INTRODUCTION TO DATA SCIENCE

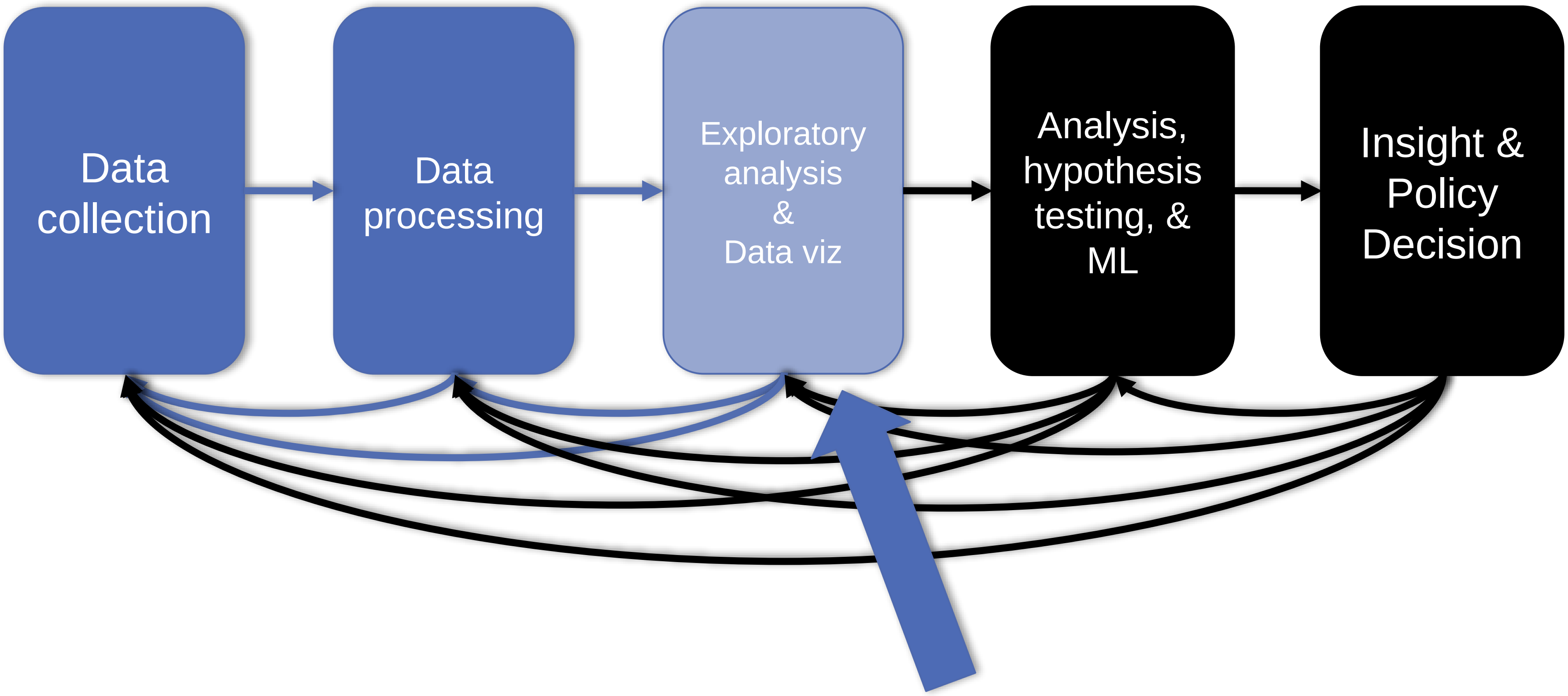
JMCT

Lecture #12 – 06/16/2021

CMSC320
Weekdays
2:00pm – 3:25pm
(… or anytime on the Internet)

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

TODAY'S LECTURE



MISSING DATA

Missing data is information that we want to know, but don't

It can come in many forms, e.g.:

- People not answering questions on surveys
- Inaccurate recordings of the height of plants that need to be discarded
- Canceled runs in a driving experiment due to rain

Could also consider missing columns (no collection at all) to be missing data …

[JA]

KEY QUESTION

Why is the data missing?

- What mechanism is it that contributes to, or is associated with, the probability of a data point being absent?
- Can it be explained by our observed data or not?

