

# On Missing Data

Drafted by Haoda Fu, Ph.D.

Statistics for Data Scientists Series

November 8, 2018

**Advanced Analytics and Data Sciences (AADS)**  
**Eli Lilly and Company**

Why this topic is important

Missing data are common problems for data scientists.

## Why this topic is important

Missing data are common problems for data scientists.

## Connections

- Missing data and causal inference are closely connected and both are important to our digital health.
- Different types of missing mechanism: missing completely at random, missing at random, and missing not at random.
- Different methods to handle missing data.

- $Y = (Y_{ij})$ : complete data matrix.
- $M = (M_{ij})$ : missing-data indicator matrix.
- $\phi$ : unknown parameter.
- $Y_{obs}$  and  $Y_{mis}$ : the observed and missing components of  $Y$ .
- $f(\cdot)$ : probability density function.

## Definition (MCAR)

MCAR is defined that the  $M$  does not depend on  $Y$ ,

$$f(M|Y, \phi) = f(M|\phi), \quad \forall Y, \phi.$$

Note: this assumption does not mean that the pattern itself is random, but rather that the missingness does not depend on the data values.

## Definition (MAR)

MAR is defined that the  $M$  only depends on  $Y_{obs}$ ,

$$f(M|Y, \phi) = f(M|Y_{obs}, \phi), \quad \forall Y_{obs}, \phi.$$

Note: This assumption is weaker than the MCAR and it is the most widely used assumption for clinical trials.

### Definition (MNAR)

The mechanism is called MNAR if the distribution of  $M$  depends on the  $Y_{miss}$ .

Note: Assumptions are not verifiable. It is often used for robustness evaluation. Methods handle MNAR include pattern mixture models and selection models.

Let  $\theta$  be the interest parameter, i.e.  $f(Y|\theta) = f(Y_{obs}, Y_{mis}|\theta)$ , and  $\psi$  be the parameter for missing mechanisms.

$$\begin{aligned}f(Y, M|\theta, \psi) &= f(Y|\theta)f(M|Y, \psi) \\f(y_{obs}, m|\theta, \psi) &= \int f(y_{obs}, Y_{mis}|\theta)f(m|y_{obs}, Y_{mis}, \psi)dY_{mis} \\&= f(m|y_{obs}, \psi) \int f(y_{obs}, Y_{mis}|\theta)dY_{mis} \\&\triangleq f(m|y_{obs}, \psi)f(y_{obs}|\theta)\end{aligned}$$

Remarks: the lower case means the observed value. The  $dY_{mis}$  can only be calculated when  $M$  is observed. Also,  $f(y_{obs}|\theta)$  is not the same pdf as if we only observed iid data, so that we use  $\triangleq$  notation instead of regular equal to, because we define  $f(y_{obs}|\theta) = \int f(y_{obs}, Y_{mis}|\theta)dY_{mis}$ .