Missing Data: Mechanisms & Approaches

Ariadne Letrou

Missing data is everywhere!

Missing data is everywhere!

Time

Missing data is everywhere!

Missing data is everywhere!

Time

“One of the most important **statistical** and **design** problems in research”
— William Shadish

Missing data is everywhere! How do we deal with it?

````
```{r}
# Select variables
data <- data %>%
  select(score, cls_perc_eval, cls_students, bty_avg)

# Confirm that there is no missing data
nomiss <- data %>%
  drop_na() # drop missing data

print(c(nrow(data), nrow(nomiss)))
```
````

— Ariadne Letrou, Lab 7, PSY 503

How do we deal with missing data?

```
```{r}
# Select variables
data <- data %>%
  select(score, cls_perc_eval, cls_students, bty_avg)

# Confirm that there is no missing data
nomiss <- data %>%
  drop_na() # drop missing data

print(c(nrow(data), nrow(nomiss)))
```
```

— Ariadne Letrou, Lab 7, PSY 503

"**Deletion** methods are among the **worst** methods available for practical applications."

— American Psychological Association task Force on Statistical Inference

How do we deal with missing data?

```{r}
# Select variables
data <- data %>%
  select(score, cls_perc_eval, cls_students, bty_avg)

# Confirm that there is no missing data
nomiss <- data %>%
  drop_na() # drop missing data

print(c(nrow(data), nrow(nomiss)))
```

— Ariadne Letrou, Lab 7, PSY 503

"**Deletion** methods are among the **worst** methods available for practical applications."

— American Psychological Association task Force on Statistical Inference

na.rm

np.isnan()

Traditional approaches to missing data

- **Deletion**

Traditional approaches to missing data

- **Deletion**
    - **Listwise** - Discard all missing data from all of our analyses
        - Reduce sample size → reduce power of significant tests

Traditional approaches to missing data

- **Deletion**
  - **Listwise** - Discard all missing data from all of our analyses
    - Reduce sample size → reduce power of significant tests
  - **Pairwise** - Incomplete cases discarded on an analysis-by-analysis basis
    - Helps preserve some power

Traditional approaches to missing data

- **Deletion**
  - **Listwise** - Discard all missing data from all of our analyses
    - Reduce sample size → reduce power of significant tests
  - **Pairwise** - Incomplete cases discarded on an analysis-by-analysis basis
    - Helps preserve some power

Biases estimates

Traditional approaches to missing data

- **Deletion**
  - **Listwise** - Discard all missing data from all of our analyses
    - Reduce sample size → reduce power of significant tests
  - **Pairwise** - Incomplete cases discarded on an analysis-by-analysis basis
    - Helps preserve some power
- **Single imputation** - "Fill in" missing data with replacement values

Biases estimates

Traditional approaches to missing data

- **Deletion**
  - **Listwise** - Discard all missing data from all of our analyses
    - Reduce sample size → reduce power of significant tests
  - **Pairwise** - Incomplete cases discarded on an analysis-by-analysis basis
    - Helps preserve some power

Biases estimates

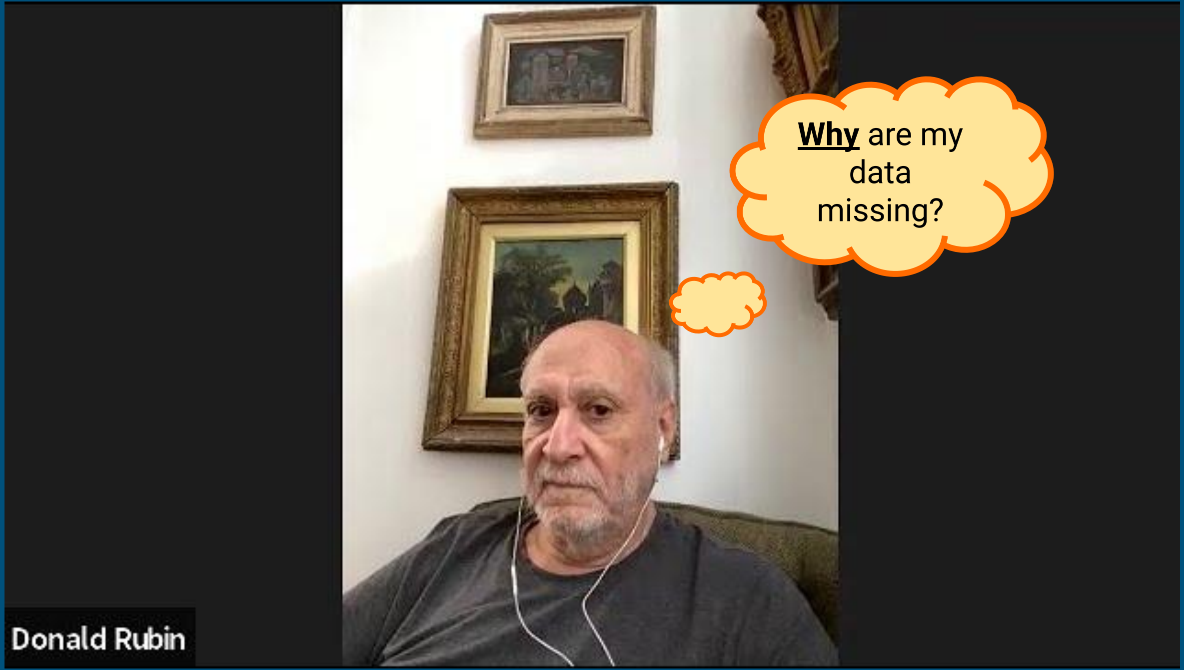
- **Single imputation** - "Fill in" missing data with replacement values
  - **Mean imputation**
  - **Regression imputation**
  - **Stochastic regression imputation** - "best" method
    - Adds random error to predicted values