The answers drastically affect what we can ultimately do to compensate for the missing-ness

[JA]

COMPLETE CASE ANALYSIS

Delete all tuples with any missing values at all, so you are left only with observations with all variables observed

```
# Clean out rows with nil values
df = df.dropna()
```

Default behavior for libraries for analysis (e.g., regression)

- We'll talk about this much more during the Stats/ML lectures

This is the simplest way to handle missing data. In some cases, will work fine; in others, ?????????????:

- Loss of sample will lead to variance larger than reflected by the size of your data
- May bias your sample

[JA]

EXAMPLE

Dataset: Body fat percentage in men, and the circumference of various body parts **[Penrose et al., 1985]**

Question: Does the circumference of certain body parts predict body fat percentage?

Given complete data, how would you answer this ?????????

One way to answer is regression analysis:

- One or more independent variables ("predictors")
- One dependent variables ("outcome")

What is the relationship between the predictors and the outcome?

What is the conditional expectation of the dependent variable given fixed values for the dependent variables?

LINEAR REGRESSION

Assumption: relationship between variables is linear:

- (We'll relax linearity, study in more depth later.)

Population Y-Intercept

Population Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

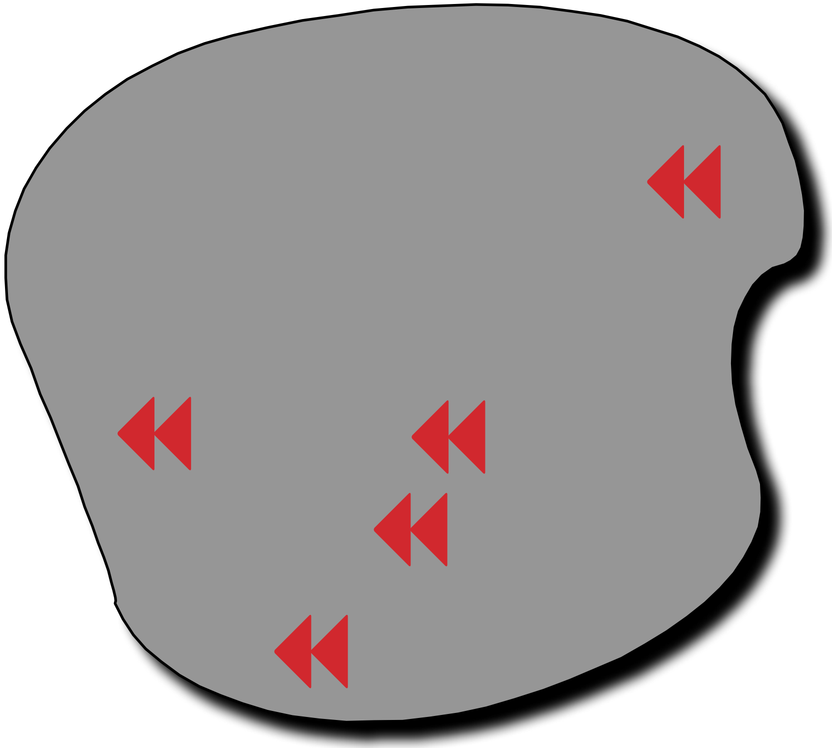
Dependent Variable (e.g., ????????)

Independent Variable(s) (e.g., ?????????)

[WF]

POPULATION & SAMPLE REGRESSION MODELS

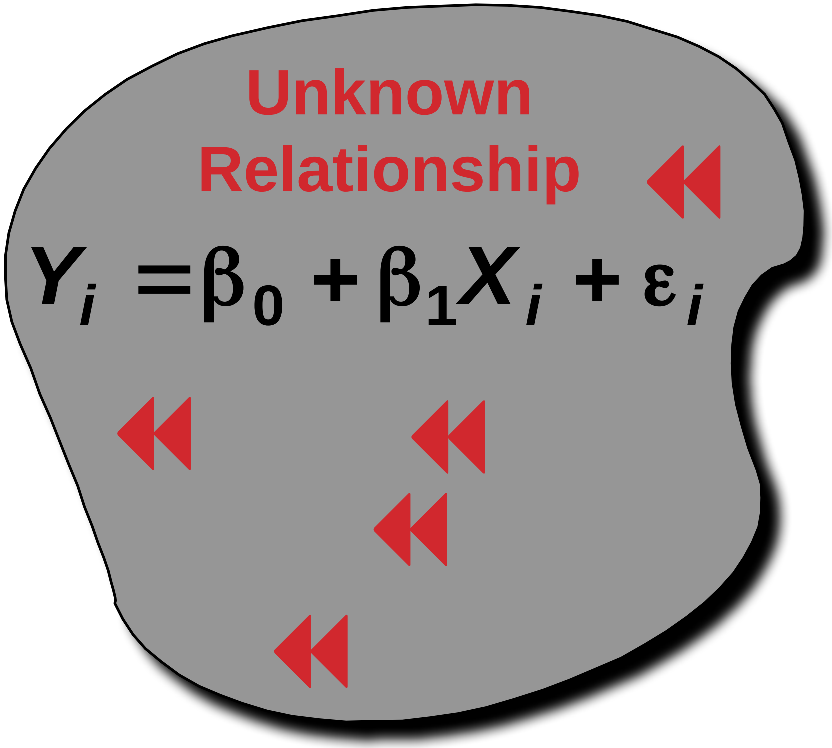
Population



[WF]

