Compensating for missing data in psychiatric surveys

GRAHAM DUNN

INTRODUCTION

Missing observations are a characteristic of all psychiatric surveys. They may arise from the fact that particular measurements were not made on some individuals, or from the subsequent discovery that the measurements were either made or recorded in error and that they should, therefore, be dropped from the data set prior to any analysis. The fact that the observations were never made might be accidental (i.e. missing by happenstance) or arise from a deliberate decision of the investigator (i.e. missing by design). 2-phase prevalence surveys provide an example of the latter (see below), whilst 'accidentally-missing' data can arise from a variety of reasons ranging, for example, from the administrative or technical incompetence of the investigator to the death or emigration of the subject. Perhaps the subject cannot be contacted at the time of the survey, or is too ill to participate in a survey interview. Or perhaps the patient or informant simply refuses to take part in the investigation.

Traditionally investigators have coped with missing value problems in multivariate data sets by analysing only that part of the data with no missing observations (complete-case analysis). Although in some situations this strategy might be warranted, in general the data analysis should use all available information. The complete-case analysis is likely to be inefficient (unless the number of subjects with missing data is relatively-small) but, more-importantly, it might lead to biased estimates due to the fact that the complete cases are not representative of the sample as a whole. In many longitudinal cohort studies, for example, the subjects who drop out, or who are otherwise lost to follow-up, will differ systematically from those who remain in the study (see, for example, Gornbein *et al.*, 1992). People do not die, or become sick, or become non-cooperative completely at random. The same can also be said for subjects with missing data in the simpler cross-sectional and case-control studies.

The purpose of this editorial is to introduce the reader to general-purpose methods of compensating for missing data in the analysis of an epidemiological survey. However, we will discuss neither special-purpose techniques that are tailor-made for a particular statistical methodology nor sophisticated models for drop-outs in longitudinal data. For the latter the reader is referred to Gornbein *et al.* (1992) and Little (1995). Here we will be particularly concerned with the use of so-called expansion weights to compensate for partial non-response (see below), whether this arises by accident or by design. More detailed reviews of the methods discussed can be found in Brick & Kalton (1996), Pickles & Dunn (1998) or Pickles *et al.* (1995).

PATTERNS OF MISSING DATA

Brick & Kalton (1996) describe four sources of missing data. *Total* or *unit non-response* is probably the most familiar. Here the patient or informant provides no survey information at all. Data on this subject is completely missing. Either the subject is unavailable for interview, for example, or cannot be traced, or refuses to take part in the investigation. Compensation for total non-response is usually made by means of *weighting adjustments* in which the respondents (i.e. those who provide the required information) are assigned greater weight in the analysis in order to represent the non-respondents. The second source of missing survey data is *incomplete*

Indirizzo per la corrispondenza: Professor G. Dunn, School of Epidemiology and Health Sciences, Medical School, University of Manchester, Oxford Road, Manchester M13 9 PT (UK).
Fax +44 - 161-275.5567.

coverage arising from the inadequacy of the survey's sampling frame. Here there are patients or informants who have no chance of selection for interview, simply because they are not listed in the sampling frame. Again, compensation is usually made through the use of weights, but here the weights have to be determined by reference to external data sources. In the case of total non-response, however, weights can be determined by reference to the actual sample.

A third source of missing data is *item non-response* in which there may be one or more variables for which there is inadequate information provided by the respondent. The pattern of missing data might also vary from one respondent to another. Item non-response can arise for a variety of reasons. A patient or informant may refuse to answer a particular question, or may not understand what the question means, or simply not know the answer. The interviewer may forget to ask the question or record the answer. The answer may be coded incorrectly, and so on. The most frequent form of compensation for item non-response is *imputation* (assigning a value for the missing response). Finally, Brick & Kalton (1996) discuss *partial non-response*. Partial non-response involves a substantial number of item non-responses, but typically they are not occurring in a haphazard way. In a 2-phase prevalence survey, for example, all subjects (the 1st-phase sample) might provide demographic information and the results of administering a screening questionnaire — but only selected sub-samples (the 2nd-phase respondents) are given a structured or semi-structured psychiatric interview. If the 2nd-phase respondents provide a single measurement (psychiatric diagnosis, for example) then this is an example of item non-response. If, however, they provide, say, a detailed breakdown of their symptoms together with further background information, then this is clearly an example partial non-response. Often, however, the distinction will be blurred. Compensation for partial non-response can be handled by either weighting or imputation. Weighting involves discarding the respondents with incomplete data and using weighting adjustments on the complete cases as in coping for total non-response. In the case of imputation the partial respondents are kept in the analysis and their missing observations are imputed with reference to similar patients providing complete data.

