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 is 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.

This section examines different visualizations using naniar (Tierney, 2018), visdat (Tierney, 2017) and ggplot2 (Wickham, 2009) R’s packages. We demonstrate how visualizations typically used with datasets outside of meta-analysis can be adapted to the realities of meta-analytic data and contribute to understanding a dataset structure. Three types of visualization of missing data are discussed: whole-data plot; bivariate plots; and comparison with effect size and error variance plots.

3.1 Whole-data plots

Aggregation plots are useful tools for identifying the number of missings in each variable and case. Overall missingness is visualized with a heatmap style using vis\_dat and vis\_miss functions from the visdat package (see Figure 1 below).

(Insert plot 1 - 2)

These simple visualizations demonstrate how severe the level of data loss is within our dataset. In this example, Plot A highlights variables with missing data. While none of the effect estimates are missing, it is clear that missing values appear in other 18 variables of interest, which, as shown in plot B, represents 6.2% of the total dataset.

Similarly, function gg\_miss\_var from naniar package provides an approach to visualizing overall missingness by identifying variables with a greater number or percentage of missing cases, ordering by missingness (see Figure 2 below).

(Insert plot 3)

Figure 2 presents variables ordered by missingness. For this particular dataset, there are 10 variables with at least 10% of missing cases. This visualization becomes relevant when deciding which variable to include in the analysis.

Different combinations of missingness across cases can be visualized using an “upset plot” (Conway et al. 2017) with the gg\_miss\_upset function in the naniar package; thus providing the number of times certain variables go missing together.  
We explore the combinations among variables with higher percentage of missing (see Figure 3 below).

(Insert plot 8)

Figure 3 details those variables that are missing together. For instance, there are a large number of cases where Group 2 Level of Care, Number of Sessions, Treatment Contact (hours per week) and Duration of Treatment (days) are missing together. This simple exploration provides valuable information for imputation.

3.2 Bivariate plots

Variable missingness in cases over some other factor variable is visualized with a heatmap style using the gg\_miss\_fct function from the naniar package (see Figure 4 below). This provides information regarding any relationship between observed values and missingness condition.

(Insert plot 7)

Figure 4 highlights the number of missings in each column, broken down by a factor variable, in this case the Level of Care for group 2. The inpatient category has 100% of missing values in at least 12 different variables, suggesting that this category could impose a problem when fitting a regression model.

3.3 Comparison with effect size and error variances plots

To explore the relationship of variables presenting a large percentage of missing data with effect size and error variances, the original data is transform in order to create a data structure that keeps track of missing values (Tierney, 2018). Using the as\_shadow and bind\_shadow functions from the naniar package a data frame with the same set of columns, but with the column names added a suffix\_NA, is bound to the original dataset.

Later, the distribution of effect size/error variance is visualized when some covariates are missing, and when they are not using the ggplot function. Figure 5 shows three scenarios with different relevant covariates.

1. Clear relationship

(Insert plot 19)

Plot A shows that the covariance Duration of Treatment (days) for Group 1 is mostly missing for larger effect size values. Further, the effect size has larger standard error, when this covariate is missing.

1. Some relationship

(Insert plot 17)

Plot B illustrates a case where the effect size tends to be closer to zero when a particular covariate is missing. Specifically, when Treatment Contact (hours per week) for group 1 is missing, both the effect size and its standard errors tend to be smaller than when the covariate is present.

1. No relationship

(Insert plot 18)

Plot C shows that both, the effect size and its standard errors, have a similar distribution either when the covariate Treatment Contact (hours per week) for group 2 is present or not.

- Numerical summaries  
  
 i. Raw counts/percents  
   
 - Raw counts are useful in telling us how many effects went into fitting a model.  
   
 ii. 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

1. Issue 2: Why are data missing?
   * Relate back to assumptions of MAR, MCAR, MNAR, MAAR, ignorability
   * Illustration in data
   * Potential tests and their limitations
2. 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.