Traditional approaches to missing data

- **Deletion**
  - **Listwise** - Discard all missing data from all of our analyses
    - Reduce sample size → reduce power of significant tests
  - **Pairwise** - Incomplete cases discarded on an analysis-by-analysis basis
    - Helps preserve some power
- **Single imputation** - "Fill in" missing data with replacement values
  - **Mean imputation**
  - **Regression imputation**
  - **Stochastic regression imputation** - "best" method
    - Adds random error to predicted values

Biases estimates

Attenuates estimates of correlation & variability


Why are my data missing?
Donald Rubin


Why are my data missing?
Not all missing data is created equal!
Donald Rubin

Theoretical background: Rubin’s missing data mechanisms

- Missing completely at random (**MCAR**)
- Missing at random (**MAR**)
- Missing not at random (**MNR**)

Theoretical background: Rubin's missing data mechanisms

- Missing completely at random (**MCAR**)
- Missing at random (**MAR**)
- Missing not at random (**MNR**)

How do you know which **one** of these mechanisms applies to your data?

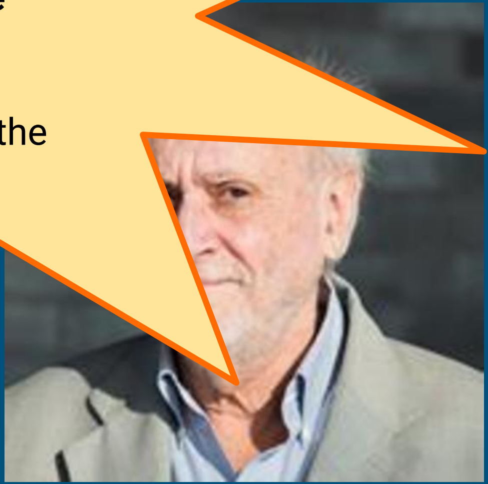
Theo tical backgroun Rubin's ng data mechanisms

- Missing comp
- Missi
- Mis

TRICK QUESTION!!!

The same dataset can have all **three** mechanisms present.

The mechanisms present depends on the variables in your **analysis**.



Missing Completely At Random (MCAR)

- Missingness is **not systematic** in any way.
  - More intuitively, “the observed data can be thought of a random subsample of the hypothetically complete data.”

Missing Completely At Random (MCAR)

- Missingness is **not systematic** in any way.
    - More intuitively, “the observed data can be thought of a random subsample of the hypothetically complete data.”

Missing Completely At Random (MCAR)

- Missingness is **not systematic** in any way.
  - More intuitively, “the observed data can be thought of a random subsample of the hypothetically complete data.”

Missing Completely At Random (MCAR)

- Missingness is **not systematic** in any way.
    - More intuitively, “the observed data can be thought of a random subsample of the hypothetically complete data.”

