SOME EMPIRICAL INVESTIGATIONS OF NONRESPONSE IN SURVEYS WITH CALLBACKS

J.H. Drew, GTE Laboratories Incorporated
and
G. Ray, GTE Laboratories Incorporated

1. Introduction

We shall develop some extensions to the models proposed by Drew and Fuller (1980) and Proctor (1977) in their studies of the response-nonresponse of individuals given repeated opportunities to respond to a questionnaire. The analysis follows in the spirit of Deming (1953) and bears some methodological resemblance to the work of Politz and Simmons (1949, 1950) and Thomsen and Siring (1979). The concept of Poisson sampling as given by Hajek (1957) is also appropriate in our development.

Suppose that the population is partitioned into K categories based on the values of a discrete random variable. The relative size of the k^{th} category is f_k, k = 1, 2,...,K. Associated with each unit in the k^{th} category is a response probability $q_k \epsilon [0, 1]$ which is the conditional probability that a unit furnishes a response when sampled. Those units which have zero response probabilities are handled as follows: a proportion $1-\gamma$ of the population is composed of hard core nonrespondents who will never answer the survey. The relative categorical composition of this group must be assumed or estimated from other data. The simplest assumption is that the categorical composition is identical to the composition of the entire population. That is,

$$\Pr \{\text{unit in category k/unit in HCNR}\} = f_k, k = 1, 2, \ldots, K.$$

Two potentially undesirable features of this model are that: (1) the response probability is required to be constant over a category; and (2) the response probability is a function of the unit's category only, and is thus not dependent on the survey circumstances under which a response is solicited. The first assumption seems to be necessary in order to avoid any parametric modeling of the response probabilities, but may be reasonable if the categorization is fine enough to admit only slight changes in the response probabilities of units in a given category. The second assumption can be largely eliminated by the incorporation of parameters into the model which represent interviewer, questionnaire, or callback effects. These parameters will not be used in the sequel, but an approach in this area was made by Drew and Fuller (1980) and Thomsen and Siring (1979).

The general survey situation requires a selection of units according to a given sampling design. If some of the units do not respond when contacted, those units are recontacted in a second call. After R calls, the number of sampled units not responding to any call of the survey is recorded. We give the appropriate notation below for simple random sampling.

2. Simple Random Sampling

Let a simple random sample of n units be selected from a population of N units. Let n_{rk} be the number of sample units observed in the K^{th} category on the r^{th} call, and let n_o be the number of sample units unobserved after R calls. Let f_k be the proportion of units in the k^{th} category. Under the assumptions given above, the data $\underset{\sim}{n} = (n_{11},\ldots, n_{RK}, n_o)$ satisfy a multinomial model with cell probabilities $\underset{\sim}{\pi} = (\pi_{11},\ldots,\pi_{RK}, \pi_o)$, where:

$$\pi_{rk} = \gamma(1-q_k)^{r-1}q_kf_k, \; r = 1,2,\ldots, r; \quad k = 1,2,\ldots,K,$$

and

$$\pi_o = 1-\gamma + \gamma \sum_{k=1}^{K} (1-q_k)^R f_k,$$

and we set

$$f_K = 1 - \sum_{j=1}^{K} f_j.$$

Thus, π_{rk} is the probability that an individual in category k will respond on call r, and π_o is the probability that a sampled individual will not have responded by the R^{th} call. The associated log likelihood differs by a constant from:

$$\text{Log L} = \sum_{r=1}^{R} \sum_{k=1}^{K} n_{rk} \log \pi_{rk} + n_o \log \pi_o \tag{1}$$

