[Leeuw, Edith D. de, and Joop Hox. (2008). Missing Data. *Encyclopedia of Survey Research Methods*. Retrieved from http://sage-ereference.com/survey/Article_n298.html]

Missing Data

An important indicator of data quality is the fraction of missing data. Missing data (also called "item non-response") means that for some reason data on particular items or questions are not available for analysis. In practice, many researchers tend to solve this problem by restricting the analysis to complete cases through "listwise" deletion of all cases with missing data on the variables of interest. However, this results in loss of information, and therefore estimates will be less efficient. Furthermore, there is the possibility of systematic differences between units that respond to a particular question and those that do not respond—that is, item nonresponse error. If this is the case, the basic assumptions necessary for analyzing only complete cases are not met, and the analysis results may be severely biased.

Modern strategies to cope with missing data are *imputation* and *direct estimation*. Imputation replaces the missing values with plausible estimates to make the data set complete. Direct estimation means that all available (incomplete) data are analyzed using a maximum likelihood approach. The increasing availability of user-friendly software will undoubtedly stimulate the use of both imputation and direct estimation techniques.

However, a prerequisite for the statistical treatment of missing data is to understand why the data are missing. For instance, a missing value originating from accidentally skipping a question differs from a missing value originating from reluctance of a respondent to reveal sensitive information. Finally, the information that is missing can never be replaced. Thus, the first goal in dealing with missing data is to have none. Prevention is an important step in dealing with missing data. Reduction of item nonresponse will lead to more information in a data set, to more data to investigate patterns of the remaining item nonresponse and select the best corrective treatment, and finally to more data on which to base imputation and a correct analysis.

A Typology Of Missing Data

There are several types of missing data patterns, and each pattern can be caused by different factors. The first concern is the randomness or nonrandomness of the missing data.

Missing At Random Or Not Missing At Random

A basic distinction is that data are (a) missing completely at random (MCAR), (b) missing at random (MAR), or (c) not missing at random (NMAR). This distinction is important because it refers to quite different processes that require different strategies in data analysis.

Data are MCAR if the missingness of a variable is unrelated to its unknown value and also unrelated to the values of all other variables. An example is inadvertently skipping a question in a questionnaire. When data are missing completely at random, the missing values are a random sample of all values and are not related to any observed or unobserved variable. Thus, results of

data analyses will not be biased, because there are no systematic differences between respondents and nonrespondents, and problems that arise are mainly a matter of reduced statistical power. It should be noted that the standard solutions in many statistical packages, those of listwise and pairwise deletion, both assume that the data are MCAR. However, this is a strong and often unrealistic assumption.

When the missingness is related to the observed data but not to the (unknown) value of the missing response itself, it is said that the data are MAR. For example, an elderly respondent may have difficulty recalling an event because of memory problems. The resulting missing datum is related to age but not to the event itself. When the data are missing at random, the missingness is a random process conditional on the observed data. If the data are missing at random *and* if the proper statistical model is used, the missingness is said to be *ignorable* with respect to inference. For example, in the case of the elderly respondent, the variable related to the missingness (age) is measured and available for inclusion in the proper analysis.

Finally, when the missingness is related to the unknown (missing) answer to the question itself, the data are NMAR. For example, a respondent perceives the real answer to a sensitive survey question as socially undesirable (e.g. she or he does have drinking problems) and refuses to respond. If the missing data are the NMAR type, the missingness is said to be *nonignorable*, and no simple solution for treating the missing data exists. A model for NMAR missingness must be postulated and included in the analysis to prevent bias.

Missing Data Patterns

Three main patterns can be discerned in item missing data: (1) the data are missing systematically by design (e.g. contingency questions); (2) all the data are missing after a certain point in the questionnaire (partial completion); and (3) data are missing for some questions for some respondents (item nonresponse).

Missing By Design

Data are missing by design when the researcher has decided that specific questions will not be posed to specific persons. There are two main reasons for items to be missing by design. First, certain questions may not be applicable to all respondents and the questionnaire routing skips these questions for these respondents, that is, these are contingency questions. Since the responses to other questions determine the missingness, the missingness mechanism is accessible to the analyst and can be incorporated in the analyses.

The second reason for items to be missing by design is when a specific design is used to administer different subsets of questions to different persons. In this case, all questions are applicable to all respondents, but for reasons of efficiency not all questions are posed to all respondents. Specific subsets of questions are posed to different groups of respondents, often following a randomized design in an experiment (i.e. random assignment) that makes the missingness mechanism MCAR. Again, since the missingness mechanism is accessible, the incomplete data can be handled statistically and the analyses give unbiased results.