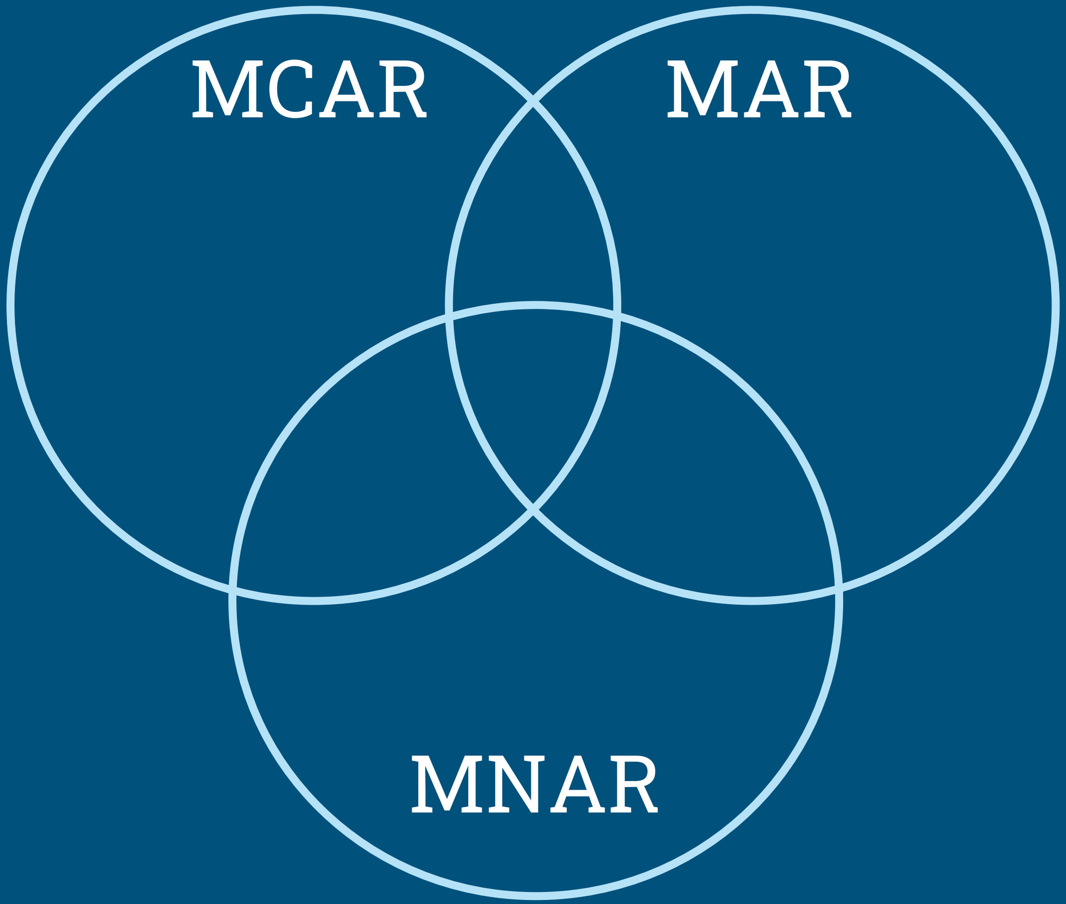
- Examples: Scheduling difficulties, administrative blunders, research design (planned missing data design)
- The ***strictest*** assumption!

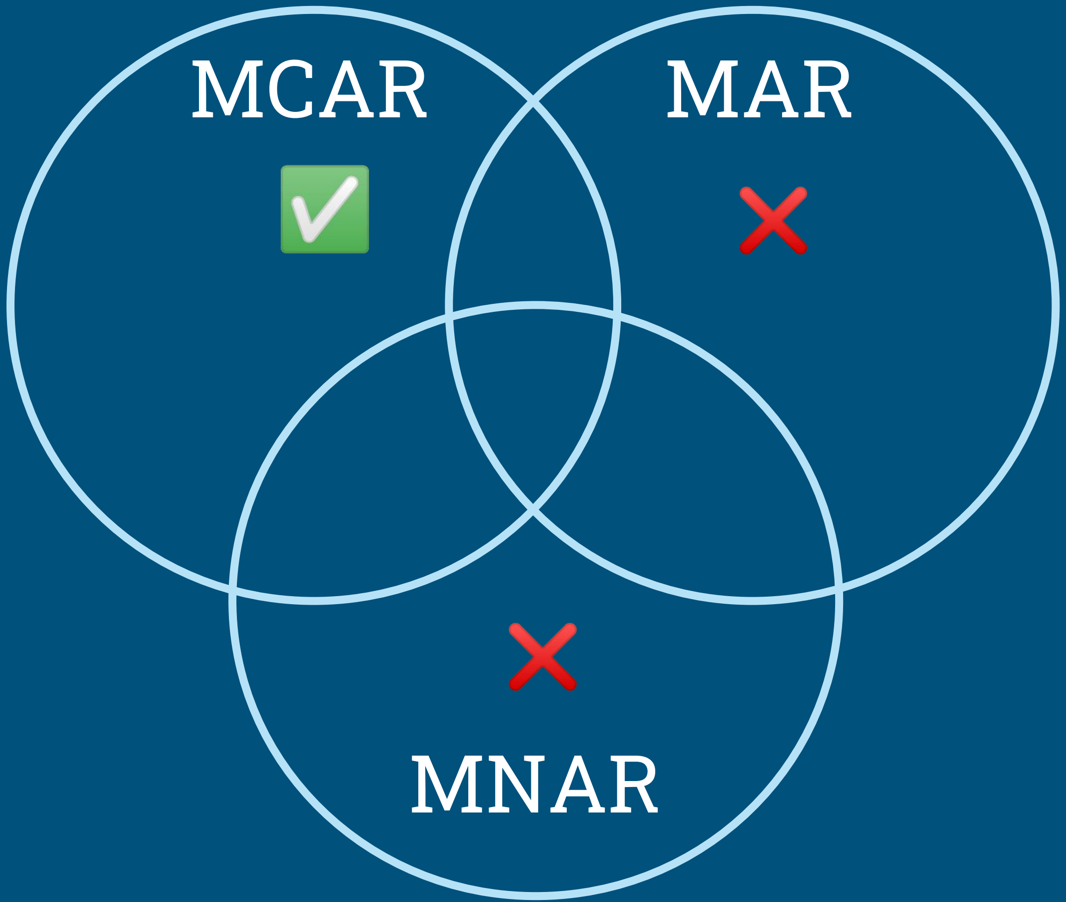
Missing At Random (MAR)