The solutions to the likelihood equations can be verified to be

$$\hat{q}_1, \hat{q}_2,\ldots,\hat{q}_K, \hat{f}_1,\ldots,\hat{f}_{K-1}, \hat{\gamma}, \tag{2}$$

where

$\widehat{q}_k$ is the solution to the R^{th} degreee polynomial equation

$$\sum_{r=1}^{R} n_{.k}^{-1} n_{rk}(1-rq_k) = [1-(1-q_k)^R]^{-1} R\, q_k(1-q_k)^R, \quad (3)$$

$$\widehat{\gamma} = n^{-1} \sum_{k=1}^{K} [1-(1-\widehat{q}_k)^R]^{-1} n_{.k}, \quad (4)$$

$$\widehat{f}_k = n^{-1} \widehat{\gamma}^{-1} [1-(1-\widehat{q}_k)^R]^{-1} n_{.k} \quad (5)$$

and

$$n_{.k} = \sum_{r=1}^{R} n_{rk}.$$

If the maximum likelihood estimates $\widehat{f}_1,\ldots,\widehat{f}_{K-1},\widehat{q}_1,\ldots,\widehat{q}_K,\widehat{\gamma}$ are in the interval (0, 1), then they are roots of the likelihood equations. Otherwise, the roots must be found numerically. It can be demonstrated that $(\widehat{q}_1,\ldots,\widehat{q}_K,\widehat{f}_1,\ldots,\widehat{f}_{K-1}\widehat{\gamma})$ is consistent for $(q_1,\ldots,q_K,f_1,\ldots,f_{K-1},\gamma)$. See Drew (1981).

In our initital development, we assumed that the categorical composition of HCNR was identical to that of the population. Alternative assumptions are also possible. It is reasonable to postulate that

Pr{unit in category k | unit in HCNR} =

$$(1-q_k)\, f_k \Big(\sum_{k=1}^{K} (1-q_k) f_k \Big)^{-1},$$

i.e., that the relative size of category k among the HCNR is proportional to $(1-q_k)f_k$. This assumption states that the categorical composition of HCNR is the same as the composition of that part of the non-HCNR population which would be expected to respond only after the first call. The modification of the model to incorporate this assumption is straightforward.

The estimates of $\{q_k\}$ for this model are identical to those of the earlier model. The estimates for $\{f_k\}$ are analogous to those of the earlier model. In particular, denoting estimates under this model by $\{f'_k\}$, we have

$$\widehat{f'_k} = n^{-1} [1-\widehat{\beta}(1-\widehat{q}_k)]^{-1} [1-)1-\widehat{q}_k)^R]^{-1} n_{.k},$$

where $\widehat{\beta}$ is chosen so that

$$\sum_{k=1}^{K} \widehat{f'_k} = 1$$

3. A Framework for Assessing the Model

The model we consider in this paper is tentative in the sense that it has not been extensively tested on a variety of data sets, and because some of the assumptions leading to the simple form given in Section 2 are controversial. To assess its performance under a variety of conditions, including several which violate assumptions on which the model is based, we have calculated the approximate expectations of two estimators based on our models, as well as the approximate expectation of a "naive" estimator. Two categories were chosen, and the object was to estimate f_1 which was chosen to be 0.5.

The levels of γ, the fraction of potential respondents, and (q_1, q_2), the response probabilities for the two categories were chosen as follows:

γ = 0.75, 0.95

(q_1, q_2) = (0.8, 0.7), (0.8, 0.5)
(0.5, 0.4), (0.5, 0.2)

Note that in addition to assessing the effect of high and low q_k values, these values of (q_1, q_2) also generate high and low bias in simple means since such bias is caused by large values of $|q_1 - q_2|$. Four callbacks were performed for each situation, so for most cases a substantial number of potential respondents would be unobserved at the last call.

The other two conditions set for these calculations attempt to explore the two controversial assumptions of the model. First, some assumption must be made about the categorical composition of the hard core nonrespondents. Our first model in Section 2 supposes that this composition is identical to that of the entire population. The extended model supposes that the fraction of hard core nonrespondents in category k, k = 1, 2, is proportional to $(1 - q_k)$, k = 1, 2, i.e., that the composition is like that part of the population which would be expected to respond on the second and subsequent calls only. Since it is difficult to organize hypotheses about hard core nonrespondents when their composition is different from any subgroup of the respondents, we chose the fraction of hard core nonrespondents in category k, k = 1, 2 to be proportional to $(1 - q_k)^{HCC}$, for HCC = 0, 4, 8. Thus, the categorical composition of the hard core is identical to those for individuals who would not be expected to respond until after the HCC^{th} call. Note that HCC = 0 corresponds to the assumption made in the first model of Section 2.

