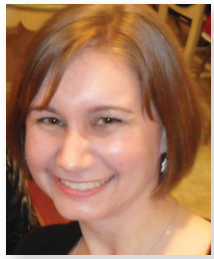
STATISTICALLY Speaking



Mi??ing Data: Should We C?re?

Across all fields of research, the issue of missing data arises. In most studies, surveys, or experiments, there are instances of nonresponse that need to be appropriately dealt with to conduct reasonable statistical analyses. There are several types of missing data as well as ways to deal with incomplete data sets.

First of all, what is meant by missing data? Missing data are simply unobserved values in a data set, but they can be of different types and may be missing for different reasons. Consider a survey that asks several questions of a participant. Maybe the participant feels uncomfortable answering a question regarding salary and leaves that question blank. This illustrates an instance of item nonresponse. Now consider a second participant who fails to return the survey at all. Because none of the questions were answered, this is an example of unit nonresponse. In longitudinal studies, dropout is another common cause of missing data.

Data may be missing according to different mechanisms of missingness. The first type of missing data mechanism is "missing completely at random" (MCAR). Statistically speaking, this means that the probability of the observation being missing is not dependent on any observed or unobserved covariates. In other words, it is what one typically thinks of as "random." Perhaps a computer malfunction erases part of a data set. That missingness is not dependent on any of the covariates or the responses measured in the study.

The second type of missing data mechanism is "missing at random" (MAR). In this case, the probability of the observation being missing is dependent only on observed values; that is, it is conditionally missing. A simple example of MAR is a survey where participants older than a certain age refuse to answer a particular survey question and age is measured in the study. The missingness is dependent on the observed covariate of age so the data are MAR.

The final type of missing data mechanism is "missing not at random" (MNAR). In this instance, the probability of an observation being missing is dependent on the missing responses or an unobserved covariate. An example of MNAR could be in a study measuring cognition. A participant's cognition might decline so low that it is necessary for the participant to move to a home and drop out of the study. In this case, the missing values are the result of the unobserved response values and so the data are MNAR.

Why should we care about missing data? Why not just take the complete cases at hand to analyze the data? Such an approach, which is the default in most statistical software, can have a drastic impact on the statistical inference drawn from the data. This phenomenon is easy to see in the following example. In a study testing a new drug, 90 of 100 participants drop out of the study because of lack of efficacy of the drug. If only the 10 participants who did not drop out are considered, the study would show that the drug is effective, even though it was only so for at best 10% of the study sample. This is an extreme and impractical example, surely, but the point remains that by eliminating the participants who dropped out, bias is introduced into the analysis of the data. In a less extreme example, this bias might tip the scales from one conclusion to another and leave the researcher with an incorrect inference.

There are several other reasons why missing data should not simply be ignored. Using only complete cases reduces sample size and, therefore, reduces power. In addition, it is costly to obtain data. To completely remove a participant with any item nonresponse would be to waste the potentially useful observed data and the money used to obtain it.

Despite the issues caused by only using complete cases, researchers use this method for data analysis regularly. Or, more likely, they simply do not know how to handle missing data and leave it to the default settings of statistical packages. Harel et al.[1] illustrated this phenomenon in a literature review of missing data in HIV prevention studies. They found that in 57 randomized controlled trials, none mentioned the missing data assumptions used in their analyses. Most of the studies (74%) used complete case analysis (CCA) where only the complete cases are used. Under relaxed assumptions, the authors expect only 12% of the studies to yield unbiased results, based on their analysis methods. Not only is CCA problematic, but it is also being used quite frequently in the literature.

What if there is only a small amount of missing data? Belin[2] addressed the assertion that small amounts of missingness can matter with a simulation study of a randomized controlled study with 5% missing values. Analyzing the full sample with no missingness or with an appropriate missing data procedure yielded a significant treatment effect. However, analyzing the data with missingness according to CCA yielded a nonsignificant treatment effect.

Even 5% missing values can lead to an erroneous conclusion.

So, what can be done? Little and Rubin[3] have discussed a wide variety of methods ranging from CCA and other ad hoc procedures to those grounded in mathematical foundations. Ad hoc procedures refer to CCA and single imputation methods that impute, or fill in, plausible missing values. These methods include "last observation carried forward," imputing unconditional means, imputing from unconditional distribution, and conditional mean imputation. The benefit to these methods is their ease of implementation, but the drawbacks can include bias, distorted correlations, and reduced variance. Other methods that are more mathematically based include multiple imputation, maximum likelihood, and Bayesian methods. These approaches are more challenging to implement but solve many of the problems seen in CCA and single imputation.

Little and Rubin also provided recommendations for which imputation method should be utilized for each type of missing data. When the data are believed to be MCAR, CCA is generally considered to be a reasonable approach. However, situations in which data are truly MCAR are rare, and this method still presents problems with efficiency. The recommendation when data are MAR is to use likelihood-based approaches, Bayesian analysis, or multiple imputation. These methods perform well under the MAR assumption and are often used under the MCAR assumption as well. Data that are MNAR are primarily concerns in clinical trials and in sensitivity analyses. In this instance, Little and Rubin recommend using either selection models or pattern–mixture models.

Missing data research is a rich field with endless applications. The problem of missing data is one that plagues everyone from biologists and chemists to psychologists and linguists. If dealt with inappropriately, the researcher might arrive at invalid inferences. So remember: the data might be missing, but the importance of dealing with missing data is always present. ■

Ofer Harel, PhD
Jennifer Boyko, MS

About the Authors

Both authors are with the Department of Statistics at the University of Connecticut, Storrs, CT. Ofer Harel is also with the Center of Intervention and Prevention, University of Connecticut.

Correspondence should be sent to Ofer Harel, PhD, Department of Statistics, University of Connecticut, 215 Glenbrook Rd Unit 4120, Storrs, CT 06269-4120 (e-mail: ofer.harel@uconn.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted May 12, 2012.
doi:10.2105/AJPH.2012.300904

Contributors

Both authors contributed equally to the creating of the article.

Acknowledgments

This project was partially supported by the National Institute of Mental Health (award K01MH087219).

Note. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Mental Health or the National Institutes of Health.

References

1. Harel O, Pellowski J, Kalichman S. Are we missing the importance of missing values in HIV prevention randomized clinical trials? Reviews and recommendations. *AIDS Behav.* 2012; 16(6):1382–1393.

2. Belin TR. Missing data: what a little can do, and what researchers can do in response. *Am J Ophthalmol.* 2009;148(6):820–822.

3. Little RJA, Rubin DB. *Statistical Analysis With Missing Data.* New York, NY: Wiley; 2002.