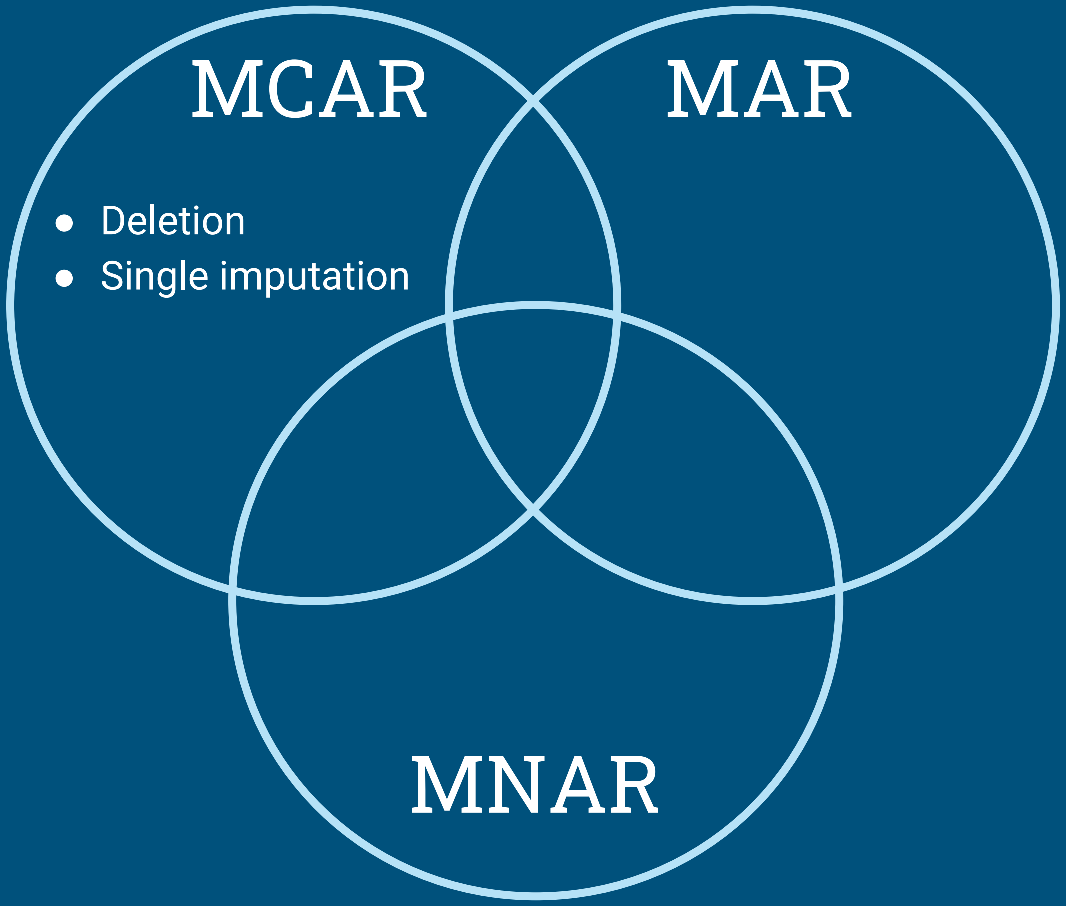
- Missingness is related to other measured variables in the analysis, but not to the underlying values of the incomplete variable itself.
  - MAR is not actually random at all, despite the name…
  - In other words, the missingness is **systematic**. The propensity of missing data is correlated with other variables in the analysis.
- Example: Substance abuse & self-esteem scores
- ***Less strict*** assumption than MCAR.

