

# Missing Data Mechanisms

Utrecht University Winter School: Missing Data in R



**Utrecht  
University**

Kyle M. Lang

Department of Methodology & Statistics  
Utrecht University

# Outline

---

Definitions

Consequences

Testing



# Missing Data Mechanisms

---

## Missing Completely at Random (MCAR)

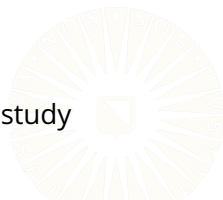
- $P(R|Y_{mis}, Y_{obs}) = P(R)$
- Missingness is unrelated to any study variables.

## Missing at Random (MAR)

- $P(R|Y_{mis}, Y_{obs}) = P(R|Y_{obs})$
- Missingness is related to only the *observed* parts of study variables.

## Missing not at Random (MNAR)

- $P(R|Y_{mis}, Y_{obs}) \neq P(R|Y_{obs})$
- Missingness is related to the *unobserved* parts of study variables.



# Simulate Some Toy Data

---

```
library(mvtnorm); library(dplyr); library(magrittr)

set.seed(235711)

nObs <- 5000 # Sample Size
pm   <- 0.3  # Proportion Missing

sigma <- matrix(c(1.0, 0.5, 0.3,
                  0.5, 1.0, 0.0,
                  0.3, 0.0, 1.0),
               ncol = 3)
dat0 <- rmvnorm(nObs, c(0, 0, 0), sigma) %>% data.frame()
colnames(dat0) <- c("x", "y", "z")

dat0 %$% cor(y, x)

[1] 0.4997145
```

# MCAR Example

---

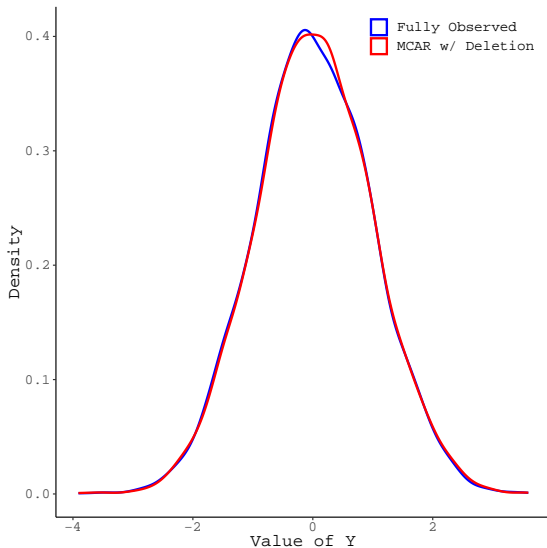
```
## Simulate MCAR Missingness:
m <- sample(1:nObs, size = pm * nObs)

## Impose MCAR missing on Y:
mcarData <- dat0
mcarData[m, "y"] <- NA

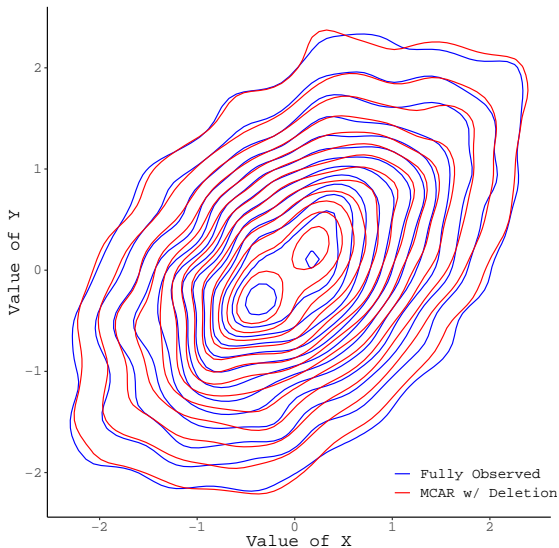
## Check the correlation between X & Y:
mcarData %$% cor(y, x, use = "pairwise")

[1] 0.5195767
```