POPULATION & SAMPLE REGRESSION MODELS

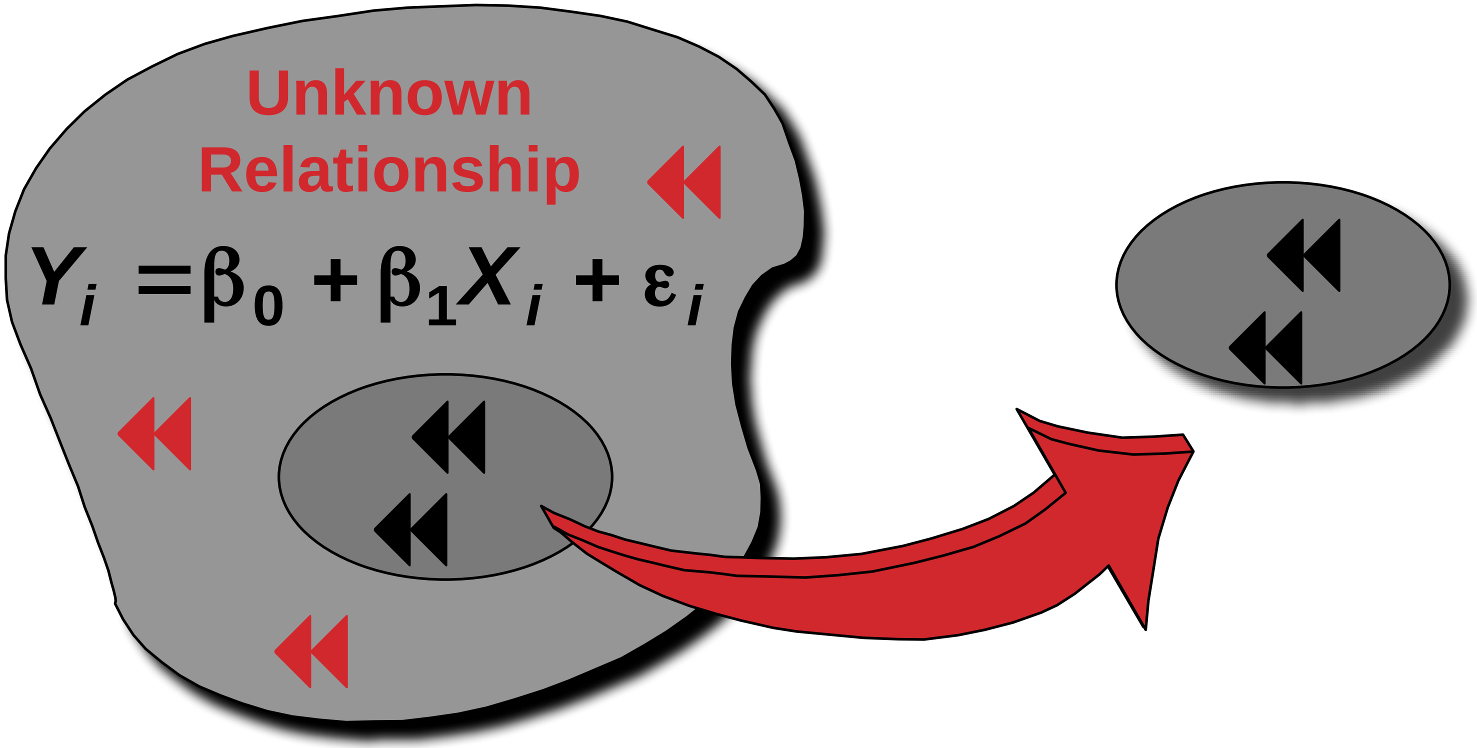
Population



[WF]

