Pentru a cita lucrarea / To cite the paper: Drane JW, Petrisor AI, Dragomirescu L (2003), Imputing Nonresponses to Mail-back Questionnaires: All Questions Have Polychotomous Responses, 1st MEDINF International Conference on Medical Informatics & Engineering, October 9-11, 2003, Craiova, Romania

Imputing Nonresponses to Mail-back Questionnaires: All Questions Have Polychotomous Responses

J. Wanzer DRANE, PhD, PE[1], **Alexandru-Ionut PETRISOR**, MSPH[2], and **Liviu DRAGOMIRESCU**, PhD[3]

[1] Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, USA; [2] Department of Environmental Health Sciences, Arnold School of Public Health, University of South Carolina, USA

[3] Department of Ecology, Faculty of Biology, University of Bucharest, Romania

Introduction

Even though the self-administered questionnaires are the least expensive of several study options because of the lack of the interviewer [1], the cost reflects in reduced accuracy and precision due to the resistance of persons receiving the questionnaires to filling them in and mailing them back [2].

"Accuracy" refers to a bias of unknown magnitudes due, in this case, to nonresponses, leading eventually to wrong conclusions, whereas "precision" relates to the statistical variation of the standard error of the estimates derived from the sample [3]. The number of mailouts to which persons respond become functionally connected strata.

Our previous studies used questionnaire with dichotomous responses [1, 2]; here, all questions may have polychotomous responses; therefore, this research represents a generalization. Furthermore, mailouts may be replaced by "waves", meaning contacts of any kind: telephone, in-person, mail, e-mail, etc.

Imputation of Nonresponses

The underlying assumption is that those who fill out the questionnaire exhibit different behaviors according to wave or number of contact to which they have responded. If there is a vested interested in the subject of the survey, the person should need little prompting to obtain a completed questionnaire. If identity is not a problem, a record may be made as to the wave number of each obtained response. The first wave might be a mailed questionnaire, the second a postcard, the third a phone call and the fourth a personal visit, face to face interview. Even then some are expected to withstand the pressures and steadfastly refuse to give up the requested information. Let us suppose for a particular question there are four possible responses such as 1="Dislike Very Much", 2="Dislike", 3="Like", 4="Like Very Much", keeping in mind that there may be an arbitrary number of choices and that they do not have to be on a Likert or ordinal Scale. Thus, each response here illustrated is a vector of three zeros and a one, where 0="Absent" for the three possibilities not checked, and 1="Present" for the possibility checked. That is Y=(1, 0, 0, 0) represents "Dislike Very Much", etc. for the other three possibilities. Further, let us suppose there are i=1,…, I choices to a particular question, and there are j=1,…, J waves with N_j respondents on the jth wave. For each wave we calculate $P'_j=(p_1, p_2,\ldots, p_I)_j$. There being J waves,

$$P_j=B_0+B_{1j} \qquad (1)$$

is the multivariate regression of the vector P, and j is the predictor variable. The constraint is, of course, P'1=1 for all values of j. After J waves there remain $N_{J+1}=N-\Sigma N_j$ who have not responded to the questionnaire. Setting j=J+1 in Equation 1 gives the imputation of P_{J+1}, which completes the sample for that question. The final estimate of the vector P is

$$\hat{\mathbf{P}}=\frac{\sum_{j=1}^{j=J+1} N_j \mathbf{P}_j}{\mathbf{N}} \qquad (2)$$

An Example

Let us assume two questions, each with three possible responses such as “No”, “Maybe”, “Yes”. Let there also be three waves with the fourth non-existent wave containing all of the non-responders. Table 1 contains the data after the zeros and ones have been converted into the vectors $P_{j,\ j}$=1,2,3. These three P_j are then regressed on j, and P4 is the extrapolation after the three waves, giving us the imputed values of the elements of P.

Q_1	j=1	j=2	j=3	j=4	$\hat{\mathbf{P}}_1$	$\hat{\mathbf{b}}$
p_1	0.60	0.51	0.39	0.2900	0.4427	0.02959
p_2	0.15	0.17	0.20	0.2233	0.1870	-0.00705
p_3	0.25	0.32	0.41	0.4867	0.3703	-0.02254
N_j	142	98	65	167	1.0000	
Q_2	j=1	j=2	j=3	j=4	$\hat{\mathbf{P}}_2$	$\hat{\mathbf{b}}$
p1	0.24	0.36	0.44	0.3800	0.3420	-0.00737
p2	0.22	0.18	0.12	0.0733	0.1460	0.01409
p3	0.54	0.46	0.44	0.5467	0.5120	-0.00673
N_j	142	98	65	167	1.0000	

Table 1. The results of tabulations of two hypothetical questions and their respective proportions for three waves with an extrapolation to obtain the respective proportions for the fourth wave. Multivariate linear regression was used. Proportion not responding is 167/472=0.3538.

The final column for each question contains estimates of bias should the fourth wave of responses had not been imputed. It can easily be shown that the bias is the prevalence of non-responses times the difference of the means (proportions) of those responding and those who are not responding. That is, bias is estimated by Equation 3. Reviewing table 1, one can see that

$$\hat{\mathbf{b}}=(1-\theta)(\hat{P}_R-\hat{P}_N) \qquad (3)$$

the relative bias is of the order of magnitude of five percent or less when omitting the fourth wave completely. Caveat emptor, at least one now has a means of judging whether to use imputation of this kind when analyzing data.

Discussion

Every contingency table can be represented as a multinomial both within the table and along its margins. Therefore the method described above can easily be extended to imputations of the internal cells of the table, and the margins are exactly as described above. The two hypothetical questions could well form the margins of a 3x3 table. Then the test of the hypothesis of independence would follow precisely as can be found in many texts when using the likelihood ratio chi-square. That is

$$G^2=2N\sum_{r=1}^{R}\sum_{c=1}^{C}\hat{p}_{rc}LOG(\frac{\hat{p}_{rc}}{\hat{p}_{r.}\hat{p}_{.c}}), \qquad (4)$$

where r=row number, and c=column number. G^2 is the likelihood ratio chi-square with (R-1)(C-1) degrees of freedom. Other chi-squares might be calculated for other sets of hypotheses. The reader is cautioned to take care when calculating the appropriate degree of freedom because of the imputation. That is, if the bias is not too severe, one might choose to forgo it and analyze the data as though they were derived from a simple random sample.

References

[1] J. W. Drane, Imputing Nonresponses to Mail-back Questionnaires, American Journal of Epidemiology 134 (1991) 474-478.

[2] J. W. Drane, D. Richter, C. Stoskopf, Improved Imputation of Non-Responses to Mailback Questionnaires, Statistics in Medicine 12 (1993) 283-288.

[3] P. S. Levy, S. Lemeshow, Sampling of Populations: Methods and Applications. John Wiley & Sons, New York, 1999, ISBN: 0-471-15575-6, p. 37-8.