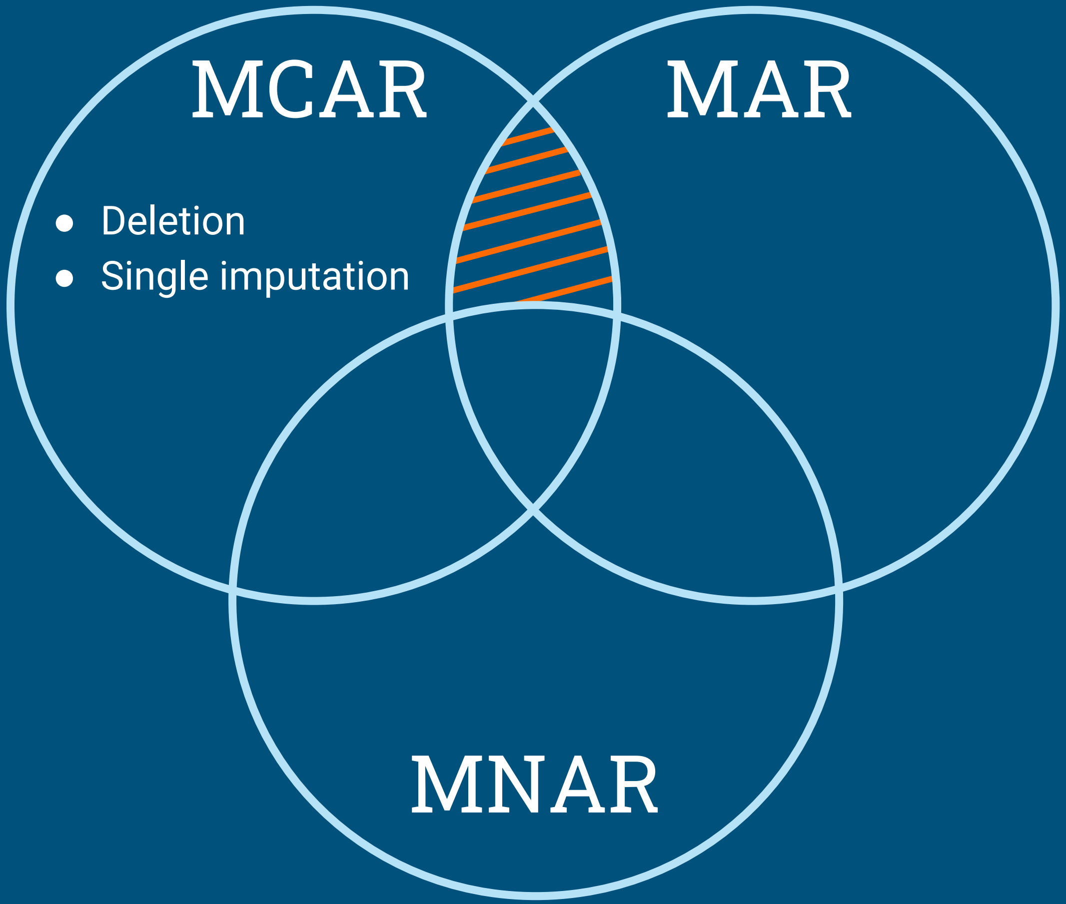
Missing Not At Random (MNAR)

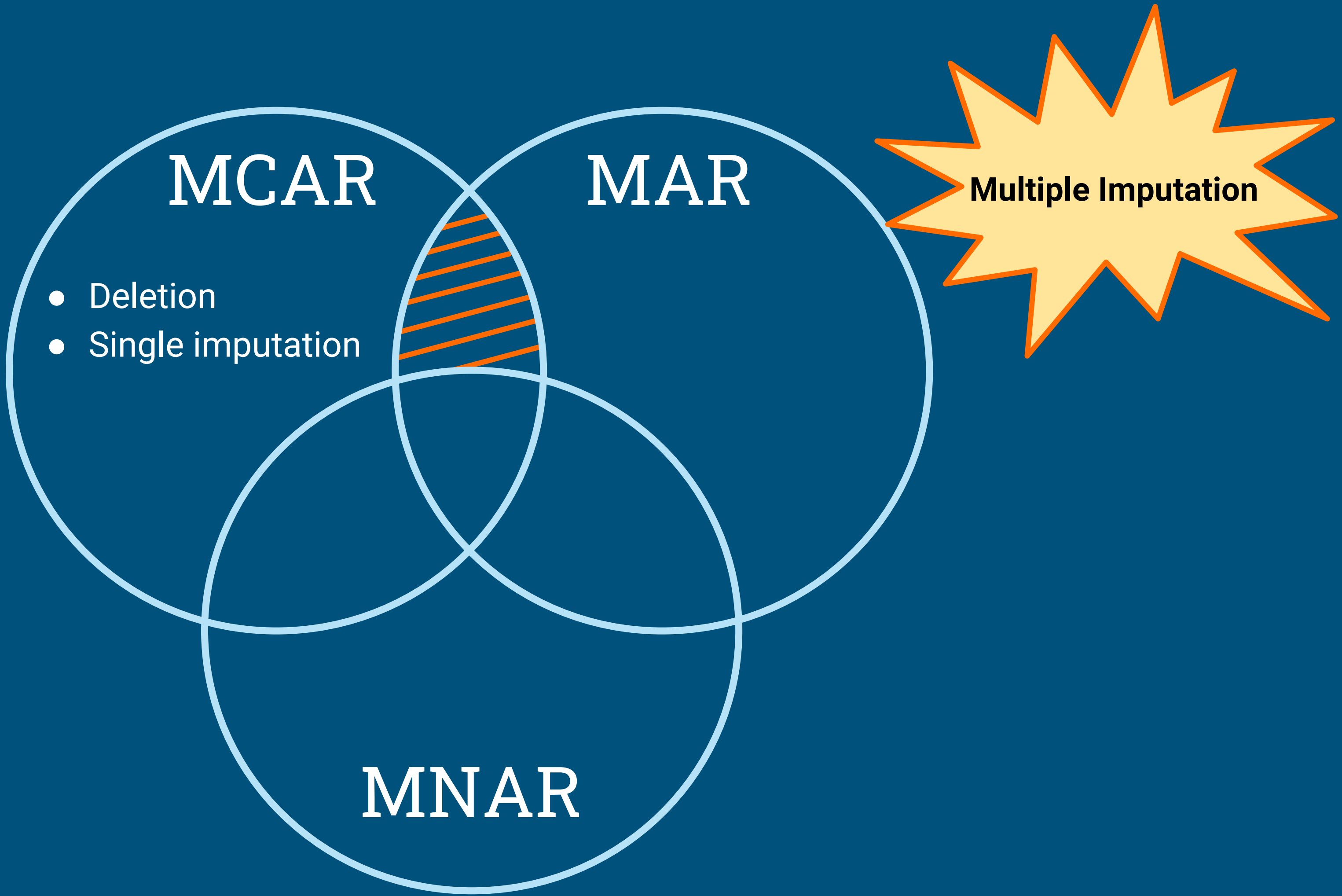
- Missingness is **systematic** and related to the hypothetical values of the incomplete variable.
- Example: Missing questions on a reading test because you fail to understand the accompanying text excerpt

