ESTIMATION PROCEDURES IN THE PRESENCE OF NONRESPONSE

J. E. Jackson -- Eastman Kodak Company
P.S.R.S. Rao -- University of Rochester

1. INTRODUCTION

The problem which we wish to address concerns estimation in the presence of survey nonresponse. We consider the situation in which the initial survey is conducted through the mail, and the subsampling of nonrespondents is done through the mail or by telephone. It is recognized that after the second attempt or wave, there may still be some nonrespondents. Further, the population may contain hardcore refusals. We consider various estimation procedures for the population mean of a characteristic, using the results of two waves and making appropriate assumptions regarding the remaining nonrespondents and the refusals.

2. MAIL - MAIL

We will begin by discussing the situation where a mail survey is followed by a second mailing to a subsample of the nonrespondents. This assumes that the nonrespondents can be identified in such a way that this follow-up may be carried out and that this identification procedure will not introduce a bias into the results due to nonresponse. Further, it is assumed that all questionnaires are delivered.

The basic framework within which we shall work consists of a population of N units. Of these, N_1 units would respond to a mail questionnaire if contacted and the mean of the response to a specific question will be μ_1. An additional N_2 units would respond to a follow-up, and their mean will be μ_2. Beyond that, there are N_3 units with mean μ_3 who would answer on subsequent follow-ups were we to make them. Finally, there are $N_4=N-N_1-N_2-N_3$ units who are hardcore refusals. We assume that the answers from these refusals, if obtainable, would be in the proportion as the $N_1+N_2+N_3$ who would respond, i.e.: $N_iN_4/(N_1+N_2+N_3)$ have mean μ_i ($i = 1,2,3$). Other assumptions regarding the refusals may also be appropriate in practice.

The overall mean for the population is:

$$(2.1) \qquad \mu = \frac{N_1\mu_1 + N_2\mu_2 + N_3\mu_3 + N_4\mu_4}{N}$$

which, because of the assumption above regarding refusals, can be simplified to:

$$(2.2) \qquad \mu = \frac{N_1\mu_1 + N_2\mu_2 + N_3\mu_3}{N_1 + N_2 + N_3}$$

since $\mu_4 = \mu$.

For most of the cases we will be discussing, we shall assume that the proportion of refusals is a known quantity, P. Therefore, $N_4 = PN$ and since $N_1+N_2+N_3 = N-N_4$, (2.2) can be rewritten as:

$$(2.3) \qquad \mu = \frac{N_1\mu_1 + N_2\mu_2 + N_3\mu_3}{N(1-P)}$$

The sampling procedure is as follows: Initially, mail the questionnaire to a random sample of size n. From these n questionnaires, n_1 responses are obtained with mean $\bar{y}_1$. From the remaining $n-n_1$ units, a follow-up second mailing is sent to $r \leq n-n_1$ units from which r_1 responses are obtained with mean $\bar{y}_2$. From this information, we wish to obtain an estimate of μ which will entail estimates of N_1, N_2 and N_3 and their corresponding means.

Our estimates for the sizes of these three groups are:

$$(2.4) \qquad \hat{N}_1 = n_1N/n$$
$$\hat{N}_2 = (n-n_1)(r_1/r)N/n$$
and $$\hat{N}_3 = [(n-n_1)(1-r_1/r)-Pn]N/n$$

The estimator for μ in (2.2) is:

$$(2.5) \qquad \bar{y} = \frac{n_1\bar{y}_1+(n-n_1)(r_1/r)\bar{y}_2+[(n-n_1)(1-r_1/r)-Pn]\bar{y}_3}{n(1-P)}$$

where $\bar{y}_i$ is an estimate of μ_i. However, it is not possible to obtain $\bar{y}_3$ without a third wave. The following types of estimators are obtained with different assumptions regarding the sizes of the third and fourth groups and the value of μ_3.

