Model is given below:

$$\bar{x}_{t-DS}=\frac{1}{n}\sum_{i=1}^{n}\delta_i Y_{i1t}+\frac{k}{n}\sum_{i=1}^{n}(1-\delta_i)\nu_i Y_{i2t}$$

where

Y_{i1t}=response obtained from the i-th sample unit in the initial sample at trial t,

Y_{i2t}=response obtained from the i-th unit in the subsample of nonrespondents at trial t,

δ_i =1, if the i-th unit responds
0, otherwise,

ν_i =1, if the i-th sample unit is selected in the subsample of nonrespondents,
0, otherwise,

n =sample size out of N total units,

k =the inverse of the subsampling rate.

The event of responding or not responding represented by δ_i is assumed random and independent from sample unit to sample unit; that is

$$P(\delta_i=1/s)=P_i \text{ and } P(\delta_i=1,\delta_{i'}=1/s)=P_iP_{i'}.$$

Further,

$$Y_{i1t}=X_i+B_{i1}+E_{i1t} \text{ and } Y_{i2t}=X_i+B_{i2}+E_{i2t}$$

where

$$E_t(Y_{i1t})=Y_{i1}=X_i+B_{i1} \text{ and } E_t(Y_{i2t})=Y_{i2}=X_i+B_{i2}.$$

B_{i1} and B_{i2} are the constant error terms or biases associated with the response obtained from unit i in the initial survey and the subsample of nonrespondents, respectively; the terms E_{i1t} and E_{i2t} are the corresponding random error terms at trial t. The biases B_{i1} and B_{i2} are assumed to be different due to inherent differences in response patterns when a unit responds in one sample as opposed to the other, differences in the methods of soliciting response and any interaction of the two.

The components of the mean square error for this model are given below. For details on the development of the MSE model, the reader is referred to reference 1.

SQBIAS = Square of the bias

Response Variance (RV)

SRV =simple response variance, which is the sum of the response variance of units responding in the initial sample and units responding in the subsample,

CRV =correlated response variance which is the sum of the correlated response variance of units responding in the initial sample and units responding in the subsample,

CRV_{12}=correlation between unit i in the initial sample and unit i' in the subsample, i≠i'.

Nonresponse Variance (NRV)

=nonresponse variance which is due to the difference in the response bias of a unit responding in the initial sample and that unit responding in the subsample of nonrespondents.

Sampling Variance (SV)

SSV=variance due to the subsample of nonrespondents,

PVAR=sampling variance due to the population.

The simplified No-Adjustment and Politz-Simmons Models are developed under similar assumptions with the exception that the survey step involving the selection of a subsample of nonrespondents is not applicable. The estimate of the mean for the No-Adjustment Model is

$$\bar{x}_{t-NA}=\frac{1}{n_1}\sum_{i=1}^{n}\delta_i Y_{i1t}.$$

For the Politz-Simmons Model, it is

$$\bar{x}_{t-PS}=\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i Y_{i1t}}{P_i}.$$

In both models the variables are defined as they are for the Double Sampling Model. The mean square error components consist of the square of the bias, the response variance, both simple and correlated, the nonresponse variance, and the sampling variance.

Again details on both the derivation and components of the MSE are given in Brooks (1982).

The No-Adjustment Model is considered simplified in that n_1 -- the number of respondents -- which is in fact a random variable in this model, is assumed fixed, resulting in a substantially biased estimate, probably downward, of the MSE for small sample sizes. For the Politz-Simmons Model, the response probabilities are assumed known when in fact they will have to be estimated. Thus, the variance of the Politz-Simmons estimator is underestimated.

III. The Superpopulation Model Assumptions and Selected Parameter Values

The expected mean square errors of the three models are studied using the superpopulation model approach, that is, it is hypothesized that the finite population is drawn at random from a larger universe or infinite population. In this case it assumed that the infinite population to which the Y_i's or responses, belong is normally distributed and that to which the response probabilities belong is distributed according to the beta distribution.

We assume that every unit in the population has response probabilities distributed according to the same beta distribution, so that the expectation of P can be considered the overall expected response rate. The parameters α and β are the shape parameters and their values indicate how the response probabilities are distributed in the population. For example, if $\alpha<\beta$ then the response probabilities are skewed to the right with more of the population having low response probabilities; if $\alpha>\beta$ the opposite holds, and if $\alpha=\beta$, then the distribution of response probabilities is symmetric about 0.5. It is assumed that the observed values and the true values are distributed normally each with the same variance but differing means and that the mean of the responses from the initial sample differs from the mean of the subsample because of different expected response bias only.