A second questionable assumption of the model in Section 2 is that responses within a category are independent from one call to another. One way to create a range

TABLE 1

APPROXIMATE EXPECTATIONS OF THREE f_1 ESTIMATORS

α	1		5		10	
HCC	γ = 0.75	γ = 0.95	γ = 0.75	γ = 0.95	γ = 0.75	γ = 0.95
0	0.502 0.500 0.469	0.502 0.500 0.495	0.508 0.492 0.461	0.508 0.492 0.487	0.512 0.484 0.453	0.512 0.484 0.479
	0.516 0.500 0.438	0.516 0.500 0.489	0.540 0.492 0.430	0.541 0.492 0.481	0.553 0.484 0.423	0.553 0.484 0.473
	0.519 0.500 0.485	0.519 0.500 0.498	0.528 0.445 0.432	0.528 0.445 0.443	0.532 0.416 0.404	0.532 0.416 0.414
	0.614 0.500 0.465	0.614 0.500 0.494	0.631 0.445 0.414	0.631 0.445 0.440	0.641 0.416 0.387	0.641 0.416 0.411
4	0.613 0.612 0.573	0.519 0.518 0.512	0.619 0.602 0.564	0.526 0.509 0.504	0.623 0.592 0.555	0.530 0.501 0.495
	0.672 0.658 0.576	0.541 0.525 0.513	0.694 0.648 0.567	0.565 0.517 0.505	0.704 0.637 0.557	0.578 0.508 0.497
	0.576 0.558 0.542	0.528 0.509 0.507	0.586 0.497 0.483	0.537 0.454 0.451	0.590 0.4650. 451	0.541 0.424 0.422
	0.724 0.623 0.578	0.632 0.519 0.513	0.738 0.555 0.515	0.649 0.463 0.457	0.746 0.518 0.481	0.658 0.433 0.427
8	0.656 0.654 0.613	0.526 0.524 0.519	0.661 0.643 0.603	0.532 0.516 0.511	0.665 0.633 0.593	0.536 0.507 0.502
	0.680 0.666 0.583	0.542 0.526 0.515	0.702 0.656 0.574	0.567 0.518 0.506	0.712 0.645 0.564	0.579 0.509 0.498
	0.621 0.604 0.586	0.535 0.516 0.514	0.630 0.538 0.522	0.544 0.460 0.458	0.634 0.503 0.488	0.548 0.430 0.428
	0.754 0.659 0.612	0.637 0.525 0.519	0.768 0.587 0.545	0.654 0.468 0.462	0.775 0.549 0.510	0.663 0.437 0.432

Note: Triplets are $(\hat{f}_1^N, \hat{f}_1, \hat{f}_1')$

The four rows within a cell correspond to (q_1, q_2) = (0.8, 0.7), (0.8, 0.5), (0.5, 0.4), (0.5, 0.2)

of violations of this assumption is to consider a 2 x 2 contingency table with entries M_{ij}, i = 1,2, j = 1,2

where

M_{ij} = number of individuals whose response decision was i on one call and j (if recontacted) on the next call.

Arbitrarily let i, j = 1 if the individual would decide to respond, and i, j = 2 if the individual would decline to respond. (Observe that these values are theoretical only, since in practice a respondent is not recontacted.) A standard measure of association is the cross product ratio $\alpha = (M_{11}M_{22})/(M_{12}M_{21})$, large values of α corresponding to high direct dependence between calls, and $\alpha = 1$ corresponding to independence between calls. For our calculations, α values of 1, 5 and 10 were chosen.

The effect of α can be seen by noting that when $\alpha = 5$, and the probability of a response on the first call is 0.50, then the conditional probability of a response on a second call given nonresponse on the first call is only 0.31. If $\alpha = 10$, then that conditional probability is 0.24.

Thus, with three levels of α, three of HCC, two of γ, and four of (q_1, q_2), there are 72 factor level combinations in all.

