Primer on Missing Data

BC Office of Patient-Centred Measurement & Improvement

What is Missing Data?

Missing data are planned observations that are not recorded so their values are unknown. Missing data in a survey may occur when there are no data for a person (unit non-response) or when some answers for a respondent are unknown (item non-response).

Why is Missing Data a Problem?

Missing responses affect the representativeness of your data and can be a source of bias in group- or population-based estimates of prevalence (point estimates) and group comparisons. Missing data can lead to biased point estimates even when no hypothesis tests are conducted. This in turn can lead to wrong conclusions. Missing data can also be problematic when running hypothesis tests as many procedures require complete data for each respondent.

Type of Missing Data

- **Unit non-response** occurs when an eligible survey respondent does not participate in the survey or does not provide enough information to be deemed usable.
  Example: Respondents refused to participate when surveyors phoned to remind them to complete the sector survey.
- **Item non-response** occurs when a respondent does not answer a specific survey item because of refusal or other reasons.
  Example: Respondents refused to answer a question and their answers were recorded as "prefer not to answer" by the telephone surveyor.
- **Planned missing data** occurs when the data are not collected by design (e.g., skip questions, randomly presented items, pilot testing of instrument with a subset of respondents).
  Example: Respondents were not asked questions about their experience with their Aboriginal Liaison Person because they did not self-identify as an Aboriginal person in an earlier question.
- **Human generated non-response** occurs when the data are missing because of human errors such as data entry errors or programming errors.
  Example: Respondents completing an online survey could only provide a single response to a question that was supposed to allow multiple responses due to a programming error.

When does Missing Data Matter?

- The higher the percentage of missing data, the more likely that results obtained based on statistical analyses will be biased.
- The reason for the data missing is related to the missing value or related variables (i.e., the missingness is not arbitrary or random).
- Missing data can be problematic when hypothesis testing because complete case analysis (analyzing only cases that have no missing data) results in a reduction of sample size, which will lead to biased point estimates and reduced statistical power (the long-run probability of detecting a statistically significant result when, in fact, a statistically significant result exists in reality).

How to assess Missing Data?

- Rule out other plausible explanations (e.g., data entry error, computer error, planned missing data, missing data due to a factor that was not captured in the survey responses).
- Document the number and percentage of missing data.
- Examine the pattern of missing data (e.g., whether certain values tend to be missing together).
- Compare demographics and related questions between respondents with and without missing data.
- Conduct sensitivity analyses, such as comparing results computed using different approaches.

How to handle Missing Data?

All approaches to handling missing data are designed to minimize bias when drawing conclusions about a group of respondents. Though some approaches can be used at the item level, imputed values are only "predicted values" that cannot replace unknown values with complete certainty. The best treatment of missing data is to minimize them by executing a well-planned survey.

It is always important to keep in mind that methods for accommodating missing data such as multiple imputation can also lead to greater standard errors and thus reduced statistical power despite the increase sample size.

Approach	Description	Advantage	Disadvantage
Complete Case analysis - listwise deletion	Respondents with data missing on any variables are dropped from all analyses. This is often the default setting in most software.	Simple to implement, sample size is consistent across analyses.	Reduces sample size, lowers statistical power, and will produce biased estimates when missingness is not completely at random.
Available case analysis - pairwise deletion	Respondents with data missing on any variables are dropped from a single analysis.	Sample size for each individual analysis is generally higher than complete case analysis. Pairwise deletion can be defensible in multivariate analysis with many variables.	Results in different sample size for each analysis, will produce biased estimates when missingness is not completely at random, and can also lead to mathematical problems in computing estimates of some parameters.
Single Imputation	Simple imputation substitutes missing values with the mean (average), median (middle), or modal (most frequently occurring) values. Conditional imputation predicts missing values using responses from related questions (typically done in a regression model).	Simple to understand and implement, and sample size is consistent across analyses.	Reduces response variation, reduces standard error, and weakens relationship between variables. Conditional imputation can overestimate relationships and model fit. In some cases, mean imputation can resulted in bias that is worse than with complete case analysis.
Multiple Imputation	Create multiple copies of data with imputed values and pool the analyzed results into a single estimate.	Can yield unbiased estimates and standard errors even when data is missing at random. Accounts for error and variations from estimating the imputed values.	Requires more complex computations and expert guidance. Imputed values can depend on selection of relevant variables used to predict the missing values.
Alternate Estimators	Use semiparametric or maximum likelihood estimation to obtain less biased estimates. Maximum likelihood estimates values that are most likely to have resulted given the observed data. This method does not impute any data.	Can yield unbiased estimates and standard errors even when data is missing at random.	Standard errors may be too small and not available in all procedures. Expert guidance is recommended. Maximum likelihood estimator also does not produce an imputed value.

Missing Data Mechanism

Determining the mechanism that caused the missing responses is an important consideration when deciding on how to handle missing data. When in doubt, consult with a statistician for recommendations.

Missing Mechanism	Description	Example	Recommended Actions
Missing Completely at Random (MCAR)	The data are missing for completely random reasons. There is no relationship between whether a data point is missing and any values in the data set, missing or observed. The missing data are just a random subset of the data.	Participants' responses to a question about their overall health care experience were not recorded because of a data entry error.	Where possible, complete case analysis (listwise deletion) should be avoided as it is most likely to produce biased results. Use any of the other approaches, keeping in mind their advantages and disadvantages. Compare results obtained from the chosen approaches against results from analyses with missing data to see if there are differences.
Missing at Random (MAR)	The reason or propensity for a data point to be missing is unrelated to the missing data, but it is related to some of the observed data. Whether or not a respondent answered a question has nothing to do with the missing value, but it does have to do with the values of other variables.	Participants were less likely to answer a question about their overall health care experience if they felt that they weren't treated with compassion and respect by care staff (another related question in the survey).	Use multiple imputation, alternate estimators or ask a statistician.
Not Missing at Random (NMAR)	The reason or propensity for a data point to be missing is directly related to the missing data.	Participants were less likely to answer a question about their overall health care experience if they had a poor care experience.	Ask a statistician.