(a) Assume that $N_3=0$ and $P=0$. This amounts to saying that all the questionnaires mailed on the second wave will be returned. Thus, $r_1 = r$ and hence:

$$(2.6) \qquad \bar{y}_a = \frac{n_1\bar{y}_1 + (n-n_1)\bar{y}_2}{n}$$

which is the Hansen-Hurwitz estimator (1). Clearly,

$$(2.7) \qquad E(\bar{y}_a) = \frac{N_1\mu_1 + (N-N_1)\mu_2}{N} = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}$$

(b) Assume that $\mu_2=\mu_3$. This implies that, apart from refusals, those who would not respond to the second wave are the same as those who would. With this assumption, the estimator in (2.5) becomes:

$$(2.8) \qquad \bar{y}_b = \frac{n_1\bar{y}_1 + [n(1-P) - n_1]\bar{y}_2}{n(1-P)}$$

which has expectation:

$$(2.9) \qquad E(\bar{y}_b) = \frac{N_1\mu_1 + [N(1-P) - N_1]\mu_2}{N(1-P)}$$

Another estimator that might be considered is:

$$(2.10) \qquad \bar{y}_{b'} = \frac{n_1\bar{y}_1 + (n-n_1)(r_1/r)\bar{y}_2}{n_1 + (n-n_1)(r_1/r)}$$

but it ignores any information regarding μ_3 and P. An approximated expression to the expectation of this estimator is:

$$(2.11) \qquad E(\bar{y}_{b'}) = \frac{N_1\mu_1 + N_2\mu_2}{N_1 + N_2}$$

the same as for the Hansen-Hurwitz estimator.

(c) Assume that $\mu_3=0$. This implies that, apart from refusals, those who did not respond to the second wave have a mean of zero and did not respond for that reason. Thus, setting $\bar{y}_3=0$ in (2.5) we obtain:

$$(2.12) \qquad \bar{y}_c = \frac{n_1\bar{y}_1 + (n-n_1)(r_1/r)\bar{y}_2}{n(1-P)}$$

with expectation:

$$(2.14)\quad E(\bar{y}_c) = \frac{N_1\mu_1 + N_2\mu_2}{N(1-P)}$$

(d) Assume that $N_3=0$. This is the same situation as (a) except that it allows for refusals. This turns out to have the same estimator as $\bar{y}_c$.

(e) This model, as well as the following one, assumes that μ_1, μ_2 and μ_3 form some sort of logical decreasing (or possibly ascending) sequence. Hendrick's "resistance functions" approach (2) is employed to obtain an estimate of μ_3 by extrapolation. The details are given in the appendix. From (A.9),

$$(2.14)\quad \bar{y}_e = \bar{y}_1 + \frac{[n(1-P) - n_1]\ (\bar{y}_2-\bar{y}_1)}{n_1 + (n-n_1)(r_1/r)}$$

which has an expectation approximately equal to:

$$(2.15)\quad E(\bar{y}_e) \doteq \mu_1 + \frac{[N(1-P) - N_1]\ (\mu_2-\mu_1)}{N_1 + N_2}$$

(f) The other version of this method does not involve N_4 at all and is the only one which does not require an assumption about refusals. This model assumes that the proportion that would respond to each wave drops off geometrically, i.e., $N_1/N_2 = N_2/N_3$. From (A.11),

$$(2.16)\quad \bar{y}_f = \bar{y}_1 + \frac{[(n-n_1)r_1]\ (\bar{y}_2-\bar{y}_1)}{n_1 r}$$

and approximately:

$$(2.17)\quad E(\bar{y}_f) \doteq \mu_1 + \frac{N_2(\mu_2-\mu_1)}{N_1}$$

$$= \frac{(N_1-N_2)\mu_1 + N_2\mu_2}{N_1}$$

3. NUMERICAL EXAMPLES

The examples to follow are intended to demonstrate the bias exhibited by the models presented in Section 2 under various conditions where μ_3 cannot be estimated directly. These examples are displayed in Table I, which contains the decomposition of the population of N=10,000 into N_1, N_2, N_3 and N_4, the means associated with the first three, the overall mean (2.3) and the expectations of the various estimators we have discussed (2.9, 2.11, 2.13, 2.15 and 2.17).

In Example 1, all of the original nonrespondents would respond to the recall. This case is suitable for the Hansen-Hurwitz procedure. For this example, all the estimates are unbiased, but in Example 12, which is similar with different values N_1, N_2 and μ_2, $E(\bar{y}_f)$ is different from the others.