For each of the 72 situations, three estimates of f_1 were considered. First the approximate expectation of the naive estimator

$$\hat{f}_1^N = n_{.1}/n_{..}$$

was calculated, and then those estimators from Section 2, namely $\hat{f}_1$ and $\hat{f}_1'$ were formed and their approximate expectations calculated. The results are given in Table 1. In these calculations, first order Taylor approximations of each f_1 estimator were formed, and expectations of these approximations were taken. Thus, the approximate expection of $\hat{f}_1^N$ is calculated as $E(n_{.1})/(E(n_{.1}) + E(n_{.2}))$, and the approximate expectation of $\hat{f}_1$ is $E(n_{.1})/(n\gamma(1-(1-q_1)^4)$.

The layout of this table makes one general point obvious: $\hat{f}_1$ and $\hat{f}_1'$ are nearly always better than $\hat{f}_1^N$, in the sense of being closer to $f_1 = 0.5$, and their improvement over $\hat{f}_1^N$ is often dramatic. Only when HCC = 0 and $\alpha > 5$ does $\hat{f}_1^N$ have smaller bias than $\hat{f}_1$ or $\hat{f}_1'$. Indeed, when (q_1, q_2) = (0.8, 0.7) or (0.8, 0.5) then most individuals have responded by the last callback and $\hat{f}_1^N$ is reasonably good. Then $\hat{f}_1$ and $\hat{f}_1'$ are only slightly better or slightly worse.

When (q_1, q_2) = (0.5, 0.4) or (0.5, 0.2) the naive estimator can be poor, and $\hat{f}_1$, $\hat{f}_1'$ can be significantly better. The relative goodness of these estimators with low response probabilities is important when one considers that other weighting techniques seem to require moderately high response probabilities to perform well. Generally, only when HCC = 0 (and especially when $\alpha > 5$) does f_1 have smaller bias than f_1'. In these situations both estimators have a negative bias and thus tend to underestimate f_1.

The value of HCC can greatly affect the quality of $\hat{f}_1^N$ and $\hat{f}_1$, especially when $\gamma = 0.75$. In fact, when HCC = 8, $\hat{f}_1$ shows only slight improvement over $\hat{f}_1^N$. In a few cases, $\hat{f}_1'$ does noticeably better than $\hat{f}_1$. It is somewhat surprising that f_1', which essentially assumes HCC = 1, does quite well when HCC = 8. This performance is particularly encouraging when the proportion of hard core nonrespondents is quite high ($\gamma = 0.75$).

4. An Empirical Test of the Model

While it is encouraging to see the relatively good performance of our estimators in the framework of the preceding section, it remains to assess these models in actual surveys, where nonresponse may operate in a more inscrutable way. Such an empirical test may also suggest modifications in our approach which reflect real survey exigencies.

We take our data from Jones [1983]. In 1976, the Australian Federal Government sponsored a mail survey in East and West regions of the Canberra suburb of Tuggeranong. A systematic random sample of addresses was selected for each region, and an interview schedule was hand-delivered to each selected address. Approximately two weeks later, a "reminder" postcard was mailed to each address, with a second "reminder" one week after that. Three weeks after that date, a final reminder letter with additional questionnaires was sent to each address. Jones perceived three waves of respondents: those responding to the initial mailing, those responding to one of the first two "reminders," and those responding to the mailing of additional questionnaires. The respondents were allocated to these waves based on the date of receipt of the completed questionnaire.

An interview survey was also conducted for an interpenetrating sample of addresses in each region. Personal interviews were conducted during the first half of the six week period over which mail survey responses were collected. Nearly all selected subjects from the interview samples furnished completed questionnaires, the response rates being given as 95.5 percent for the East region and 90.3 percent for the West region.

The analysis of these data considered only factual, and not attitudinal variables, to guard against the contamination of this study by interviewer response error. Each such variable was an attribute variable, and Jones' data consists of the proportions of respondents possessing the attribute in question. Three estimates for each proportion were given in the original study.

In addition to the usual ratios based on all responses to the mail survey, and on all responses to the personal interview, Jones gives an extrapolated estimate obtained from a linear regression of cumulative attribute proportion on cumulative response over the three mailing waves.

