Missing Data Mechanisms

1

Patterns Versus Mechanisms

The missing data pattern describes the configuration of observed and the missing values in a data set

The pattern describes the location of the holes in the data but says nothing about why the data are missing

The missing data mechanism describes how the probability of missing data on a particular variable is related to other variables, if at all

2

General Missing Data Pattern

A general pattern has missing values dispersed throughout the data matrix

Missingness may or may not be systematic

Modern missing data procedures handle general patterns

X	Y	Z
NA	Y_1	NA
NA	NA	Z_2
NA	Y_3	Z_3
X_4	Y_4	NA
X_5	Y_5	NA
X_6	Y_6	Z_6
NA	Y_7	Z_7
NA	Y_8	Z_8
X_9	NA	Z_9
X_{10}	Y_{10}	NA

3

Motivating Example

20 participants from a smoking cessation study

Participants report the number of years smoking and number of cigarettes smoked

20% of respondents do not report the number of cigarettes smoked

Years	Cigarettes
7	9
8	12
1	11
4	3
6	10
8	5
8	7
10	11
15	12
5	NA
9	NA
11	11
14	10
13	NA
12	15
11	8
10	NA
10	NA
7	10
11	10

4

Notation and Terminology

Y_{com} is the hypothetically complete data set

Y_{obs} denotes the observed portions of Y_{com}

Y_{mis} denotes the unseen (latent) scores in Y_{com} that are missing

R is a missing data indicator (or matrix of indicators) where $R = 0$ if Y is observed and $R = 1$ if Y is missing

Y_{com}	
Years	Cigs
7	9
8	12
1	11
4	3
6	10
8	5
8	7
10	11
15	12
5	11
9	12
11	11
14	10
13	19
12	15
11	8
10	13
10	8
7	10
11	10

Y	
Years	Cigs
7	9
8	12
1	11
4	3
6	10
8	5
8	7
10	11
15	12
5	NA
9	NA
11	11
14	10
13	NA
12	15
11	8
10	NA
10	NA
7	10
11	10

$Y_{com} = (Y_{obs}, Y_{mis})$	
Years	Cigs
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{mis}
Y_{obs}	Y_{mis}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{mis}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{mis}
Y_{obs}	Y_{mis}
Y_{obs}	Y_{obs}
Y_{obs}	Y_{obs}

R	
Years	Cigs
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	0
0	1
0	1
0	0
0	0
0	1
0	0
0	0
0	1
0	1
0	0
0	0

Missing Data Mechanisms

Rubin's (1976) missing data mechanisms describe relations between the probability of missing data and Y_{obs} and Y_{mis}

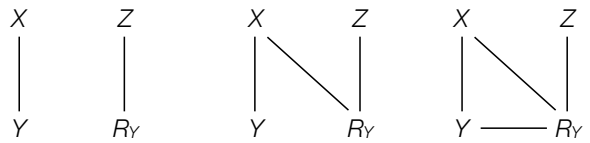
Missing completely at random (MCAR)

Missing at random (MAR)

Not missing at random (NMAR)

Diagram of Mechanisms

X represents a set of observed variables correlated with Y, Z represents a set of observed variables uncorrelated with X and Y, and R_Y is the missing data indicator for Y



Motivating Example

20 participants enroll in a smoking cessation study

Participants report age, number of years smoking, number of cigarettes smoked, and self-efficacy to quit

Number of cigarettes and self-efficacy have missing values

Age	Years	Cigs	Efficacy
29	7	9	NA
39	8	NA	NA
25	1	11	16
41	4	NA	21
39	6	10	17
41	8	5	10
46	8	7	13
40	10	NA	10
51	15	NA	11
43	5	11	13
26	9	12	11
51	11	11	16
36	14	NA	10
51	13	19	9
41	12	15	5
28	11	8	7
30	10	13	10
41	10	8	NA
23	7	10	7
33	11	10	6

