COMPARISON OF SEVERAL IMPUTATION METHODS

Choiril Maksum and William D.Warde, Oklahoma State University
William D.Warde, Stat. Dept., 301 MS, OSU, Stillwater, OK 74078

KEY WORDS: Imputation, Monte Carlo, Non Response

INTRODUCTION

One major problem which is encountered by practitioners in the field of sample surveys is non–response. However hard they try to obtain a response from every element of the sample who was selected, there will come a time when efforts must cease and analysis of the collected data must begin. At this point in time, there will usually be some individuals for whom there has been no data collected. The proportion of these non–respondents for any survey has been shown to vary as a function of a number of variables.

Among these are the method of data collection (O'Niell, 1975, Jordan, Marcus and Reeder, 1978), the type of population being studied, the credibility of the organization conducting the survey (Brunner and Carroll, 1969, Houston and Nevin, 1977), etc. It has also been shown in many studies that characteristics of the respondent will also affect the probability of a response by that individual (Sharp and Feldt, 1959, Greenlees, Reece and Zieschang, 1982).

Many techniques are used to attempt to improve the response rate in surveys, and then it is common in some areas for results to be imputed for the non–respondents. Essentially, imputation is an attempt to estimate a value for a response variable for those individuals who have not responded, and to use this value as if the individual had responded and had given the imputed value for his/her answer. Most imputation techniques rely on assigning as the imputed value some function of the values for a group of respondents who are chosen by the survey researcher to be as "similar" as possible to the non–respondent.

In many cases, this similarity is determined from demographic variables which are available from the listing or frame from which the sample was selected.

In others, such as "hot deck" imputation, randomly selected respondents are used as replacements for the missing data.

Another case in which imputation commonly occurs is whenever the response suffers from item non–response. Most of the data in the survey is available but there are certain respondents who have failed to answer a few of the questions on the questionnaire for some reason. In these cases, their responses are often matched to other respondents in the survey based upon those items which are completed, and then the values for the missing items imputed from the values for those individuals deemed to be sufficiently similar to those with missing values based upon the data available.

This paper discusses a method of imputation which is a modification of the technique suggested by Hendricks (1949). It is based upon the estimation of the mean of the non–responding group based on data obtained in multiple attempts to contact the individuals selected as members of the sample. Other statistical imputation techniques which exist in the literature are the original empirical method suggested by Hendricks (1949), the maximum likelihood technique developed by Drew (1981) (see also Drew and Fuller (1980, 1981), and the weighting adjustment method proposed by Oh and Sheuren (1983).

In this paper, these four methods will be compared using Monte Carlo techniques.

NEW IMPUTATION TECHNIQUE

Suppose we select a sample of size n from our population. On the first contact, we obtain n_1 responses. We then attempt to contact the $(n-n_1)$ non–respondents at this stage and obtain n_2 respondents on our second contact attempt. We repeat this technique J times, terminating our interviewing attempts after the J–th attempt.

At this point we can categorize our respondents into one of the j, j=1,2, ..., J, groups dependent upon when they responded, and can regard all of the non–respondents as belonging to the (J+1)–th group. If we also define $p_j = n_j/n$ to be the proportion of individuals who responded on each contact attempt, with $p^* = p_{J+1}$ being the proportion of non–respondents, then we can estimate the mean of the population for some desired attribute Y by

$$\bar{y}_p = \sum_{j=1}^{J} p_j \bar{y}_j + p_{J+1}\bar{y}^*$$

where $\bar{y}_j$ represents the mean of the Y variable for each of the partitions of the sample defined by the interview attempt on which they responded, and $\bar{y}^*$ represents some estimate of the mean response for those selected individuals who have not responded by the time that the survey was terminated.

It can be seen in many data sets that there is some trend relating $\bar{y}_j$ to j. This trend was discussed by Hendricks (1949) and is shown in data sets published by Clausen and Ford (1947), Hilgard and Payne (1944), and Finkner (1950) among others. Table 1 shows the data from Finkner as an example. In this data, Finkner reports the number of fruit trees for the non–respondents based upon data which was available from the Census of Agriculture. Hendricks (1949) referenced this data set and commented on the perils of either terminating a mail survey with too few mailings or of ignoring the effect of bias due to assuming that non–respondents are the same as respondents in any survey.

We suppose that for each response obtained there is some true response which may have been contaminated by a measurement error. We therefore assume

$$y_{ij} = Y_{ij} + e_{ij}$$

for $i = 1, 2, ..., n_j$;

$j = 1, 2, ..., J+1$.