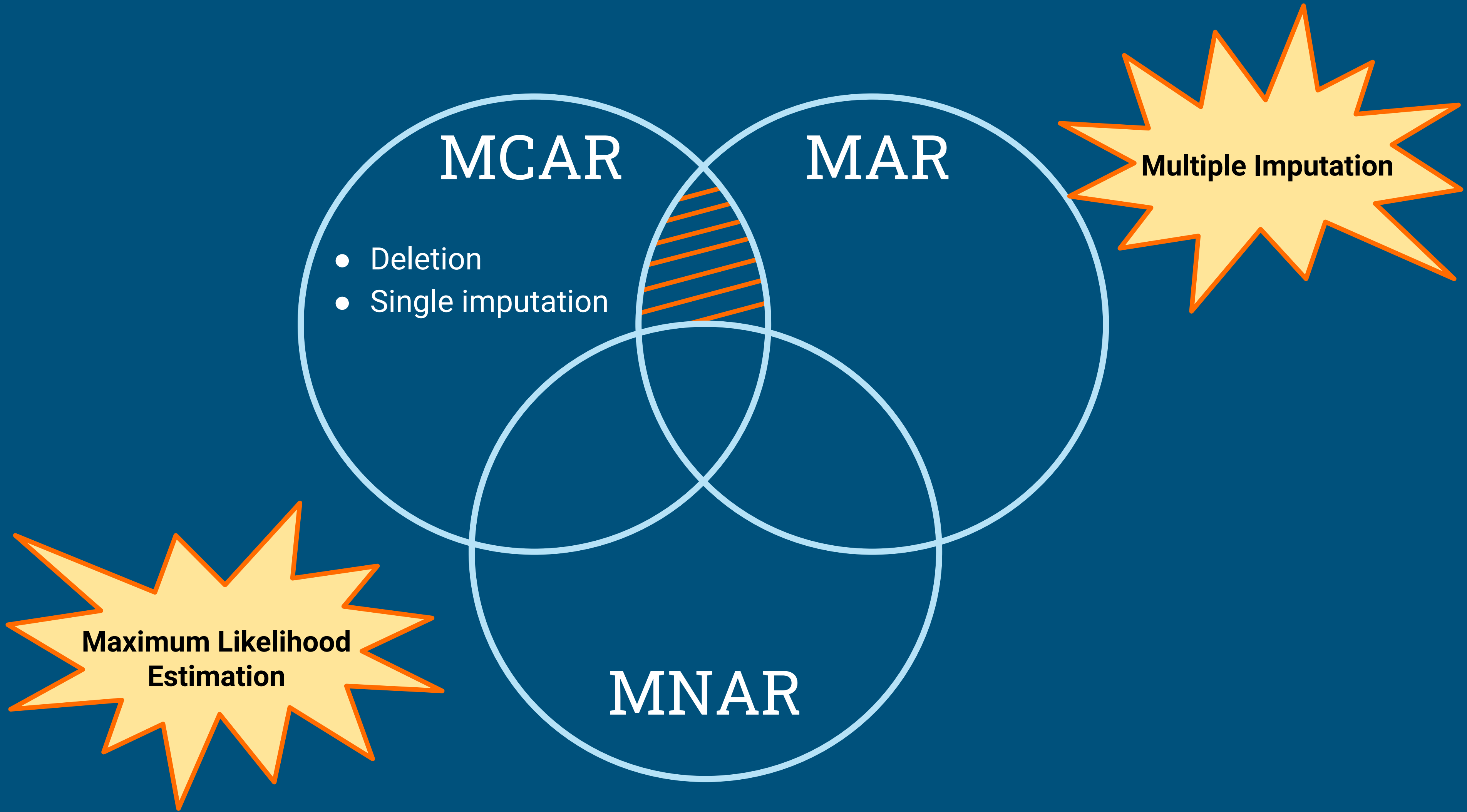

MCAR
MAR
MNAR


MCAR
MAR
MNAR


MCAR
MAR
Deletion
Single imputation
MNAR


MCAR
MAR
MNAR
Deletion
Single imputation


MCAR
Deletion
Single imputation
MAR
MNAR
Multiple Imputation


MCAR
Deletion
Single imputation
MAR
Multiple Imputation
MNAR
Maximum Likelihood Estimation

Approach #1: Multiple Imputation

- **Assumptions**: MAR/MCAR & multivariate normality.
- Create **multiple copies** of the data set, each with different “filled in” values.
- Analyses are carried out on each data set separately and results are pooled together.

Approach #1: Multiple Imputation

- **Assumptions**: MAR/MCAR & multivariate normality.
- Create **multiple copies** of the data set, each with different “filled in” values.
- Analyses are carried out on each data set separately and results are pooled together.
- Three phases:
    - **Imputation**
        - Generate multiple (~20) datasets, each with different estimates of the missing data
        - By generating multiple datasets with imputed values, we are not “putting all our eggs in one basket” and are accounting for the uncertainty in the missing data.

Approach #1: Multiple Imputation

- **Assumptions**: MAR/MCAR & multivariate normality.
- Create **multiple copies** of the data set, each with different “filled in” values.
- Analyses are carried out on each data set separately and results are pooled together.
- Three phases:
  - **Imputation**
    - Generate multiple (~20) datasets, each with different estimates of the missing data
    - By generating multiple datasets with imputed values, we are not “putting all our eggs in one basket” and are accounting for the uncertainty in the missing data.
  - **Analysis**
    - Using the different copies, get multiple parameter estimates & their standard errors.

Approach #1: Multiple Imputation

- **Assumptions**: MAR/MCAR & multivariate normality.
- Create **multiple copies** of the data set, each with different "filled in" values.
- Analyses are carried out on each data set separately and results are pooled together.
- Three phases:
  - **Imputation**
    - Generate multiple (~20) datasets, each with different estimates of the missing data
    - By generating multiple datasets with imputed values, we are not "putting all our eggs in one basket" and are accounting for the uncertainty in the missing data.
  - **Analysis**
    - Using the different copies, get multiple parameter estimates & their standard errors.
  - **Pooling**
    - Combine results from different copies.

Approach #1: Multiple Imputation

- **Imputation**: Data augmentation
    - **Imputation step** (“I-step”):
        - Similar to stochastic regression imputation, we use regression equation to predict values for incomplete variables & add random noise to add variability to the data