It is also hypothesized that the response probabilities are not correlated with either the random or constant response error terms, but that in general, the response probabilities and the observed values are correlated; it is through this correlation that nonresponse bias is realized. In order to obtain some estimate of the correlation of functions of Y and P, then, it was hypothesized that the conditional distribution of Y given P is normally distributed with constant variance σ^2 and mean dependent on P; the conditional mean is approximated by a polynomial function of P, in this case only up to the linear term.

The expected mean square error of the three models is studied using certain selected values of the parameters. Since the choice of values can greatly affect the conclusions, an effort was made to select what were considered moderate values. The selected value for the expected true mean was 100 while those for the expected mean of responses in the initial sample and in the subsample are 110 and 106, respectively. These values were selected such that the relative expected biases of the initial sample responses and subsample responses are 10 and 6 percent, respectively, under the assumption that the improved procedure used in obtaining interviews from the subsample of nonrespondents would produce lower response biases. The sampling variance was chosen to be such that the coefficient of variation for a sample size of 200 would be approximately 15-16 percent and the ratio of the simple and correlated response variance to sampling would be 15 and 25 percent, respectively. The correlation between P and Y is small, approximately 1 to 4 percent, depending on E(P).

IV. Results and Implications to Survey Research

Table 1 shows the effect of shape of the beta distribution on the E(MSE) and its components for three different values of E(P) --0.2, 0.5, and 0.8, representing distributions skewed to the right, symmetric, and skewed to the left. Only two terms in the E(MSE) were affected more than negligibly --the expected nonresponse variance and the expected population variance. Within each of the groupings of E(P) the sum of the percent of the population variance and the nonresponse variance are of the total expected mean square error remains constant but increases as E(P) increases. The nonresponse variance increases as α and β increase by the same amount that the population variance decreases. Also, as the expected response probability increases the percent the expected response variance and the percent the expected squared bias is of the expected mean square error decrease. However, the expected response variance itself remains constant; also the increasing expected squared bias and the percent it is of the E(MSE) is due to the choice of response bias terms. Had the assumption been that the expected response bias from units in the subsample was larger than that of the initial sample instead of vice versa, then the squared bias would be a decreasing function of E(P).

Tables 2 and 3 present results from the No-Adjustment and Politz-Simmons Models. The following can be noted: the expected mean square error of the No-Adjustment Model decreases with increasing expectation of P, difference from that of the double sampling model; however, this is due to the correlation between P and Y which greatly affects this model through the expected squared bias while the correlation of P and Y has negligible effect on the E(MSE) of the Double Sampling Model and its components. When the correlation of P and Y is assumed zero, then the squared bias term of the No-Adjustment Model, is due to response bias only, and the E(MSE) is unaffected by differing shapes of the beta distribution. This, however, is not shown in the tables. In the Politz-Simmons model, the E(MSE) is greatly affected by changing α and β with a dramatic drop in E(MSE) between α close to 1 and α=2. For E(P)=0.2 and 0.5, the major contributor to the E(MSE) is the expected nonresponse variance while for E(P)=0.8, the largest contributor is the expected sampling variance.

Tables 4 and 5 compare the E(MSE's) for the three models for fixed cost. For the Double Sampling Model the optimum n and inverse of the subsampling rate k were derived using the Hansen and Hurwitz cost model and the usual principle of Lagrangian multipliers considering total expected variance, while for the other two models which involve no subsampling, n is determined from cost and E(P) only. c_1 is the cost of obtaining and processing responses from the initial sample, c_2 the cost of soliciting responses from nonrespondents, and c_3 is cost of obtaining and processing responses from subsample units. In making comparisons it should be kept in mind that the E(MSE) of the No-Adjustment Model and the Politz-Simmons Model are understated.

Table 4 shows that based on the size of the E(MSE) the Double Sampling Model is preferred in most cases. However, as α and β increase, the E(MSE) of the other two models decreases while that of the Double Sampling Model more or less remains the same. For E(P)=0.2 α and β would have to be very large and thus the variance of P very small, for the E(MSE) of the Double Sampling Model to be the larger of the three models. However, for E(P)=0.5 and 0.8, the E(MSE) of the Politz-Simmons Model is approximately the same as that of the Double Sampling Model for α=8 and above.

Table 5 with different cost ratios shows different relationships among the models when the cost of obtaining and processing sampling units in the initial sample is only 20 percent as high as for the subsample of nonrespondents. Then the other two models compare more favorably to the Double Sampling Model. For E(P)=0.2 the E(MSE) of the No-Adjustment Model is approximately the same as that of the Double Sampling Model for $\alpha \geq 2$; as E(P) gets larger and the differential between the size of n for the two models closes, then the differential in E(MSE) widens, with that of the Double Sampling Model having the lower E(MSE). In comparing the expected mean square error of the Politz-Simmons type estimator with that of the double sampling estimator, for all value of E(P), the Politz-Simmons generally has E(MSE)'s comparable to those of the Double Sampling Model.