Note that in this representation, group (J+1) represents the non–respondents for whom we will NOT actually have an observed y_{ij} value, but for whom there is an actual Y_{ij} value.

For the proposed technique (a formalized version of the Hendricks method) we suppose that

$$\bar{y}^* = \bar{y}_J + \exp(1-J)/\alpha$$

for some unknown parameter value α. We compute the mean square error of $\bar{y}_p$ and thence obtain an estimator for α by minimizing that mean square error. This results in the following

$$\hat{\alpha} = \{gp^*(n-1) - g\}/\{np^*c)\}$$

where

$$c = ap_J n^{-1} \sum_{i=1}^{N_J} Y_{iJ} - p^* n^{-1} \sum_{i=1}^{N^*} Y_{ij}$$

$a = E(n^*/n_J)$

and

$g = p^* \exp(1-J)$.

MONTE CARLO STUDY

In order to make a comparison among the several imputation methods, a simulation study was conducted. Several populations of size N=1000 were simulated using the Gamma distribution (shape parameter r = 2, 3, 4, 7; scale parameter λ = 10, 25, 50) and the exponential distribution with parameter θ = 0.1, 0.25, 0.5, 1, 2, 3, 4, 5, 6, 7. These distributions were selected in order to represent "typical" populations for which the variable of interest is always positive and highly skewed. A simple random sample of size n=100 was selected from each of these populations, and a uniform random deviate generated and associated with each of these selected observations. The value of the uniform random deviate was used to determine whether the individual "responded" on the j–th attempt or was categorized as a non–respondent. This was achieved by comparing the value of the uniform random deviate to values obtained from the probability of response model

$$\pi_{ij} = (mb_j)^{-1} \exp[-y_{ij}/(mb_j)]$$

where π_{ij} denotes the probability of a response by the i–th individual on attempt j; y_{ij} denotes the value of the variable of interest for the i–th individual on the j–th call; m is the population mean; and b_j denotes an appropriately chosen constant to ensure that the probability of a response is different on different calls.

The simulations allowed for three distinct models for the response probabilities based upon the choices of the parameters b_j as follows:

(1) The b_j values increase with j.

(2) The b_j values decrease with j.

(3) The b_j values increase and then decrease (with j=2 as the maximum).

Model (3) has been shown to be a reasonable model by Kish (1965) and Warde (1987). This effect is common in personal and telephone interviews in which the respondent is unavailable at the time of the first contact, but information can be obtained from a relative, neighbor, etc. which will increase the probability of locating that individual on the second attempt to contact them, thus causing an improved response probability on the second attempt.

Assuming there is no measurement error in the model, then the estimator of the population mean is computed for the resulting sample using the unadjusted method (mean of all respondents), Hendricks' (1949) method, Drew's (1981) method, and the Oh and Sheuren (1983) method as well as by the new method described above. The number of calls, j, was varied from 3 to 5 in these simulations.

Tables 2 and 3 summarize some of the results obtained from these Monte Carlo simulations. Additional details are available in Maksum (1990). For the populations studied, Drew's technique seems best for all situations in which the response probability followed model (3), and it was also best for the exponential populations with 4 or 5 calls where the response probability was decreasing with j (model (2)). The technique proposed by Oh and Sheuren (1983) performs well for some situations with decreasing probability of response (model (2)), particularly the Gamma populations with intermediate parameter values, and the exponential populations when only 3 calls were made. The Hendricks model and the proposed variant of that model perform well for increasing probability of response (model (1)) situations in the exponential populations, and for most of the gamma populations. Drew's technique does supplant these techniques for the rapid response gamma populations.

The unadjusted technique (assuming non–respondents are the same as respondents) rarely performed well. This is with the models analysed since the population models studied were designed in such a manner as to make the assumptions for this model untenable.

Since experience would indicate that in most practical situations, models (2) or (3) are likely to match real response probabilities, our overall conclusion is that the technique suggested by Drew (1981) is the best general purpose imputation method among the methods examined in this study. This conclusion is based upon the assumption that the exponential or gamma type populations are realistic models for the population, and that the probability of an individual responding on any particular contact attempt does depend upon the magnitude of the value of the variable of interest in the study.

BIBLIOGRAPHY

Brunner, M.J. & Carroll, S.J. (1969) "The Effect of Prior Notification on the Refusal Rate in Fixed Address Surveys." J of Advertizing Research,9#1:42–44.