    - **Posterior step** (“P-step”):
        - We use Bayesian estimation principles are used to get *new* estimates of the means & covariances and add random noise

Repeat to get multiple copies of data set

- This is a “**two-step iterative algorithm**”
    - We use the updated parameter estimates to construct a new set of imputations for the next copy of the dataset
    - Important to not use consecutive iterations to ensure that copies are independent

Approach #1: Multiple Imputation

- **Analysis**
  - Analyze each copy of the dataset in an identical manner.
  - Ultimately, we will obtain multiple estimates for a given parameter (with multiple standard errors)
- **Pooling**
  - Parameter estimates: take the average!
  - Standard error: Must account for standard errors from imputed datasets (**within-imputation variance**) and also the extent to which the estimates vary across the datasets (**between-imputation variance**).

Approach #1: Multiple Imputation

- **Analysis**
  - Analyze each copy of the dataset in an identical manner.
  - Ultimately, we will obtain multiple estimates for a given parameter (with multiple standard errors)
- **Pooling**
  - Parameter estimates: take the average!
  - Standard error: Must account for standard errors from imputed datasets (**within-imputation variance**) and also the extent to which the estimates vary across the datasets (**between-imputation variance**).

$$W = \frac{\sum SE_t^2}{m}, \qquad B = \frac{\sum(\hat{\theta}_t - \overline{\theta})^2}{m-1}, \longrightarrow SE = \sqrt{W + B + B/m}.$$

Approach #1: Multiple Imputation

- **Analysis**
  - Analyze each copy of the dataset in an identical manner.
  - Ultimately, we will obtain multiple estimates for a given parameter (with multiple standard errors)
- **Pooling**
  - Parameter estimates: take the average!
  - Standard error: Must account for standard errors from imputed datasets (**within-imputation variance**) and also the extent to which the estimates vary across the datasets (**between-imputation variance**).

$$W = \frac{\sum SE_t^2}{m}, \qquad B = \frac{\sum(\hat{\theta}_t - \overline{\theta})^2}{m-1}, \quad \longrightarrow \quad SE = \sqrt{W + B + B/m}.$$

Approach #2: Maximum Likelihood Estimation

- **Assumptions**: MCAR/MAR & multivariate normality
- Rather than "filling in" missing data, we seek to identify the **population parameter** values that have the highest probability of producing the sample data.
- **Method**:
  - Use log likelihood to quantify the standardized distance between observed data and the parameters of interest (e.g. mean).
  - Goal: Minimize these distances!
    - What is this conceptually similar to?