Missing Completely At Random (MCAR)

The probability of missing data on a variable *Y* is unrelated to observed responses of other variables and is unrelated to the would-be values of *Y* itself

$$P\left(R|Y_{obs}, Y_{mis}\right) = P\left(R\right)$$

All participants have the same probability of missing data

MCAR = observed and missing scores are the same, on average

Testing MCAR with Missing Data Indicators

MCAR is the only mechanism with testable propositions

Create a missing data indicator *R* for each incomplete variable (e.g., 0 = complete, 1 = missing) and examine mean differences across missing data patterns

This strategy can rule out MCAR but says nothing about the plausibility of MAR and NMAR mechanisms

MCAR Example

The absence of large mean differences supports MCAR

Pattern	Mean	SD	n
Age			
Complete	37.5	8.6	15
Missing	38.2	10.1	5
Years			
Complete	8.9	3.7	15
Missing	9.4	2.9	5
Self-Efficacy			
Complete	11.5	4.9	13
Missing	10.8	1.7	4

Age	Years	Cigs	Efficacy	R_{Cigs}
29	7	9	NA	0
39	8	12	NA	0
25	1	11	16	0
41	4	3	21	0
39	6	10	17	0
41	8	5	10	0
46	8	7	13	0
40	10	11	10	0
51	15	12	11	0
43	5	NA	13	1
26	9	NA	11	1
51	11	11	16	0
36	14	10	10	0
51	13	NA	9	1
41	12	15	5	0
28	11	8	7	0
30	10	NA	10	1
41	10	NA	NA	1
23	7	10	7	0
33	11	10	6	0

13

Missing At Random (MAR)

The probability of missing data on a variable *Y* is related to observed responses of other variables but is unrelated to the would-be values of *Y* itself

$$P\left(R|Y_{obs}, Y_{mis}\right) = P\left(R|Y_{obs}\right)$$

The probability of nonresponse varies across different observed score profiles

14

MAR = observed and missing scores are the same, on average, after conditioning on (controlling for) other variables

15

MAR Example

The presence of large mean differences refutes MCAR

Pattern	Mean	SD	n
Age			
Complete	36.5	9.4	15
Missing	41.4	5.7	5
Years			
Complete	8.1	3.2	15
Missing	11.8	2.9	5
Self-Efficacy			
Complete	12.0	4.5	13
Missing	9.0	2.7	4

Age	Years	Cigs	Efficacy	R_{Cigs}
29	7	9	NA	0
39	8	NA	NA	1
25	1	11	16	0
41	4	3	21	0
39	6	10	17	0
41	8	5	10	0
46	8	7	13	0
40	10	NA	10	1
51	15	NA	11	1
43	5	11	13	0
26	9	12	11	0
51	11	11	16	0
36	14	NA	10	1
51	13	19	9	0
41	12	NA	5	1
28	11	8	7	0
30	10	13	10	0
41	10	8	NA	0
23	7	10	7	0
33	11	10	6	0

16

MAR is Untestable

MAR implies that the missing data indicator R_Y is unrelated to the missing values of Y

After controlling for observed variables, the distribution of Y is the same for complete and incomplete cases

MAR is inherently untestable because it involves propositions about the missing Y values

Not Missing At Random (NMAR)

The probability of missing data on a variable Y is related to observed responses of other variables and is related to the would-be values of Y itself

$$P\left(R|Y_{obs}, Y_{mis}\right)$$

The unseen values in Y_{mis} predict nonresponse above and beyond the observed scores in Y_{obs}

NMAR = observed and missing scores are different, on average, after conditioning on (controlling for) other variables

NMAR Example

The presence of large mean differences refutes MCAR

Pattern	Mean	SD	n
Age			
Complete	35.5	8.5	15
Missing	44.4	6.1	5
Years			
Complete	8.1	3.3	15
Missing	11.6	2.7	5
Self-Efficacy			
Complete	12.1	4.5	13
Missing	8.8	2.6	4