Clausen, J.A. & Ford, R.N. (1947) "Controlling Bias in Mail Questionaires" JASA, 42, 497–511.

Drew, J.H. (1981) "Nonresponse in Surveys with Callbacks" Unpub. Ph.D. Dissertation, Iowa State University, Ames, IA.

Drew, J.H. & Fuller, W.A. (1980) "Modelling Nonresponse in Surveys with Callbacks" Proc of the Survey Research Methods Section of the ASA, 639—642.

Finkner, A.L. (1950) "Methods of Sampling for Estimating Commercial Peach Production in North Carolina." North Carolina Agricultural Experiment Station Technical Bulletin #91.

Greenlees, J.S., Reece, W.S. & Zieschang, K.D. (1982) "Imputation of Missing Values when the Probability of Response Depends upon the Variable being Imputed" JASA, 77, 251—261.

Hendricks, W.A. (1949) "Adjustment for Bias in Nonresponse in Mailed Surveys" Agricultural Economics Research, 1, 52—56.

Hilgard, E.R. & Payne, S.L. (1944) "Those Not at Home: Riddle for Pollsters." Public Opinion Quarterly, 8, 254—261.

Houston, M.J. & Nevin, J.R. (1977) "The Effect of Source and Appeal in Mail Survey Response Patterns" J of Marketing Research, 14:374— 378.

Jordan, L.A., Marcus, A.C. & Reeder, L.G. (1978) "Response Styles in Telephone and Personal Interviewing: A Field Experiment from the Los Angeles Health Survey" Proc of the Survey Research Methods Section of the ASA, 362—366.

Maksum, C. (1990) "A New Method for Imputing Missing Values when the Probability of Response Depends on the Variable Being Imputed." Unpub. Ph.D. dissertation, Oklahoma State University, Stillwater, OK.

O'Niell, M.J. (1979) "Estimating the Non—response Bias Due to Refusals in a Telephone Survey." Public Opinion Quarterly, 43:218— 237.

Oh, H.L. & Sheuren, F.J. (1983) "Weighting Adjustment for Unit Nonresponse" In: Madow, W.G., Olkin, I. & Rubin, D.B. (eds) Incomplete Data in Sample Surveys, Vol 2. Theory and Bibliography. New York, NY: Academic Press, 249—333.

Sharp, H. & Feldt, A. (1959) "Some Factors in a Probability Sample of a Metropolitan Community" American Sociological Review, 24, 650—661.

Warde, W.D. (1987) "Time of Day for C.A.T.I. Contacts in Agricultural Surveys." Proc of the Survey Research Methods Section of the ASA, 621—626.

Table 1. Response to a mail enquiry to fruit growers in North Carolina

	# Growers	% response	Mean # fruit trees/grower
Response to first mailing	300	9.6	456
Response to second mailing	543	17.4	382
Response to third mailing	434	13.9	340
Non–response after three	1839	59.0	290
Total	3116	99.9	329

Table 2. Summary of Best Methods for the Gamma Population.

	Response Probability Model		
# of Calls	Increase	Decrease	Increase–decrease

3	D(2,10) to D(2,50)
	P(3,10) to P(4,10)
	D(4,25) to D(7,50)
		D(2,10) to D(3,50)
		OS(4,10) to OS(7,50)
			D(all)

4	D(2,10) to D(2,50)
	P(3,10) to P(4,25)
	D(4,50) to D(7,50)
		D(all)	D(all)

5	D(2,10) to D(3,10)
	P(3,25) to P(7,50)
		D(2,10)
		OS(2,25) to OS(7,25)
		D(7,50)
			D(all)

D = Drew's method; P = Proposed Method; H = Hendricks' Method
OS = Oh and Sheuren's method
Parenthetic values are parameter values for the Gamma distribution for which the indicated technique performs best.

Table 3. Summary of Best Methods for the Exponential Population.

	Response Probability Model		
# of Calls	Increase	Decrease	Increase–decrease

3	H(0.1) to H(0.5)
	P(1) to P(5)
	H(4) to H(7)
		OS(0.1) to OS(4)
		P(5) to P(7)
			D(all)

4	P(all)	D(all)	D(all)
5	P(all)	D(all)	D(all)

D = Drew's method; P = Proposed Method; H = Hendricks' Method
OS = Oh and Sheuren's method
Parenthetic values are parameter values for the Exponential distribution for which the indicated technique performs best.