Partial Completion

A partial completion (breakoff) is characterized by time or place dependency. After a certain point in time or place within the questionnaire, *all* data are missing. Partial completions mostly occur in telephone interviews and Web surveys. At a certain time point in the interview, the respondent stops and disconnects. As a result, the remainder of the questionnaire is not

answered. When the breakoff occurs early in the questionnaire and only a few questions have been answered, it is usually treated as unit nonresponse. When the breakoff occurs at the end of the questionnaire, the remaining unanswered questions are usually treated as item nonresponse. In that case, information on earlier questions and the interview process is used to investigate the missingness mechanism and adjust for it in the analyses.

Item Nonresponse

Item nonresponse is characterized by blanks in the data for some respondents on some variables. Not every blank in the data matrix originates in the same way. One can distinguish three forms of item non-response: (1) the information is not provided by a respondent for a certain question (e.g. a question is overlooked by accident, an answer is not known, a refusal to respond); (2) the information provided by a respondent for a certain question is not usable (e.g. a given answer is not a possible answer, it falls outside the range of permissible responses, multiple responses are given when only one is allowed, it cannot be coded, and/or it is unreadable/illegible); and/or (3) usable information is lost (e.g. error in data entry or data processing). The first two of these mechanisms (information is not provided and information is not usable) originate in the data collection phase. The third is the result of errors in the data processing phase.

The most problematic form of item nonresponse occurs when a respondent does not provide information, because in this case different missing data mechanisms may be at work. When the respondent accidentally overlooks an item, the data are MCAR. The missingness mechanism is ignorable and almost all simple statistical treatments may be used, even list-wise deletion. When a respondent is willing but unable to respond—for example, because of memory problems—the missingness depends on an observed variable (age), but not on the answer to the question itself and is thus missing at random. If the data are MAR and if the variable related to the missingness is available, the missingness can be handled adequately with relatively simple solutions. However, when not responding is related to the (unknown) answer to the question itself, the missingness mechanism is NMAR. When a respondent *refuses* to respond, the missingness is probably NMAR and the mechanism is non-ignorable. In this case, simple solutions no longer suffice, and an explicit model for the missingness must be included in the analysis.

When item nonresponse is due to unusable responses that are coded as missing, it is generally problematic. The reasons for inadequate responses (e.g. outside the range of possible answers or nonsubstantive responses) are related to the question format and the real value of the answer, pointing to NMAR. If the real answer is partly revealed (e.g. through interviewer notes), the missingness mechanism is at least partly known.

Finally, losing information because of errors in coding, editing, or storing is usually not systematic and therefore normally MCAR. It arises by accident and is not related to questionnaire and respondent characteristics, so the mechanism is ignorable and the solutions are simple.

Analyzing Incomplete Data Sets

Inspecting The Structure And Patterns Of Missing Data

For an optimal treatment of item nonresponse, knowledge of the missing data mechanism is valuable. First, one should investigate whether the data are MCAR or not. When incomplete data are MCAR, analyses will not be biased, because there are no systematic differences between

respondents who completed the question and respondents who have a missing value for that question.

The first step in the analysis of incomplete data is to inspect the data. This can provide very practical information. For instance, one may find that most of the missing values concern only one specific variable (e.g. household or personal income). But if that variable is not central to the analysis, the researcher may decide to delete it. The same goes for a single respondent with many missing values. In general, however, missing values are scattered throughout the entire data matrix. In that case, a researcher would like to know if the missing data form a pattern and if missingness is related to some of the observed variables. If one discovers a system in the pattern of missingness, one may include that in the statistical analyses or imputation procedures.

The mere inspection of missing data patterns cannot tell the researchers with certainty whether or not the missingness is independent of the (unknown) value of the variable (question). Extra information is needed to test the MAR hypothesis and help to determine the causes of item nonresponse. This information may be available in the data set, but often additional information (information from other sources than the actual sample) is needed, such as theory, logic, or auxiliary data from registers, sampling frames, reinterviews, or other special nonresponse studies.