MISSING DATA MECHANISMS

In compensating for missing values it is vital that we bear in mind either our explicit or implicit assumptions concerning the way in which the missing data have arisen. Using the terminology of Little & Rubin (1987), the simplest assumption is that the missing observations are *missing completely at random* (MCAR). That is, there is no information which we have collected, or might have collected, which would enable us to predict who might have missing information. The novice might be tempted to shorten this description to label the missing data mechanism as missing at random. Little & Rubin, however, use the phrase *missing at random* (MAR) for situations in which data are missing at random, but conditional on the values of the non-missing observations. An example should clarify the distinction. Consider a 2-phase survey to validate a screening questionnaire. All 1st-phase respondents provide fallible screening information. Only a sub-sample of these (the 2nd-phase sample), however, are given a validation interview. If the 2nd-phase sample is chosen without reference to the screening (or any other) information, by the toss of a coin, for example, then the validation data are MCAR. An alternative (and much more commonly used) strategy is to interview a high proportion of the screen positives (say 80%) but only a small sub-sample (say 20%) of the screen negatives. Providing the sampling mechanism is random within strata (the outcome of the screening questionnaire) then this produces data that are MAR. If the probability of a missing validation interview is dependent on the subject's true psychiatric status, even after conditioning on the screening questionnaire outcome, then the missing data mechanism is referred to as being non-ignorable (Little & Rubin, 1987). In addition to any 2nd-phase sampling fractions chosen by the investigator, the probability of the selected subject agreeing to be interviewed is likely to be influenced by the severity or sub-type of their illness, for example. But this can only determined by the interview. The concept of ignorability as discussed by Little and Rubin (1987) is rather a difficult one, but readers should simply interpret it as a statistical mechanism that is *neither* MCAR *nor* MAR. Often, however, we will approximate reality by MAR — the justification for the weighting and imputation methods to be introduced below. MAR is a justifiable assumption for 2-phase designs since the missing data mechanism is, in fact, known — it is part of the design.

WEIGHTING AND IMPUTATION

«In most sample surveys, weights are attached to each respondent record and then used in analyses to produce approximately unbiased estimates of parameters of the target population. These weights compensate for the facts that sampled elements may be selected at unequal sampling rates and have different probabilities of responding to the survey, and that some population elements may not be included in the list or frame used for sampling. The main objective of weighting is to reduce bias in survey estimates by making each respondent represent a different fraction of the target population.» (Brick & Kalton, 1996). Essentially, one can ask «What proportion of each relevant sub-group within the target population provides non-missing observations?». The reciprocal of this proportion corresponds to the expansion or *probability weight* to be attached to these non-missing observations. If the non-missing fraction is one fifth, for example, then each actual observation is representing five potential observations. The probability weight is 5.

Pickles *et al.* (1995) provide an example of the results of a 2-phase psychiatric survey in Cantabria (Northern Spain). In the 1st phase a sample of consecutive primary care attenders was given a screening interview. Of the 514 screen negatives, 42 were subsampled for the 2nd-phase interview. The sampling fraction was 42/514 and the corresponding probability weight was therefore 514/42 = 12.238. Of the 309 screen positives, 161 were interviewed in the 2nd-phase. Here the sampling fraction was 161/309 and the corresponding probability weight was 309/161 = 1.919. When calculated separately for the two sexes, the probability weights for men were 12.611 and 2.366 for screen negatives and screen negatives, respectively. The corresponding weights for women were 11.958 and 1.767.

The use of these weights will be illustrated in the next section. Before moving on to the next section, however, we will briefly discuss one particular method of imputation. Essentially, imputation involves filling the gap (the missing observation) with an estimate based on a knowledge of the variables that have been observed. Some form of regression equation might be used, for example. Consider the particularly simple situation where the only information available is the subject's screen status together with their sex. Here one would like to impute missing interview responses (case or non-case) for four possible classes of subject (screen status crossed by sex) — the four *imputation classes*. One can replace the missing responses by the proportion of cases amongst the subjects with non-missing data in the appropriate imputation class. For example, the proportion of cases in the 2nd-phase screen positive men in the Spanish survey is 22/41 = 0.537. We can replace the missing values for caseness in those 1st-phase screen positive men who were not interviewed by the value 0.537, and similarly for the other imputation classes, and then proceed with the analysis as if we have no missing values. This *deterministic* imputation would yield unbiased estimates and is, in fact, equivalent to the above method of probability weighting. A subtle variation on this theme is *stochastic* imputation: to randomly select the value 1 for each missing observation (1 = case) with a probability of 0.537 (and similarly with the three other imputation classes) and again estimate the overall prevalence of cases. This is an example of what is known as *hot-deck imputation*. (Brick & Kalton, 1996). The similarities of hot-deck imputation and weighted estimation are discussed by Reilly & Pepe (1997) who recommend the weighting methods (as a simple method of carrying out the equivalent to hot-deck imputation) for their ease of use with standard software.