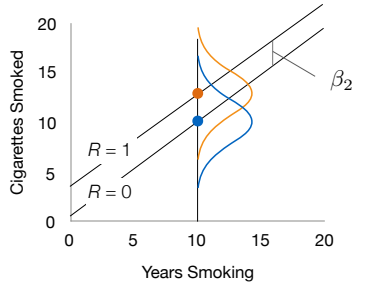
Age	Years	Cigs	Efficacy	R_{Cigs}
29	7	9	NA	0
39	8	NA	NA	1
25	1	11	16	0
41	4	3	21	0
39	6	10	17	0
41	8	5	10	0
46	8	7	13	0
40	10	NA	10	1
51	15	NA	11	1
43	5	11	13	0
26	9	12	11	0
51	11	11	16	0
36	14	10	10	0
51	13	NA	9	1
41	12	NA	5	1
28	11	8	7	0
30	10	13	10	0
41	10	8	NA	0
23	7	10	7	0
33	11	10	6	0

NMAR-Based Regression

The missing data indicator carries information about the missing values above and beyond years smoking

Valid inference requires a model that includes the indicator variable

$$Cigs = \beta_0 + \beta_1(Years) + \beta_2\,(R) + e$$



Difficulty with NMAR Modeling

The NMAR model is inestimable because it includes a parameter for which there is no data

$$\begin{aligned} Cigs &= \beta_0 + \beta_1(Years) + \beta_2\,(R) + e \\ &= \beta_0 + \beta_1(Years) + (\mu_{mis} - \mu_{obs})\,(R) + e \end{aligned}$$

Estimating the model requires the user to specify a value for the inestimable mean of the incomplete cases

Why Mechanisms Matter

Mechanisms function as analysis assumptions

Some older approaches require MCAR (others make no attempt to satisfy any mechanism)

Modern approaches like multiple imputation, maximum likelihood, and Bayes assume MAR (or MCAR)

Estimates are biased when assumptions are violated

Computer Simulation Procedure

Generate 1000 samples of bivariate data with *N* = 250

Delete 50% of *Y* scores according to an MCAR, MAR, or NMAR mechanism

Apply complete-case analysis (exclude all incomplete cases) or maximum likelihood (an MAR-based procedure that uses all data) to each of the 1000 data sets

Compute the average estimates for each method and compare these to the true population values

MCAR Simulation Results

Complete-case analysis and maximum likelihood are unbiased because mechanism assumptions are satisfied

Parameter	True Value	ML	CCA
M_X	100.00	100.02	99.98
SD_X	13.00	12.97	13.38
M_Y	12.00	11.99	12.00
SD_Y	3.00	2.99	3.00
R_{XY}	0.50	0.50	0.50

MAR Simulation Results

Complete-case analysis is biased because deleting cases requires MCAR, but maximum likelihood is unbiased because it requires an MAR mechanism

Parameter	True Value	ML	CCA
M_X	100.00	100.01	110.35
SD_X	13.00	12.98	7.86
M_Y	12.00	12.01	13.21
SD_Y	3.00	2.99	2.76
R_{XY}	0.50	0.49	0.14

NMAR Simulation Results

Both complete-case analysis and maximum likelihood are biased because the necessary mechanisms do not hold, but complete-case estimates are generally worse

Parameter	True Value	ML	CCA
M_X	100.00	100.00	105.51
SD_X	13.00	13.00	11.90
M_Y	12.00	14.12	14.40
SD_Y	3.00	1.82	1.81
R_{XY}	0.50	0.36	0.32

Practical Recommendations

MAR-based methods are usually a good starting point but are not necessarily perfect solutions

We cannot test for an MAR mechanism and thus we must rely on logical arguments and knowledge about our data collection and participants to justify these methods

NMAR-based procedures are available but are difficult to implement and require other tenuous assumptions