Example 2 is the same as Example 1 except for the inclusion of some refusals (Model (d)) and again all the estimators are unbiased exect $\bar{y}_f$.

Examples 3 and 4 are designed to conform to Model (b). Only $\bar{y}_b$ is unbiased. $\bar{y}_{b'}$ overestimates μ as it does for most of the remaining examples. The remaining three expectations are low.

Examples 5 and 6 are designed to conform to Model (c). Only $\bar{y}_c$ is unbiased; the others, except $\bar{y}_f$ overestimate μ.

Examples 7 and 8 conform to Model (e) while Examples 9 and 10 conform to Model (f). In general, $\bar{y}_b$ and $\bar{y}_{b'}$ will overestimate μ and $\bar{y}_c$ will underestimate it. As expected, the bias of $\bar{y}_e$ or $\bar{y}_f$ would depend on the presumed model for extrapolation.

Examples 11 and 13 conform to none of these models and are included for comparison. In particular, the means for Example 13 are in ascending order; this situation is suitable only for Models (e) and (f).

The purpose of these examples is to demonstrate the nature and magnitude of the biases which can occur using these various estimation procedures. Obviously, no one model can handle all situations, nor would any other estimators that might be derived. One should keep this in mind and pay particular attention to the assumptions that must be made when only a single recall wave is employed.

4. MAIL - TELEPHONE

An increasingly popular procedure is to use telephone calls for the follow-up phase of a mail survey. It has the advantage of a higher response rate from those who are contacted, but has the disadvantage of the noncontact problem and the possible biases associated with it. For the mail-telephone case, the composition of the population of size N will be defined differently. N_1 units would still respond to the first wave of a mail survey and they will have mean μ_1. N_2 units would be contacted in a telephone follow-up and would respond with mean μ_2. N_4 units would be contacted in a telephone follow-up and would refuse to respond. N_3 units would not be at their telephones at the times the call would be made; denote their mean by μ_3.

We shall assume that all of the population have working telephones. Further, we assume that the refusals obtained, N_4, are composed of two groups. $N_1N_4/(N_1+N_2)$ will have the same mean, μ_1, as those who answered the mail survey while the remainder, $N_2N_4/(N_1+N_2)$ will have the same mean, μ_2, as those who cooperated with the telephone follow-up. We shall assume the same refusal rate, $P^* = N_4/(N_1+N_2+N_4)$ for the N_3 units which were not contacted. If they had been contacted, N_3P^* units would have refused and $N_3(1-P^*)$ units would have responded. The population mean would then be:

$$\mu = \frac{1}{N}\left\{\left[N_1 + \frac{N_1N_4}{N_1+N_2}\right]\mu_1 + \left[N_2 + \frac{N_2N_4}{N_1+N_2}\right]\mu_2 + N_3\mu_3\right\}$$

The procedure is as follows: Take an initial mail sample of size n from which n_1 responses will be obtained with mean $\bar{y}_1$. From the $n-n_1$ nonresponses, make $r \le (n-n_1)$ telephone calls. Of these r_1 will respond with mean $\bar{y}_2$, r_f will refuse and $(r-r_1-r_f)$ will not be contacted. The sample estimate of (4.1) is:

$$(4.2)\quad \bar{y} = \frac{1}{n}\left\{\left[n_1 + \frac{n_1(n-n_1)(r_f/r)}{n_1+(n-n_1)(r_1/r)}\right]\bar{y}_1 + \left[(n-n_1)(r_1/r) + \frac{(n-n_1)^2(r_1/r)(r_f/r)}{n_1+(n-n_1)(r_1/r)}\right]\bar{y}_2 + \left[(n-n_1)\left[\frac{r-r_1-r_f}{r}\right]\bar{y}_3\right\}$$

Note that $\bar{y}_3$ represents the portion of the sample which neither responded to the mail questionnaire nor were contacted by telephone. If one were to assume that these units were no different than the units which did not respond when contacted by telephone ($\mu_2=\mu_3$), the solution would be:

(4.3)

$$\bar{y} = \frac{1}{n}\left\{n_1[n_1(r-r_1-r_f)+n(r_1+r_f)]\bar{y}_1+(n-n_1)[(n-n_1)r_1+ n_1(r-r_f)]\bar{y}_2\right\}/\left\{n_1 r + (n-n_1)r_1\right\}$$

If this assumption is not valid, then this problem cannot be solved without additional information such as could be obtained from a second wave of telephone calls aimed at the noncontacts. These noncontacts may represent a different life style, buying pattern, etc. and hence one cannot assume that $\mu_3=0$ or that μ_3 may be related to μ_1 and μ_2 by some sort of regression approach as was done in some of the mail - mail models in Section 2. This would imply that if a one-wave telephone follow-up is employed, it must be carried out with such thoroughness that the remaining noncontacts may safely be assumed to be the same as the telephone contacts.

REFERENCES

1. Hansen, M. H. and Hurwitz, W. N. (1946). "On the Theory of Nonresponse in Sample Surveys", J. Amer. Stat. Assn. 41, 517-529.

2. Hendricks, W. A. (1949). "Adjustment for Bias by Nonresponse in Mailed Surveys", Agr. Econ. Res. 1, 52-56.

APPENDIX

EXTRAPOLATION METHODS FOR MAIL - MAIL SURVEYS

The estimation models (e) and (f) in Section (2) on mail - mail surveys use an extrapolation procedure to estimate μ_3, the mean of the portion of the population that would have responded to a third wave if one had been conducted. As in that section, a mail sample of size n is taken; n_1 respond. If another questionnaire is sent to the $n-n_1$ nonrespondents, n_2 would respond; n_3 more would respond to the third wave, if there were one, and finally $n_4 = n - n_1 - n_2 - n_3$ would refuse altogether. What we wish to do in this appendix is estimate n_3 and from this μ_3 based on the assumption that μ_1, μ_2 and μ_3 form some logical sequence. Estimates of μ_1 and μ_2 would have been obtained from the first and second wave. We continue the assumption that the refusals also have means μ_1, μ_2 and μ_3 in the same proportion as those who respond or would have on the third wave.

Although the main body of this paper allows for subsampling on the recall, the development of this appendix makes use of the entire sample. The number of initial respondents, n_1, is given. The size of the potential second wave is $\hat{n}_2 = (n-n_1)(r_1/r)$. $\hat{n}_4$ is a function of the assumption of the particular model employed and we will now proceed to obtain $\hat{n}_3$.

First define:

$$(A.1)\quad x_{[1]} = \frac{n_1}{n_1 + \hat{n}_2 + \hat{n}_3}$$

$$(A.2)\quad x_{[2]} = \frac{n_1 + \hat{n}_2}{n_1 + \hat{n}_2 + \hat{n}_3}$$

$$(A.3)\quad x_{[3]} = \frac{n_1 + \hat{n}_2 + \hat{n}_3}{n_1 + \hat{n}_2 + \hat{n}_3} = 1$$

$$(A.4)\quad \bar{y}_{[1]} = \bar{y}_1$$

$$(A.5)\quad \bar{y}_{[2]} = \frac{n_1\bar{y}_1 + \hat{n}_2\bar{y}_2}{n_1 + \hat{n}_2}$$

$$(A.6)\quad \bar{y}_{[3]} = \frac{n_1\bar{y}_1 + \hat{n}_2\bar{y}_2 + \hat{n}_3\bar{y}_3}{n_1 + \hat{n}_2 + \hat{n}_3}$$

where $\bar{y}_2$ is the mean of the second wave and $\bar{y}_3$ is the estimated mean of the third wave if one had been carried out. If we assume that $x_{[i]}$ and $y_{[i]}$ are linearly related, then the final estimate of μ is:

$$\bar{y} = \bar{y}_{[3]} = a + bx_{[3]} = a + b$$

since $x_{[3]} = 1$. b could be expressed as:

$$b = \frac{\bar{y}_{[2]} - \bar{y}_{[1]}}{x_{[2]} - x_{[1]}}$$

and hence, with a little algebra we obtain:

$$(A.7)\quad \bar{y} = \bar{y}_{[3]} = \bar{y}_{[1]} - \left[\frac{\bar{y}_{[2]} - \bar{y}_{[1]}}{x_{[2]} - x_{[1]}}\right]x_{[1]} + \frac{\bar{y}_{[2]} - \bar{y}_{[1]}}{x_{[2]} - x_{[1]}}$$

$$= \bar{y}_1 + \frac{(\hat{n}_2 + \hat{n}_3)(\bar{y}_2 - \bar{y}_1)}{n_1 + \hat{n}_2}$$