WEIGHTED ANALYSES IN PRACTICE

For the *i*th subject with complete data we determine a probability weight w_i. Consider the estimation of the prevalence of a psychiatric disorder, for example. For the complete-data subjects let $Y_i = 1$ if the subject is a psychiatric case, 0 otherwise. An estimate of the prevalence of disorder, π, is provided by the following ratio:

$$\pi = \Sigma\, w_i Y_i / w_i$$

This is the well-known Horvitz-Thompson estimator (see, for example, Lehtonen & Pahkinen, 1995). If one is using one of the common general-purpose statistical packages then one analyses the complete-data cases only and proceeds to calculate the appropriate weighted proportion of cases, having already declared the w_i as the required weight. But, in general, *the users of commercial statistical packages should take great care in the use of the*

weighting procedures which they might provide. The use of weights within most packages will give the correct estimates of prevalence, for example, but, unfortunately, neither the correct standard errors nor valid confidence intervals and significance tests. This arises from the fact that the weights are typically interpreted as *frequency weights* (an indicator of the number of observations with identical data to that provided in a given record). The package accordingly treats the ith subject in the complete-data sample as if it had actually been observed w_i times and accordingly produces P-values, standard errors and confidence intervals that are far too small. The appropriate use of a probability weight, on the other hand, recognises that the observation has only occurred once, but that the observed subject is representative of w_i subjects who might have provided observations. The software package *Stata* (StataCorp, 1997), for example, clearly distinguishes between probability and frequency weighting and therefore allows one to produce valid confidence intervals and so on. Technical details are beyond the scope of this editorial, but one subtle difference between the various appropriate *Stata* programs depends on whether the weights are regarded as fixed (as in the *Huber* procedures) or as random variables (as in the bootstrap methods, allowing for the weights to change with each bootstrap sample) — see Clayton *et al.* (1997). On the whole, however, this latter distinction seems to have little practical significance.

Generalisations of the weighting method to logistic regression and other modelling techniques are discussed in Binder (1983), Pickles *et al.* (1995) and Clayton *et al.* (1997). Note that the use of weighting is a general purpose method to compensate for missing data. On the whole it will perform reasonably well in most applications, but should not be regarded as the optimum strategy in all or even in most circumstances. Discussion of the relative performance of probability weighting and other estimation methods is discussed in detail, for example, by Robins *et al.* (1994), Breslow & Holubkov (1997), Clayton *et al.* (1997) and Schill & Drescher (1997). From the dates of the references cited in this editorial it should be obvious to the reader that missing value problems are currently an area of very active research.

REFERENCES

Breslow N.E. & Holubkov, R. (1997). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* 16, 103-116.

Binder D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 51, 279-292.

Brick J.M. & Kalton G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research* 5, 215-238.

Clayton D., Dunn G., Pickles A. & Spiegelhalter D. (1997). Analysis of longitudinal binary data from multiphase sampling (with discussion). *Journal of the Royal Statistical Society*, Series B (in press).

Gornbein J.A., Lazaro C.G. & Little R.J.A. (1992). Incomplete data in repeated measures analysis. *Statistical Methods in Medical Research* 1, 275-295.

Lehtonen R. & Pahkinen E.J. (1995). *Practical Methods for Design and Analysis of Complex Surveys*. Wiley & Sons: Chichester.

Little R.J.A. (1995). Modelling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association* 90, 1112-1121.

Little R.J.A. & Rubin D.B. (1987). *Statistical Analysis With Missing Data*. Wiley & Sons: New York & Chichester.

Pickles A. & Dunn G. (1998). Estimation of disease prevalence from screening data. In *Encyclopedia of Biostatistics* (ed. P. Armitage and T. Colton). Wiley & Sons: New York & Chichester.

Pickles A., Dunn G. & Vàzquez-Barquero J.L. (1995). Screening for stratification in two-phase ('two-stage') epidemiological surveys. *Statistical Methods in Medical Research* 4, 73-89.

Reilly M. & Pepe M. (1997). The relationship between hot-deck multiple imputation and weighted likelihood. *Statistics in Medicine* 16, 5-19.

Robins J.M., Rotnitsky A. & Zhao L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* 89, 846-866.

Schill W. & Drescher K. (1997). Logistic analysis of studies with two-stage sampling: a comparison of four approaches. *Statistics in Medicine* 16, 117-132.

StataCorp (1997). *Stata Statistical Software: Release 5.0*. Stata Corporation: College Station, Texas.