

Missing Data Descriptives

Utrecht University Winter School: Missing Data in R



**Utrecht
University**

Kyle M. Lang

Department of Methodology & Statistics
Utrecht University

Outline

Missing Data Patterns

Nonresponse Rates

Coverage Measures



What are Missing Data?

Missing data are empty cells in a dataset where there should be observed values.

- The missing cells correspond to true population values, but we haven't observed those values.



What are Missing Data?

Missing data are empty cells in a dataset where there should be observed values.

- The missing cells correspond to true population values, but we haven't observed those values.

Not every empty cell is a missing datum.

- Quality-of-life ratings for dead patients in a mortality study
- Firm profitability after the company goes out of business
- Self-reported severity of menstrual cramping for men
- Empty blocks of data following “gateway” items



A Little Notation

Y := An $N \times P$ Matrix of Arbitrary Data

Y_{mis} := The *missing* part of Y

Y_{obs} := The *observed* part of Y

R := An $N \times P$ response matrix

M := An $N \times P$ missingness matrix

The R and M matrices are complementary.

- $r_{np} = 1$ means y_{np} is observed; $m_{np} = 1$ means y_{np} is missing.
- $r_{np} = 0$ means y_{np} is missing; $m_{np} = 0$ means y_{np} is observed.
- M_p is the *missingness* of Y_p .

Missing Data Pattern

Missing data (or response) patterns represent unique combinations of observed and missing items.

- P items $\Rightarrow 2^P$ possible patterns.

	X	Y
1	x	y
2	x	.
3	.	y
4	.	.

Patterns for $P = 2$

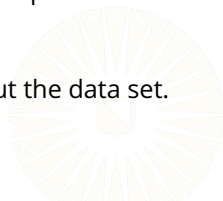
	X	Y	Z
1	x	y	z
2	x	y	.
3	x	.	z
4	.	y	z
5	x	.	.
6	.	.	z
7	.	y	.
8	.	.	.

Patterns for $P = 3$

Missing Data Pattern

The concept of a “missing data pattern” can also be used to classify the spatial arrangement of missing cells on a data set.

- Univariate
 - Missing data occur on only one variable
- Monotone
 - The proportion of complete elements, in both rows and columns, decreases when traversing the data set.
 - The observed cells can be arranged into a “staircase” pattern.
- Arbitrary
 - Missing values are “randomly” scattered throughout the data set.



Example Missing Data Patterns

	X	Y	Z
1	x	y	z
2	x	y	z
3	x	y	z
4	x	y	z
5	x	y	z
6	x	.	z
7	x	.	z
8	x	.	z
9	x	.	z
10	x	.	z

Univariate Pattern

	X	Y	Z
1	x	y	z
2	x	y	z
3	x	y	z
4	x	y	.
5	x	y	.
6	x	y	.
7	x	.	.
8	x	.	.
9	x	.	.
10	.	.	.

Monotone Pattern

	X	Y	Z
1	x	.	z
2	x	y	z
3	x	y	z
4	x	.	z
5	x	y	z
6	x	.	z
7	.	y	z
8	x	y	z
9	x	.	.
10	x	y	.

Arbitrary Pattern

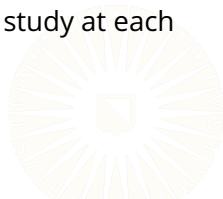
Nonresponse Rates

PROPORTION MISSING

- The proportion of cells containing missing data
- Good early screening measure
- Should be computed for each variable, not for the entire dataset

ATTRITION RATE

- The proportion of participants that drop-out of a study at each measurement occasion



Nonresponse Rates

PROPORTION OF COMPLETE CASES

- The proportion of observations with no missing data
- Often reported but nearly useless quantity

FRACTION OF MISSING INFORMATION

- Associated with an estimated parameter, not with an incomplete variable
- Like an R^2 for the missing data
- Most important diagnostic value for missing data problems
- Can only be computed after treating the missing data



Coverage Measures

COVARIANCE COVERAGE

$$CC_{jk} = N^{-1} \sum_{n=1}^N r_{nj} r_{nk}$$

- The proportion of cases available to estimate a given pairwise relationship (e.g., a covariance between two variables)
- Very important to have adequate coverage of the parameters you want to estimate

Coverage Measures

INBOUND STATISTIC

$$I_{jk} = \frac{\sum_{n=1}^N (1 - r_{nj}) r_{nk}}{\sum_{n=1}^N (1 - r_{nj})}$$

- The proportion of missing cases in Y_j for which Y_k is observed

OUTBOUND STATISTIC

$$O_{jk} = \frac{\sum_{n=1}^N r_{nj} (1 - r_{nk})}{\sum_{n=1}^N r_{nj}}$$

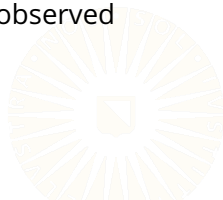
- The proportion of observed cases in Y_j for which Y_k is missing

Coverage Measures

INFLUX COEFFICIENT

$$I_j = \frac{\sum_{k=1}^P \sum_{n=1}^N (1 - r_{nj}) r_{nk}}{\sum_{k=1}^P \sum_{n=1}^N r_{nk}}$$

- The proportion of observed cells in Y that exists in cases for which Y_j is missing
- How well the missing values in Y_j connect to the observed values in Y_{-j}



Coverage Measures

OUTFLUX COEFFICIENT

$$O_j = \frac{\sum_{k=1}^P \sum_{n=1}^N r_{nj}(1 - r_{nk})}{\sum_{k=1}^P \sum_{n=1}^N (1 - r_{nk})}$$

- The proportion of missing cells in Y that exists in cases for which Y_j is observed
- How well the observed values in Y_j connect to the missing values in Y_{-j}