# MCAR Example



# MCAR Example



# MAR Example

---

```
## Simulate MAR Missingness:
m <- with(dat0, x < quantile(x, probs = pm))

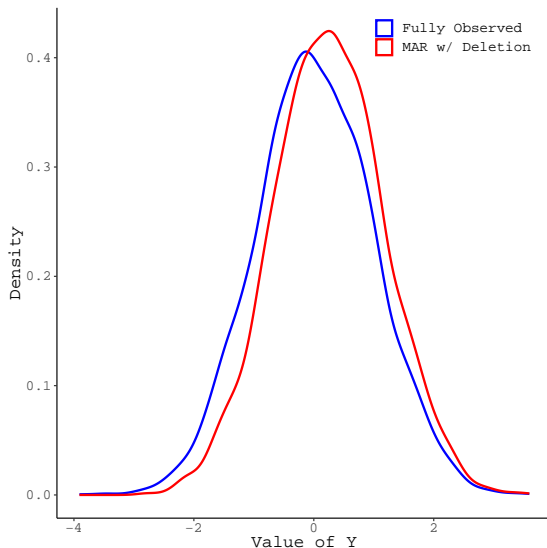
## Impose MAR missing on Y:
marData      <- dat0
marData[m, "y"] <- NA

## Check the correlation between X & Y:
marData %$% cor(y, x, use = "pairwise")

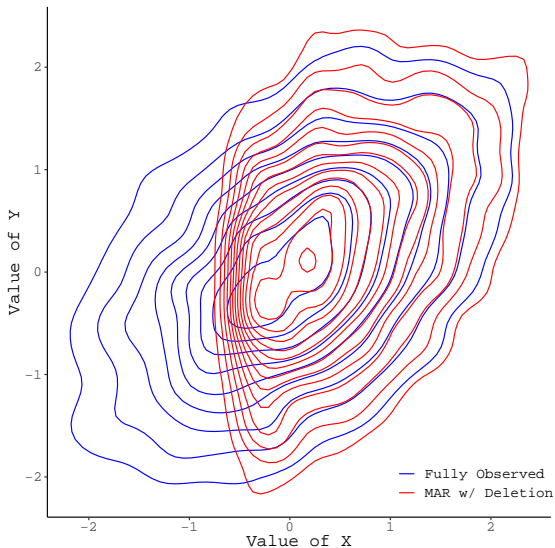
[1] 0.3822143
```



# MAR Example



# MAR Example



# MNAR Example

---

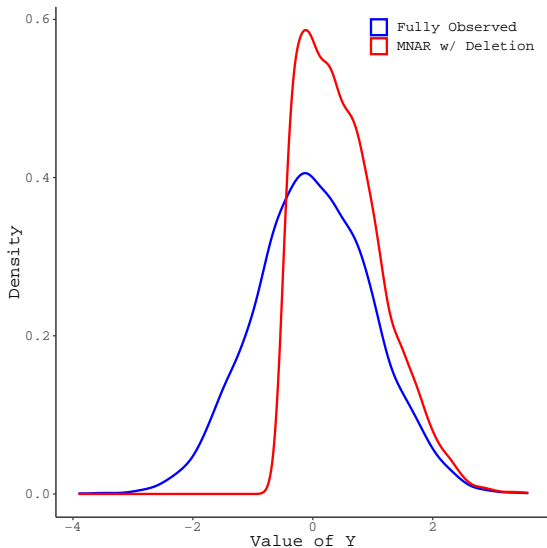
```
## Simulate MNAR Missingness:
m <- with(dat0, y < quantile(y, probs = pm))

## Impose MNAR missing on Y:
mnarData <- dat0
mnarData[m, "y"] <- NA

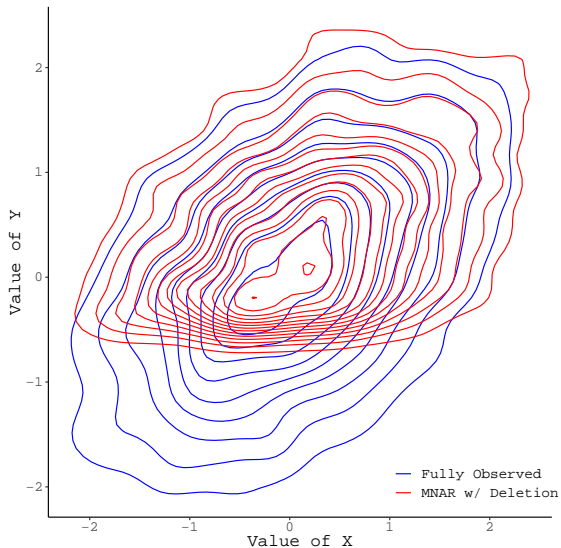
## Check the correlation between X & Y:
mnarData %$% cor(y, x, use = "pairwise")

[1] 0.3902962
```

# MNAR Example



# MNAR Example



# Crucial Nuance

---

In our previous MAR example, ignoring the predictor of missingness actually produces *Indirect MNAR*.

# Crucial Nuance

---

In our previous MAR example, ignoring the predictor of missingness actually produces *Indirect MNAR*.

**QUESTION:** What happens if we ignore the predictor of missingness, but that predictor is independent of our study variables?

# Crucial Nuance

---

In our previous MAR example, ignoring the predictor of missingness actually produces *Indirect MNAR*.