In summary, then, the E(MSE) of the three models are all affected by the E(P); each gets smaller as E(P) increases. In all three cases this was due to an overall decrease in expected variance which offset the increase in expected squared bias. The effect the components had on the E(MSE) also differed in some cases. Now the shape of the beta distribution, or in other words, the size of α and β affected the Politz-Simmons and No-Adjustment Models, but the Double Sampling Model only negligibly. The effect on the No-Adjustment Model was through the expected squared bias. Thus in designing a survey we find that it is important to have some idea as to the distribution of response probabilities in the population as well as the overall response rate; for, if the response probabilities could be reasonably estimated, for unimodal distributions with relatively large α's and β's, or in other words, when the response probabilities are concentrated around E(P), then a Politz-Simmons type estimator might be more appropriate. Also, when the cost of taking a subsample is considerably higher than the cost of taking the initial sample, then for small E(P) and large α and β the No-Adjustment Model may be preferred over double sampling; however, the correlation between P and Y is crucial to this decision. Thus, further research is needed to look at the effect of the correlation of functions of P and Y on the E(MSE). For other than a positive linear relationship between P and Y, the results may have been very different. In addition, a look at the No-Adjustment Model and a Politz-Simmons type model without the simplifying assumptions is in order.

REFERENCES

Brooks, C.A., "Probabilistic Survey Error Models to Correct for Nonresponse," Dissertation, University of North Carolina at Chapel Hill, 1982.

Hansen, M.H. and Hurwitz, W.H., "The Problem of Nonresponse in Sample Surveys," Journal of the American Statistical Association, 41(1946) pp. 517-529.

Johnson, N.L. and Kotz, S., Continuous Univariate Distributions-2, Boston: Houghton Mifflin Company, 1970.

Lessler, J.T., "An Expanded Survey Error Model," Paper presented at the Symposium on Incomplete Data, August 11-19, 1979, Washington, D.C.

Platek, R. and Gray, G.B., "Imputation Methodology: Total Survey Error," Unpublished manuscript, 1980.

Platek, R., Singh, M.P., and Tremblay, V., "Adjustment for Nonresponse in Surveys," Survey Methodology, Vol. 3, No. 1, 1977, pp. 1-24.

Politz, A.N. and Simmons, W.R., "An Attempt To Get the 'Not-at-Homes' into the Sample Without the Callbacks," Journal of the American Statistical Association, 44, 1949, pp. 9-31.

Table 1: E(MSE) Components as Percent of Total E(MSE) for Varying Shapes of the Beta Distribution—Double Sampling Model (n = 200)

			Components as Percent of Total E(MSE)							TOTAL
E(P)	α	β	E(RV)	E(NRV)	E(SSV)	E(PVAR)	E(SV)	E(VAR)	E(SQBIAS)	E(MSE)
.2	$\frac{1}{8}$	$\frac{1}{2}$	16.8	2.6	33.5	39.4	92.2	92.2	7.8	596
	$\frac{1}{2}$	2	16.8	4.8	33.5	37.1	92.2	92.2	7.8	596
	2	8	16.8	6.1	33.5	35.8	92.2	92.2	7.8	596
	8	32	16.8	6.5	33.5	35.4	92.2	92.2	7.8	596
.5	$\frac{1}{8}$	$\frac{1}{8}$	18.6	2.3	23.1	44.1	88.1	88.1	11.9	539
	$\frac{1}{2}$	$\frac{1}{2}$	18.6	5.8	23.1	40.6	88.1	88.1	11.9	539
	2	2	18.6	9.3	23.2	37.1	88.1	88.1	11.9	539
	8	8	18.6	10.9	23.2	35.5	88.1	88.1	11.9	539
.8	$\frac{1}{2}$	$\frac{1}{8}$	20.7	3.2	10.0	48.6	58.6	82.5	17.5	483
	2	$\frac{1}{2}$	20.7	5.9	10.0	45.9	55.9	82.5	17.5	483
	8	2	20.7	7.5	10.1	44.2	54.3	82.5	17.5	483
	32	8	20.7	8.1	10.1	43.7	53.7	82.5	17.5	483

Table 2: E(MSE) Components as Percent of Total E(MSE) for Varying Shapes of the Beta Distribution-No-Adjustment Model (n = 200)

