Missing Data Analysis

Goals

- Understand definition of missing data and what constitutes a missing data problem.
- Understand pros and cons of simple missing data methods.
- Understand the definition of missing at random, and its implications for statistical analysis.
- Understand basic principles of weighting and imputation methods.
- Understand R computational tools for missing data.

Missing data patterns

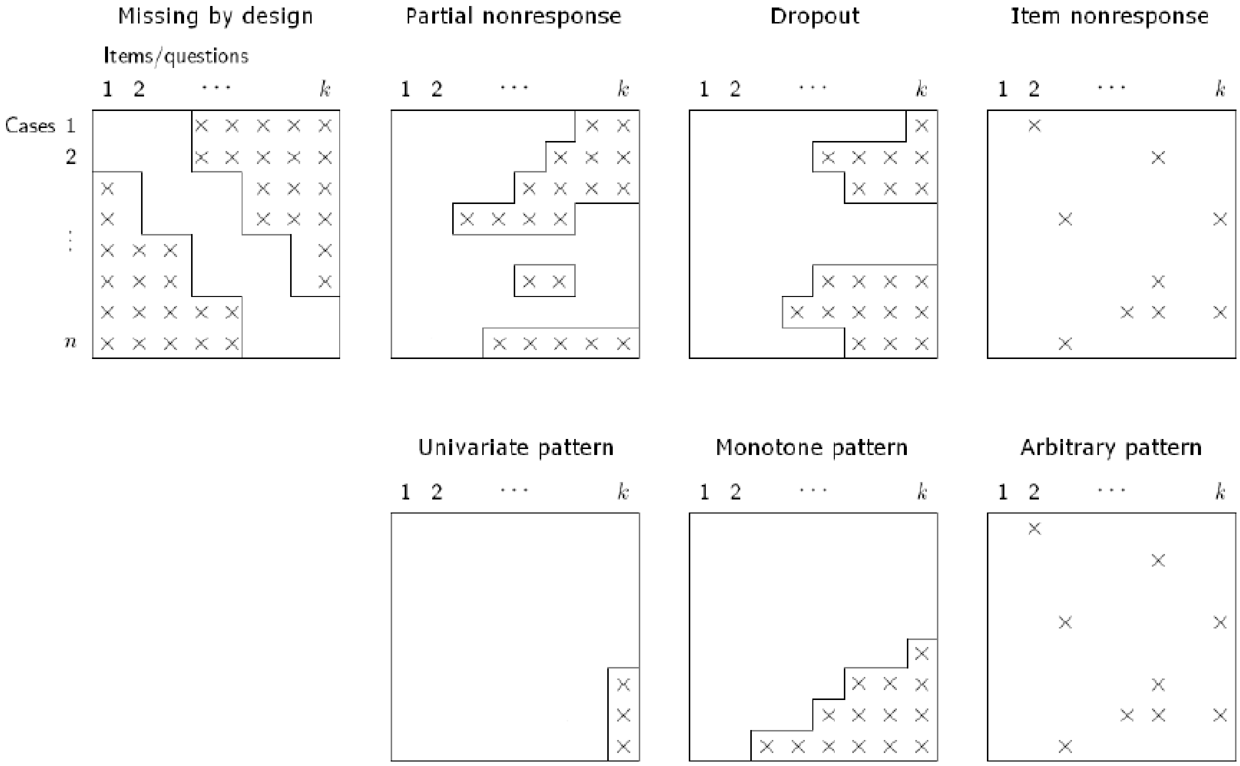

Figure 1: Diagram illustrating missing data patterns

Missingness Mechanisms

Let Y be the data matrix. Let Y_{obs} be the observed components of Y and let Y_{mis} be the missing components of Y. Let R be a missing data indicator matrix where the (i, j)th element of R is 1 if the corresponding element of Y is missing and 0 if it observed. In other words, R is a missing data indicator matrix where each element in Y is replaced by either 0 or 1 depending on whether it is observed or missing in Y.

The missing data pattern concerns the distribution of R whereas the missing data mechanism concerns the distribution of R given Y. That is, the missing data pattern concerns which data are missing whereas the missing data mechanism concerns *why* the data are missing.

Missing Completely at Random

Missing data are missing completely at random (MCAR) if the missing data (and hence the observed data) can be regarded as a simple random sample of the complete data.

The probability that a data value is missing (missingness) is unrelated to the data value itself or to any other value, missing or observed, in the data set.

MCAR if missingness is independent of Y:
$P(R|Y) = P(R)$ for all Y

Missing at Random

If missingness is related to the observed data but - conditioning on the observed data - not to the missing data, then data are missing at random (MAR).

MAR if missingness only depends on observed components Y_{obs} of Y
$P(R|Y) = P(R|Y_{obs})$ for all Y

MAR if dropout depends on values recorded prior to drop-out.

MCAR is a stronger condition and a special case of MAR.

Missing Not at Random

If missingness is related to the missing values themselves even when the information in the observed data is taken into account, then missing data are missing not at random (MNAR).