**QUESTION:** What happens if we ignore the predictor of missingness, but that predictor is independent of our study variables?

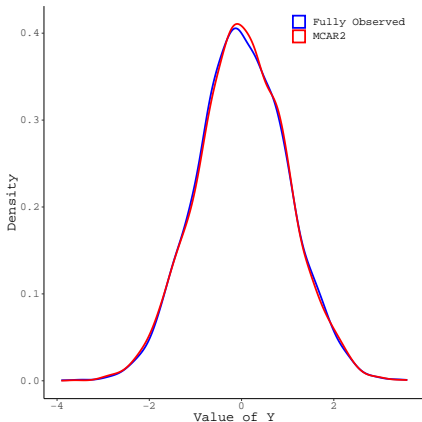
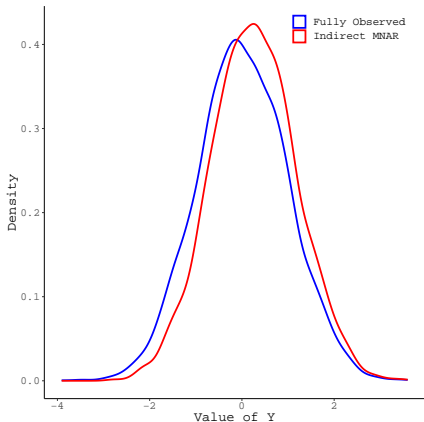
```
m <- with(dat0, z < quantile(z, probs = pm))  
  
mcarData2          <- dat0  
mcarData2[m, "y"] <- NA  
  
mcarData2 %$% cor(y, x, use = "pairwise")  
  
[1] 0.5118075
```

**ANSWER:** We get back to MCAR :)



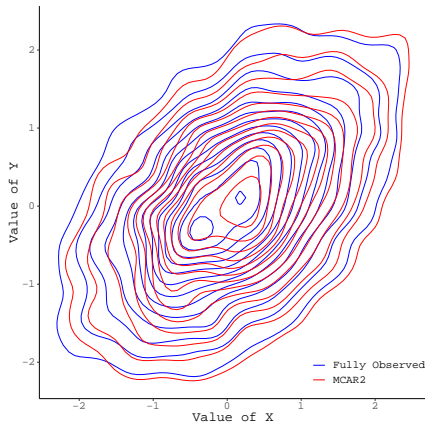
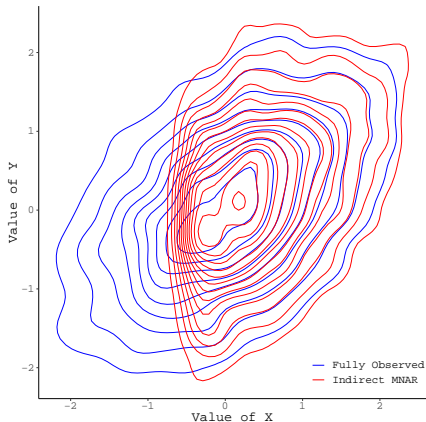
# Crucial Nuance

The missing data mechanisms are not simply characteristics of an incomplete dataset; we also need to account for the analysis.



# Crucial Nuance

The missing data mechanisms are not simply characteristics of an incomplete dataset; we also need to account for the analysis.

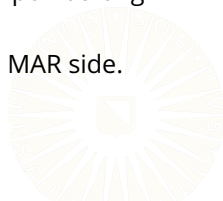


# Testing the Missing Data Mechanism

---

We cannot fully test the MAR or MNAR assumptions.

- To do so would require knowing the values of the missing data.
- We can find observed predictors of missingness.
  - Use classification algorithms to predict missingness from  $Y_{obs}$ .
  - We can never know that we have discovered all MAR predictors.
- In practice, MAR and MNAR live on the ends of a continuum.
  - Our missing data problem exists at some unknown point along this continuum.
  - We can do a lot to nudge our problem towards the MAR side.



# Testing the Missing Data Mechanism

---

We can (partially) test the MCAR assumption.

- With MCAR, the missing data and the observed data should have the same distribution.
- We can test for MCAR by testing the distributions of *auxiliary variables*,  $\mathbf{Z}$ .
  - Use a t-test to compare the subset of  $\mathbf{Z}_p$  that corresponds to  $\mathbf{Y}_{mis}$  to the subset corresponding to  $\mathbf{Y}_{obs}$ .
  - The Little (1988) MCAR test is a multivariate version of this.

These procedures actually test if the data are *observed* completely at random.

# References

---

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198–1202. doi: 10.1080/01621459.1988.10478722