Approach #2: Maximum Likelihood Estimation

- **Assumptions**: MCAR/MAR & multivariate normality
- Rather than “filling in” missing data, we seek to identify the **population parameter** values that have the highest probability of producing the sample data.
- **Method**:
  - Use log likelihood to quantify the standardized distance between observed data and the parameters of interest (e.g. mean).
  - Goal: Minimize these distances!
    - What is this conceptually similar to?
      - **OLS!**

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2}\right]$$

Approach #2: Maximum Likelihood Estimation

- **Assumptions**: MCAR/MAR & multivariate normality
- Rather than "filling in" missing data, we seek to identify the population parameter values that have the highest probability of producing the sample data.
- **Method**:
  - Use log likelihood to quantify the standardized distance between observed data and the parameters of interest (e.g. mean).
  - Goal: Minimize these distances!
    - What is this conceptually similar to?
      - **OLS!**

PDF of normal distribution - describes shape of normal curve

$$logL = \sum_{i=1}^{N} log \boxed{\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2}\right]}$$

Approach #2: Maximum Likelihood Estimation

- **Assumptions**: MCAR/MAR & multivariate normality
- Rather than “filling in” missing data, we seek to identify the population parameter values that have the highest probability of producing the sample data.
- **Method**:
  - Use log likelihood to quantify the standardized distance between observed data and the parameters of interest (e.g. mean).
  - Goal: Minimize these distances!
    - What is this conceptually similar to?
      - **OLS!**

PDF of normal distribution - describes shape of normal curve

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\left(\frac{y_i-\mu}{\sigma}\right)^2}\right]$$

Adds relative probabilities to get **sample log likelihood** (probability of drawing entire sample from normally distributed population)

Approach #2: Maximum Likelihood Estimation

- **Assumptions**: MCAR/MAR & multivariate normality
- Rather than “filling in” missing data, we seek to identify the population parameter values that have the highest probability of producing the sample data.
- **Method**:
  - Use log likelihood to quantify the standardized distance between observed data and the parameters of interest (e.g. mean).
  - Goal: Minimize these distances!
    - What is this conceptually similar to?
      - **OLS!**

PDF of normal distribution - describes shape of normal curve

$$logL = \boxed{\sum_{i=1}^{N}} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\boxed{\left(\frac{y_i-\mu}{\sigma}\right)^2}}\right]$$

Adds relative probabilities to get **sample log likelihood** (probability of drawing entire sample from normally distributed population)

Most important component: Substituting parameter values gives us the distances we are trying to minimize

Approach #2: Maximum Likelihood Estimation

- **Assumptions**: MCAR/MAR & multivariate normality
- Rather than "filling in" missing data, we seek to identify the population parameter values that have the highest probability of producing the sample data.
- **Method**:
  - Use log likelihood to quantify the standardized distance between observed data and the parameters of interest (e.g. mean).
  - Goal: Minimize these distances!
    - What is this conceptually similar to?
      - **OLS!**

Most important component: Substituting parameter values gives us the distances we are trying to minimize

Score values closer to mean
=
Smaller z-score
=
Larger log-likelihood

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\boxed{\left(\frac{y_i-\mu}{\sigma}\right)^2}}\right]$$

Approach #2: Maximum Likelihood Estimation

- How do we find the parameter that gives use the largest log-likelihood (maximum likelihood estimates)?
  - Remember: the population parameters are unknown!
  - We "**audition**" different parameter values by substituting them into the function below.
  - For each audition, we compute the sample log likelihood and see which values give us the largest sample log likelihood.

Most important component: Substituting parameter values gives us the distances we are trying to minimize

Score values closer to mean
=
Smaller z-score
=
Larger log-likelihood

$$logL = \sum_{i=1}^{N} log\left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-.5\boxed{\left(\frac{y_i-\mu}{\sigma}\right)^2}}\right]$$

Approach #2.5: Auxiliary variables 💪

- Adding auxiliary variables to our analysis can help “fine-tune” the missing data approaches.
    - Increase power
    - Reduce bias
- Auxiliary variables: Variables that are **related** to our variable of interest, but do not answer the research question directly
    - Highly correlated with incomplete variable
- Example: 9th grade math performance → 12th grade math exam score