StatNews #46

Missing Data Mechanisms
October 2001

As almost any researcher can attest, missing data are a widespread problem. Data from surveys, experiments, and secondary sources are often missing some data. The impact of the missing data on the results of statistical analysis depends on the mechanism which caused the data to be missing and the way in which the data analyst deals with it. This is the first in a series of four newsletters that discusses issues surrounding missing data. This newsletter outlines the mechanisms of missing data and some of their impacts. Subsequent newsletters will explain common but problematic solutions to missing data, new and better solutions, and the software available for implementing these solutions.

Data are missing for many reasons. Subjects in longitudinal studies often drop out before the study is completed because they have moved out of the area, died, no longer see personal benefit to participating, or do not like the effects of the treatment. Surveys suffer missing data when participants refuse, or do not know the answer to or accidentally skip an item. Some survey researchers even design the study so that some questions are asked of only a subset of participants. Experimental studies have missing data when a researcher is simply unable to collect an observation. Bad weather conditions may render observation impossible in field experiments. A researcher becomes sick or equipment fails. Data may be missing in any type of study due to accidental or data entry error. A researcher drops a tray of test tubes. A data file becomes corrupt. Most researchers are very familiar with one (or more) of these situations.

Missing data are problematic because most statistical procedures require a value for each variable. When a data set is incomplete, the data analyst has to decide how to deal with it. The most common decision is to use complete case analysis (also called listwise deletion)--analyzing only the cases with complete data. Individuals with data missing on any variables are dropped from the analysis. It has advantages--it is easy to use, is very simple, and is the default in most statistical packages. But it has limitations. It can substantially lower the sample size, leading to a severe lack of power. This is especially true if there are many variables involved in the analysis, each with data missing for a few cases. It can also lead to biased results, depending on why the data are missing.

All of the causes for missing data fit into four classes, which are based on the relationship between the missing data mechanism and the missing and observed values. These classes are important to understand because the problems caused by missing data and the solutions to these problems are different for the four classes.

The first is Missing Completely at Random (MCAR). MCAR means that the missing data mechanism is unrelated to the values of any variables, whether missing or observed. Data that are missing because a researcher dropped the test tubes or survey participants accidentally skipped questions are likely to be MCAR. If the observed values are essentially a random sample of the full data set, complete case analysis gives the same results as the full data set would have. Unfortunately, most missing data are not MCAR.

At the opposite end of the spectrum is Non-Ignorable (NI). NI means that the missing data mechanism is related to the missing values. It commonly occurs when people do not want to reveal something very personal or unpopular about themselves. For example, if individuals with higher incomes are less likely to reveal them on a

survey than are individuals with lower incomes, the missing data mechanism for income is non-ignorable. Whether income is missing or observed is related to its value. Complete case analysis can give highly biased results for NI missing data. If proportionally more low and moderate income individuals are left in the sample because high income people are missing, an estimate of the mean income will be lower than the actual population mean.

In between these two extremes are Missing at Random (MAR) and Covariate Dependent (CD). Both of these classes require that the cause of the missing data is unrelated to the missing values, but may be related to the observed values of other variables. MAR means that the missing values are related to either observed covariates or response variables, whereas CD means that the missing values are related only to covariates. As an example of CD missing data, missing income data may be unrelated to the actual income values, but are related to education. Perhaps people with more education are less likely to reveal their income than those with less education.

A key distinction is whether the mechanism is ignorable (i.e., MCAR, CD, or MAR) or non-ignorable. There are excellent techniques for handling ignorable missing data. Non-ignorable missing data are more challenging and require a different approach. Subsequent newsletters will discuss these techniques.

Author: Karen Grace-Martin

Back to StatNews Table of Contents