If the value of $\bar{y}_3$ is of interest, by substituting (A.7) in (A.6), it is given by:

$$(A.8)\quad \bar{y}_3 = \frac{\bar{y}_{[3]}(n_1 + \hat{n}_2 + \hat{n}_3) - n_1\bar{y}_1 - \hat{n}_2\bar{y}_2}{\hat{n}_3}$$

Examples:

Model (e) in Section 2 assumes a constant refusal rate P based on experience. This makes $\hat{n}_4 = Pn$. Then $\hat{n}_3 = n - n_1 - \hat{n}_2 - \hat{n}_4$. Substituting $\hat{n}_3$ in (A.7),

$$(A.9)\quad \bar{y}_e = \bar{y}_1 + \frac{(n-n_1-\hat{n}_4)(\bar{y}_2-\bar{y}_1)}{n_1 + \hat{n}_2}$$

$$= \bar{y}_1 + \frac{[n(1-P) - n_1](\bar{y}_2-\bar{y}_1)}{n_1 + (n-n_1)(r_1/r)}$$

Model (f) assumes that $n_1/\hat{n}_2 = \hat{n}_2/\hat{n}_3$. Then $\hat{n}_3 = \hat{n}_2^2/n_1$ and no assumptions need be made directly about $\hat{n}_4$. Substituting in (A.7),

$$(A.10)\quad \bar{y}_f = \bar{y}_1 + (\hat{n}_2/n_1)(\bar{y}_2-\bar{y}_1)$$

$$= \bar{y}_1 + \frac{[(n-n_1)r_1](\bar{y}_2-\bar{y}_1)}{n_1 r}$$

There is no guarantee that these models would work well in practice but if either of the above assumptions are valid, models (e) or (f) would be appropriate. If some other functional relationship among μ_1, μ_2 and μ_3 exists which may be linearized, then (A.7) may be suitably modified.

TABLE 1

Examples For Mail -- Mail Models

	N_1	N_2	N_3	N_4	P	μ_1	μ_2	μ_3	μ	$E(\bar{y}_b)$	$E(\bar{y}_{b'})$	$E(\bar{y}_c)$	$E(\bar{y}_e)$	$E(\bar{y}_f)$
1	5000	5000	0	0	0	100	50	--	75	75	75	75	75	--
2	5000	5000	0	1000	.10	100	50	--	77.78	77.78	77.78	77.78	77.78	60
3	5000	3000	2000	0	0	100	50	50	75	75	81.25	65	68.75	70
4	5000	3000	1000	1000	.10	100	50	50	77.79	77.79	81.25	72.22	75	70
5	5000	3000	2000	0	0	100	50	0	65	75	81.25	65	68.75	70
6	5000	3000	1000	1000	.10	100	50	0	72.22	77.79	81.25	72.22	75	70
7	5000	3000	2000	0	0	100	50	18.75	68.75	75	81.25	65	68.75	70
8	5000	3000	1000	1000	.10	100	50	25	75	77.79	81.25	72.22	75	70
9	5303	3000	1697	0	0	100	50	21.72	71.72	76.52	81.93	68.03	71.78	71.72
10	5000	2500	1250	1250	.125	100	50	25	75	78.57	83.33	71.43	75	75
11	5000	3000	1000	1000	.10	100	50	20	74.44	77.78	81.25	72.22	75	70
12	6000	4000	0	0	0	100	55	--	82	82	82	82	82	70
13	5000	3000	1000	1000	.10	20	50	100	38.99	33.33	31.25	27.78	35	38