We supplement this work by calculating $\hat{f}_1$ from (3)-(5) in Section 2. Here there are K = 2 categories and R = 3 callbacks. Note that we treat the sample as a simple random sample. In addition, we reconstructed respondent counts from the proportions given in Table 10 of Jones [1983] by simple multiplication of those proportions by the given respondent numbers in each wave. This latter step was modified slightly to force the estimated counts to be integers consistent with the stated proportions to be achieved by rounding. (In addition, the counts were to be consistent with Jones' assertion that there were no more than a "few" omissions of any given questionnaire item.)

One aspect of the survey situation suggests the need for a modification of our first model. Since the third wave of the mail survey consisted of a second copy of the schedule, it is reasonable to suppose that the response probabilities for each category are higher for this wave than for the first two waves. Since these data can accommodate only one more parameter in a model of this sort, and still afford a goodness of fit test, we include the parameter δ in the model. This parameter can be interpreted as a measure of the multiplicative decrease in the probability of an individual's not responding on the third call. Let

$$\pi_{rk} = \gamma(1-q_k)^{r-1}(1-\delta(1-q_k))f_k$$

for r = 3, and let

$$\pi_0 = (1-\gamma) + \gamma\delta \sum_{k=1}^{K} (1-q_k)^R f_k$$

and let the other expressions for π_{rk} be identical to those of first model of Section 2. Denote the maximum likelihood estimate of f_1 based on this model by $\hat{f}_\delta$.

These five estimators, $\hat{f}_1, \hat{f}_\delta$, Jones' extrapolation estimator, and the simple ratios from the mail and interview surveys are shown in Tables 2 and 3 for each of 16 variables from each survey region, East and West.

Note first that the simple ratio estimator and the $\hat{f}_1$ estimator perform poorly on these data. In many cases Jones' extrapolation estimator improves the situation in the sense that these estimators are closer to the interview estimates than the simple ratio, but in at least seven of the East variables and six of the West variables, the extrapolation estimator is still more than ten percentage points away from the interview data. The $\hat{f}_\delta$ estimator shows some improvement, for there are three variables (Age<30, Res. Age<30, No Neighbors as Friends) in the East for which this estimator is clearly better than the extrapolation estimator, and five such variables (Income, Res. Income, No Neighbors as Friends, See

TABLE 2

EAST REGION: ESTIMATED PROPORTION POSSESSING ATTRIBUTE

Variable	Simple	$\hat{f}_1$	Extrap'tion	$\hat{f}_\delta$	Interview
Tert. Ed.	0.464	0.419	0.488	0.467	0.316
PTEM Occ.	0.436	0.224	0.328	0.448	0.251
Age<30	0.500	0.318	0.434	0.514	0.560
Inc.>10K	0.601	0.440	0.551	0.609	0.549
Res. Ter. Ed.	0.362	0.279	0.336	0.367	0.227
Res. PTEM	0.305	0.190	0.242	0.309	0.166
Unemp.	0.302	0.295	0.293	0.302	0.166
Res. Age<30	0.564	0.304	0.502	0.575	0.631
Res.Inc.>10K	0.320	0.281	0.296	0.293	0.313
Male	0.489	0.501	0.501	0.488	0.499
Married	0.924	0.927	0.924	0.923	0.944
Gov. H'sing	0.367	0.487	0.391	0.361	0.380
No N'bors as Friends	0.611	0.728	0.677	0.603	0.577
0-2 Friends	0.562	0.709	0.622	0.529	0.620
See Friends Weekly	0.656	0.471	0.622	0.661	0.707
Made New Friends	0.494	0.482	0.479	0.495	0.512