Effective Methods To Analyze Incomplete Data Sets

The default options of statistical software are usually listwise or pairwise deletion or some simple imputation technique such as *mean substitution*. These solutions are generally inadequate. Listwise deletion removes all units that have at least one missing value and is clearly wasteful because it discards information. Pairwise deletion removes cases only when a variable in a specific calculation is missing. It is less wasteful than listwise deletion, but it can result in inconsistent correlation matrices in multivariate analyses, because different elements in the correlation matrix may be based on different subsamples. Simplistic imputation techniques (e.g. mean substitution) often produce biased point estimates and will always underestimate the true sampling variances. Listwise and pairwise deletion and simple imputation are likely to be biased, because these methods are all based on the strong assumption of MCAR, which seldom is warranted. Therefore, the best policy is to prevent missing data as much as possible, and when they occur to employ an analysis strategy that uses (a) all available information to investigate the missing data patterns and (b) an analysis method that correctly adjusts for missing data.

Only when the data can be considered MCAR do simple solutions like listwise deletion not result in bias. If the fraction of missing data is small, listwise deletion is useful. If the fraction of missing data is large, the MAR-based techniques described following are more efficient.

When the data are assumed MAR, two distinct analysis approaches can be used: direct estimation and imputation.

Direct Estimation

Direct estimation means that the incomplete data are fully analyzed using a maximum likelihood approach. Direct estimation requires specialized software, but this is increasingly becoming available. For instance, several programs for structural equation modeling can include incomplete cases in the analysis. Since analysis of (co)variance, multiple regression analysis, and discriminant analysis can all be formulated as a structural equation model, these analyses can now be done using all available information, under the assumption of MAR. Another example is

using multi-level models for incomplete longitudinal data. Such analyses view the repeated measures as hierarchically nested within cases. Since multi-level models do not assume that all measurement occasions are available for analysis, missing data due to panel dropout (attrition) are not a problem.

While direct estimation is powerful, it requires access to and knowledge of specialized software. Imputation fills the gaps in the data set with plausible values, and after the data are made complete, standard software then is used. At this point, the researcher can simply ignore the missingness problem and proceed to analyze the completed data set using any standard method with which she or he is familiar.

Imputation

In imputation, the missing values are replaced by "plausible" values. Many imputation methods exist, which mainly differ in the way they define *plausible*. A problem is that most simple imputation methods, such as replacing missing values with the overall mean or using regression to estimate the missing values, result in biased estimates. However, the popular and reasonably simple *hot-deck* method results in unbiased estimates under the assumption of MAR. In the hot-deck method, the data file is sorted into a number of imputation classes according to a set of auxiliary variables. Missing values are then replaced by observed values taken at random from other respondents in the same imputation class.

There are two fundamental problems associated with imputation. First, using the information in the observed data to predict the missing values emphasizes the structure in the completed data. Second, analyzing the completed data set uses a spuriously high number of cases and thus leads to biased significance tests. Donald Rubin proposes to solve both problems by using *multiple imputation:* Each missing value is replaced by two or more (M) plausible estimates to create *M* completed data sets. The plausible values must include an error term from an appropriate distribution, which solves the problem of exaggerating the existing structure in the data. Analyzing the *M* differently completed data sets and combining the estimates into an overall estimate solves the problem of the biased significance test.

In the multiple imputation approach, analyzing *M* data sets and having to combine the results is cumbersome but not especially complex. What is difficult is generating the *M* data sets in a proper manner. A non-parametric method is to (a) compute for each respondent the propensity to have missing values on a specific variable, (b) group respondents into imputation classes based on this propensity score, and (c) use hot-deck imputation with these imputation classes. Parametric imputation methods assume a model for the data and use Bayesian methods to generate estimates for the missing values. These methods are described in detail by Joseph L. Schafer. When multiple imputation is used, it is important that the model for the data generation is very general and includes those variables that are important for predicting either missingness or the variables of interest.

Further Readings

Arbuckle, J. L. (1996). *Full information estimation in the presence of incomplete data.* In G. A. Marcoulides , ed. & R. E. Schumacker (Eds.), *Advanced structural equation modeling (pp. 243–277)* . Mahwah, NJ: Lawrence Erlbaum.

de Leeuw E. D. , Hox J. J. , and Huisman M. *Prevention and treatment of item nonresponse. Journal of Official Statistics* vol. 15 (2003) pp. 153–176. Retrieved April 14, 2008, from http:/ / www.jos.nu

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Lawrence Erlbaum.

Little, R. J. A. , & Rubin, D. B. (1987). *Statistical analysis with missing data.* New York: Wiley.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer, J. (1997). *Analysis of incomplete multivariate data.* London: Chapman & Hall.

Schafer J. and Olsen M. K. *Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavior Research* vol. 33 (1998) pp. 545–571.