- If conditional on all of the observed data, individuals with higher incomes are more likely than others to withhold information about their incomes, then the missing income data are MNAR.
- MNAR if dropout depends on values that are missing (that is, after drop-out).
- MNAR if missingness depends on missing (as well as perhaps on observed) components of Y.

If the data are MCAR or MAR then it is not necessary to model the process that generates the missing data in order to accommodate the missing data. The mechanism that produces the missing data is *ignorable.*

When data are MNAR, the missing-data mechanism is *non-ignorable.* It is necessary to model this mechanism to deal with the missing data in a valid manner.

Except in some special situations (e.g., when missing data are missing by design), it is not possible to know whether data are MCAR, MAR, or MNAR.

Showing that missingness on some variable in a data set is related to observed data on other variables, proves that the missing data are not MCAR.

Demonstrating that missingness on a variable is not related to observed data on other variables does not prove that the missing data are MCAR. Non-respondents may be different from respondents in some unobserved manner.

Generally Inappropriate Strategies for Dealing with Missing Data

Complete-case analysis

In complete-case analysis, only observations containing valid data values on every variable are retained for further analysis. Practically, this involves deleting any row with one or more missing values, and is also known as *listwise*, or *case-wise*, deletion. Deleting all observations with missing data can reduce statistical power by reducing the available sample size.

Loss of information in incomplete cases has two aspects:
– Increased variance of estimates
– Bias when complete cases differ systematically from incomplete cases (i.e., respondents may differ from nonrespondents)

Complete-case analysis is biased if drop-outs differ from non-drop-outs.

Assumption: MCAR; that is, complete-case analysis provides consistent estimates and valid inferences only when the missing data are MCAR.

Generally inappropriate.

Pairwise deletion

Pairwise deletion (also called available-case analysis) is often considered an alternative to listwise deletion when working with datasets that have missing values. In pairwise deletion, observations are deleted only if they are missing data for the variables involved in a specific analysis.

Although pairwise deletion appears to use all available data, in fact each calculation is based on a different subset of the data. This can lead to distorted and difficult to interpret results.

By basing different statistics on different subsets of the data, available-case analysis can lead to nonsensical results, such as correlations outside the range from -1 to +1 (called non-positive definite covariance matrices).

I recommend staying away from this approach. I only mention it because people still do it.

Last observation carried forward

LOCF is an imputation strategy for repeated measures with dropouts:
– impute using the last recorded value
– implicit model: values are unchanged after drop-out

LOCF is mistakenly considered to be valid under MCAR or MAR, but it is **generally inappropriate**.

Unconditional mean imputation

In *unconditional mean imputation*, the missing values in a variable are replaced with the mean of that variable (missing values of categorical variables can be replaced by the mode for that variable).

Using *mean substitution* is likely to underestimate standard errors, distort correlations among variables, and produce incorrect p-values in statistical tests. Mean imputation preserves the means of variables, but it makes their distributions less variable and tends to weaken relationships between variables. It produces biased results for data that are not MCAR.

By treating the missing data as if they were observed, mean imputation exaggerates the effective size of the data set, further distorting statistical inference - a deficiency that it shares with other single imputation methods. The substitution is nonstochastic, meaning that random error is not introduced (unlike with multiple imputation).

I recommend avoiding this approach for most missing-data problems.

Conditional mean imputation

Conditional-mean imputation replaces missing data with predicted values, obtained, for example, from a regression equation (thus, it is sometimes called regression imputation).

The imputed observations tend to be less variable than real data, because they lack residual variation.

Another problem is that we have failed to account for uncertainty in the estimation of the regression coefficients used to obtain the imputed values.

Regression imputation improves on unconditional mean imputation, but it still produces biased results for data that are not MCAR.

The first problem with regression imputation (lack of residual variation) can be addressed by adding a randomly sampled residual to each imputed value, in which case it is referred to as stochastic conditional mean imputation.

The second problem (uncertainty in the regression coefficients used for prediction of missing data) leads naturally to Bayesian multiple imputation of missing values.

Principled Approaches to Missing Data

Inverse Probability Weighting

Assigns a missingness (also called a non-response) weight to the complete cases to make them more representative of all cases (including those that are missing). Standard errors can be estimated analytically or using the bootstrap.

This approach is primarily used for dropout (i.e., the univariate or monotone missingness pattern). It is not as useful for the arbitary missingness pattern.

Analyze the incomplete data

Methods can be applied to incomplete data – that is, there are methods that do not require rectangular (or balanced) data sets. This is often accomplished using full information maximum likelihood (FIML) and the EM (Expectation Maximization) algorithm.

Likelihood inference ignoring the missing data mechanism is valid if the
– Model for Y is correctly specified
– Mechanism is MAR (or MCAR)

This approach works if there is missingness only on the outcome variable. This approach cannot handle missingness on the predictor variables.

Multiple Imputation (MI)

Bayesian multiple imputation (MI) is a flexible and general method for dealing with data that are MAR.