			Components as Percent of Total E(MSE)					TOTAL
E(P)	α	β	E(RV)	E(NRV)	E(SV)	E(VAR)	E(SQBIAS)	E(MSE)
.2	$\frac{1}{8}$	$\frac{1}{2}$	12.1	23.6	50.4	86.1	13.9	2060
	$\frac{1}{2}$	2	12.7	45.7	31.2	89.6	10.4	1965
	2	8	13.0	59.5	19.0	91.5	8.5	1916
	8	32	13.2	64.2	14.8	92.2	7.8	1900
.5	$\frac{1}{8}$	$\frac{1}{8}$	12.9	5.9	47.5	66.2	33.8	1069
	$\frac{1}{2}$	$\frac{1}{2}$	13.6	15.6	40.6	69.8	30.2	1012
	2	2	14.3	26.3	32.7	73.3	26.6	960
	8	8	14.7	31.7	28.7	75.1	24.9	938
.8	$\frac{1}{2}$	$\frac{1}{8}$	13.5	3.8	37.0	54.3	45.7	809
	2	$\frac{1}{2}$	13.9	7.2	34.8	56.0	44.0	784
	8	2	14.2	9.4	33.4	57.1	42.9	770
	32	8	14.3	10.2	32.9	57.4	42.6	765

Table 3: E(MSE) Components as Percent of Total E(MSE) for Varying Shapes of the Beta Distribution-Politz-Simmons Model (n = 200)

			Components as Percent of Total E(MSE)					TOTAL
E(P)	α	β	E(RV)	E(NRV)	E(SV)	E(VAR)	E(SQBIAS)	E(MSE)
.2	1.05	4.2	10.9	87.9	0.8	99.7	0.3	29,693
	2	8	12.3	76.9	7.7	96.9	3.1	3,245
	4	16	12.9	71.9	10.8	95.7	4.3	2,317
	8	32	13.2	69.7	12.2	95.1	4.9	2,052
.5	1.05	1.05	11.4	84.1	3.2	98.7	1.3	7,769
	2	2	15.1	54.6	21.6	91.4	8.6	1,157
	4	4	16.2	46.0	27.0	89.2	10.8	925
	8	8	16.6	42.6	29.1	88.4	11.6	859
0.8	4.2	1.05	19.5	19.7	43.4	82.6	17.4	576
	8	2	19.7	17.9	44.6	82.2	17.8	561
	16	4	19.8	17.0	45.1	82.0	18.0	554
	32	8	19.9	16.6	45.4	81.9	18.1	551

Table 4: Comparison of E(MSE) of Three Models for Fixed Cost $C-c_0 = 5000$, $c_3 = 50$, $c_1/c_3 = .5$, $c_2/c_3 = .1$

			Double Sampling			No-Adjustment		Politz-Simmons	
E(P)	α	β	n	k	E(MSE)	n	E(MSE)	n	E(MSE)
0.2	$\frac{1}{2}$	2	106	1.05	672	555	879	555	*
	2	8	106	1.05	672	555	834	555	1,273
	8	32	106	1.05	672	555	820	555	843
	32	128	106	1.05	672	555	816	555	788
0.5	$\frac{1}{2}$	$\frac{1}{2}$	145	1.29	574	333	756	333	*
	2	2	145	1.29	574	333	704	333	760
	8	8	145	1.29	574	333	682	333	581
	32	32	145	1.29	574	333	675	333	558
0.8	2	$\frac{1}{2}$	177	1.40	494	238	724	238	560
	8	2	177	1.40	494	238	710	238	497
	32	8	177	1.40	494	238	705	238	489
	128	32	177	1.40	494	238	704	238	487

*Contains negative E(MSE) components.

Table 5: Comparison of E(MSE) of Three Models for Fixed Cost $C-c_0 = 5000$, $c_3 = 50$, $c_1/c_3 = .2$, $c_1/c_3 = .1$

			Double Sampling			No-Adjustment		Politz-Simmons	
E(P)	α	β	n	k	E(MSE)	n	E(MSE)	n	E(MSE)
0.2	$\frac{1}{2}$	2	135	1.29	621	833	674	833	*
	2	8	135	1.29	621	833	631	833	903
	8	32	135	1.29	621	833	617	833	616
	32	128	134	1.29	621	833	613	833	579
0.5	$\frac{1}{2}$	$\frac{1}{2}$	236	1.83	458	666	562	666	*
	2	2	236	1.83	458	666	511	666	461
	8	8	236	1.83	458	666	489	666	372
	32	32	236	1.83	458	666	482	666	360
0.8	2	$\frac{1}{2}$	366	2.14	335	555	543	555	333
	8	2	366	2.14	335	555	529	555	306
	32	8	366	2.14	335	555	524	555	303
	128	32	366	2.14	335	555	523	555	302

*Contains negative E(MSE) components.