Outline

Many statistical analyses will have to contend with the fact that some of the data are missing. For instance, individuals drop out of clinical trials before their outcomes can be measured. In meta-analyses, missingness comes in many forms. Studies may not report the information required to extract effect estimates, variances, or covariates. Some studies may not report all of the effects they studied for a variety of reasons, including because those effect estimates were not statistically significant. Entire studies may even be relegated to the metaphorical file drawer, and hence their effect estimates would be missing from a meta-analylsis.

The question of how best to handle a missing data problem has no single answer. Instead, answers will depend on how much data are missing, why it is missing, and what sort of analyses are to be conducted. An analysis method with missing data that is appropriate under one set of conditions may be wholly inappropriate under another set of conditions. It is critical, then, that analysts have some understanding of their missing data problem.

In this tutorial, we review some standard missing data concepts and discuss their implications for meta-analysis. We then walk through a series of exploratory analyses using the R programming languate that can shed greater light on missingness. Finally, we conclude with a discussion of potential corrections for missing data, their use in meta-analysis, and some practical guidelines for reporting missingness in a meta-analysis.

1. **Describe data and prior analysis of these data.**

As part of this tutorial we use data analyzed by Tanner-Smith et al. (2016) on the effects of substance abuse inverventions for adolescents on subsequent substance use. These data were extracted from 61 randomized trials and quasi-experiments, and include 95 different effects of or contrasts between interventions. These effects include contrasts between a given treatment condition and a control condition within a study, or between two different treatement conditions in the same study.

There are a range of intervention types and venues that have been studied on individuals who use different substances and who differ in a variety of ways. For instance, interventions might focus on cognitive behavioral therapy (CBT), family therapy, or pharmacological therapy. Interventions could be in- or out-patient. Individuals in studies might present using marijuana, which is most common among adolescents, or alcohol or opioids. They may come from wealthy families or poor families. Thus, Tanner-Smith et al. fit a series of meta-regression models to examine how treatment effects varied according to the type of therapies and individuals studied. They found that assertive continuing care (ACC), behavioral therapy, CBT, motivational enhancement therapy (MET), and family therapy tended to be more effective than generic “practice as usual” interventions that often involved referrals to community services. However, they did not find strong relationships between the characterisitcs of adolescents in the studies and the effectiveness of interventions (net of intervention type).

A complicating factor in conducting these analyses was that some of the data were missing. Not every study reported the requisite information for extracting covariates for every effect size. For instance, [INSERT EXAMPLE]. As a result, not all effect estimate had information about the types of individuals in the study or the [INSERT OTHER COVARIATE]. Thus, when it came time to run meta-regressions, Tanner et al. were faced with a decision about how to address effects for which they had no covariates.

Tanner-Smith et al. ultimately opted for a sophisticated statistical procedure called the expectation-maximization (EM) algorithm to estimate their models. However, that was not their only option. They could have simply omitted effects with missing covariates from their analysis, imputed values that were missing, or augmented their meta-regression models so that they included a model for the missing covariates.

1. **Primer on missing data terms and why missingness can matter.**

As noted in the introduction, how to address missing data is not a one-size-fits-all problem. Instead, the methods one *should* use depend on a variety of factors. Thus, it is usually a good idea to conduct some exploratory analyses to help understand the nature and the scale of missingness in a dataset.

* Size of the problem is related to how much data is missing.
* Size of the problem is also related to why data is missing.
  + Define MAR/MCAR/MNAR and give plain language explanations
* Examples from data: Let’s see if we can get 2-3 variables with different amounts of missingness to walk through how missingness and mechanism interact.

1. **Issue 1: What data is missing and how much? Knowing which and how much data is missing is important for at least two reasons.** First, any potential biases are related to the amount of missing data. When a greater amount of data are missing and excluded from an analysis, then any potential biases can be larger. Conversely, if only a small amount of data are missing, then any potential biases will be small. Second, any corrections one might make will depend on and be limited by which variables are missing and how frequently. Strategies that impute missing values for variable X tend to perform better if imputations can make use of important related variables. If those related variables are also likely to be missing when X is missing, this can limit how “good” imputations are.
   * Plots
     1. Whole-data plots
     2. Bivariate (or sets of variables)
     3. Comparison with effect size and error variances
   * Numerical summaries
     1. Raw counts/percents
        + Raw counts are useful in telling us how many effects went into fitting a model.
     2. Precision averages
        + Precision averages are useful in telling us whether the big or small studies are used to fit a model.
        + Potential sticking points: dependent effect sizes and study-level predictors. If a study has several effect estimates (as in the data), and the predictor is at the study level (and not effect size level), then precision averages (and percents) may count this multiple times. But that doesn’t seem right.
   * Practical guidance for reporting
2. Issue 2: Why are data missing?
   * Assumptions of MAR, MCAR, MNAR, MAAR, ignorability
   * Illustration in data
   * Potential tests
3. Potential corrections, and what is known about them in meta-analysis
   * Likelihood-based approaches like EM, FIML
   * Imputation-based approaches like MI, SI, LVCF
   * Selection models
   * Pattern mixture models
   * Note that implementation in meta-anlytic software is not widespread.