TABLE 3

WEST REGION: ESTIMATED PROPORTION POSSESSING ATTRIBUTE

Variable	Simple	$\hat{f}_1$	Extrap'tion	$\hat{f}_\delta$	Interview
Tert. Ed.	0.488	0.448	0.460	0.485	0.260
PTEM Occ.	0.381	0.317	0.232	0.376	0.222
Age<30	0.607	0.567	0.638	0.518	0.613
Inc.>10K	0.518	0.261	0.259	0.481	0.569
Res. Ter. Ed.	0.412	0.398	0.405	0.412	0.200
Res. PTEM	0.262	0.220	0.186	0.262	0.162
Unemp.	0.273	0.247	0.233	0.272	0.312
Res. Age<30	0.645	0.561	0.630	0.640	0.679
Res.Inc.>10K	0.322	0.203	0.180	0.313	0.308
Male	0.504	0.374	0.505	0.503	0.497
Married	0.917	0.967	0.912	0.915	0.927
Gov. H'sing	0.348	0.401	0.385	0.351	0.437
No N'bors as Friends	0.563	0.740	0.721	0.576	0.565
0-2 Friends	0.481	0.716	0.583	0.492	0.580
See Friends Weekly	0.649	0.512	0.529	0.647	0.785
Made New Friends	0.634	0.217	0.505	0.616	0.624

Friends Weekly, and Made New Friends) for the West data. In contrast, there are only two variables in the East (PTEM and Res. PTEM) and three variables in the West (Res. Educ., PTEM, and Res. PTEM) for which extrapolation is better than $\hat{f}_\delta$.

In both regions, there are five variables (Some Tert. Educ., Res. Tert. Educ., PTEM, Res. PTEM, Not Employed(east) and See Friends Weekly(west)) for which $\hat{f}_\delta$ is significantly different from the interview estimate. In performing these tests, we have exploited the asymptotic normality of the proportions, and have used $\alpha = 0.05$. Furthermore, we note that the first four of these variables for the East region would give better $\hat{f}_\delta$ estimates if the proportions from the third call could be given greater importance. For data with more categories, it would be possible to incorporate this feature by postulating that the categorical composition of the hard core nonrespondents be proportional to the expected number who would be observed after the second call. A minor modification of the models in Section 4 would do this.

5. References

1. Cassel, C.M., Sarndal, C.E., and Wretman, J.H. (1983), "Some Uses of Statistical Models in Connection with the Nonresponse Problem," in *Incomplete Data in Sample Surveys* (Vol. 3), ed. W. Madow and I. Olkin, New York: Academic Press, 143-160.

2. Deming, W.E. (1953), "On a Probability Mechanism to Attain an Economic Balance Between the Resultant Error of Nonresponse and the Bias of Nonresponse, *Journal of the American Statistical Association 48,* 743-772.

3. Drew, J.H. (1981), "Nonresponse in Surveys with Callbacks," unpublished Ph.D. thesis, Iowa State University, Dept. of Statistics.

4. Drew, J.H. and Fuller, W.A. (1980), "Modeling Nonresponse in Surveys with Callbacks," *Proceedings of the Section on Survey Research Methods of the American Statistical Association.*

5. Hajek, J. (1957), "Some Contributions to the Theory of Probability Sampling," *30th Session of the International Statistical Institute,* 127-134.

6. Jones, R.G. (1983), "An Examination of Methods of Adjusting for Nonresponse to a Mail Survey: A Mail-Interview," in *Incomplete Data in Sample Surveys* (Vol. 3) ed. W. Madow and I. Olkin, New York: Academic Press, 271-290.

7. Little, R.J. (1982), "Models for Nonresponse in Sample Surveys," *Journal of the American Statistical Association* 77, 378, 237-250.

8. Politz, A.N. and Simmons, W.R. (1949), "An Attempt to Get the 'Not-at-Homes' into the Sample Without Callbacks," *Journal of the American Statistical Association 44,* 136-137.

9. Proctor, C. (1977), "Two Direct Approaches to Survey Nonresponse: Estimating a Proportion with Callbacks and Allocating Effort to Raise the Response Rate," *Proceedings of the Social Statistics Section of the American Statistical Association,* 284-290.

10. Thomsen, I. and Siring, E. (1983), "On the Causes and Effects of Nonresponse: Norwegian Experiences," in *Incomplete Data in Sample Surveys,* (Vol. 3), ed. W. Madow and I. Olkin, New York: Academic Press, 25-59.