POPULATION & SAMPLE REGRESSION MODELS

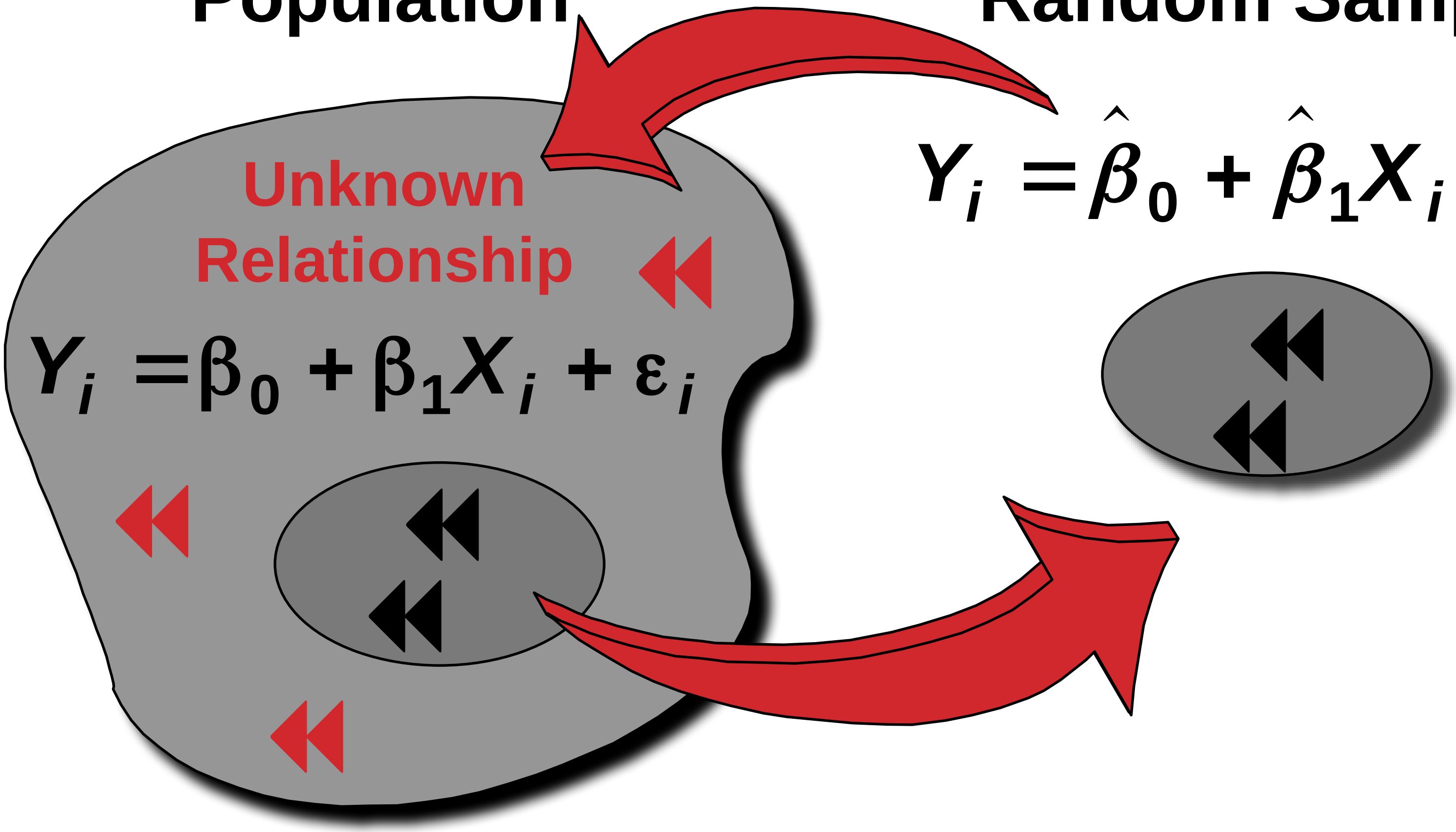
Population **Random Sample**



[WF]

POPULATION & SAMPLE REGRESSION MODELS

Population **Random Sample**



[WF]

SINGLE IMPUTATION

Mean imputation: imputing the average from observed cases for all missing values of a variable

Hot-deck imputation: imputing a value from another subject, or “donor,” that is most like the subject in terms of observed variables

- Last observation carried forward (LOCF): order the dataset somehow and then fill in a missing value with its neighbor

Cold-deck imputation: bring in other datasets

Old and busted:

- All fundamentally impose too much precision.
- Have uncertainty over what unobserved values actually are
- Developed before cheap computation

MULTIPLE IMPUTATION

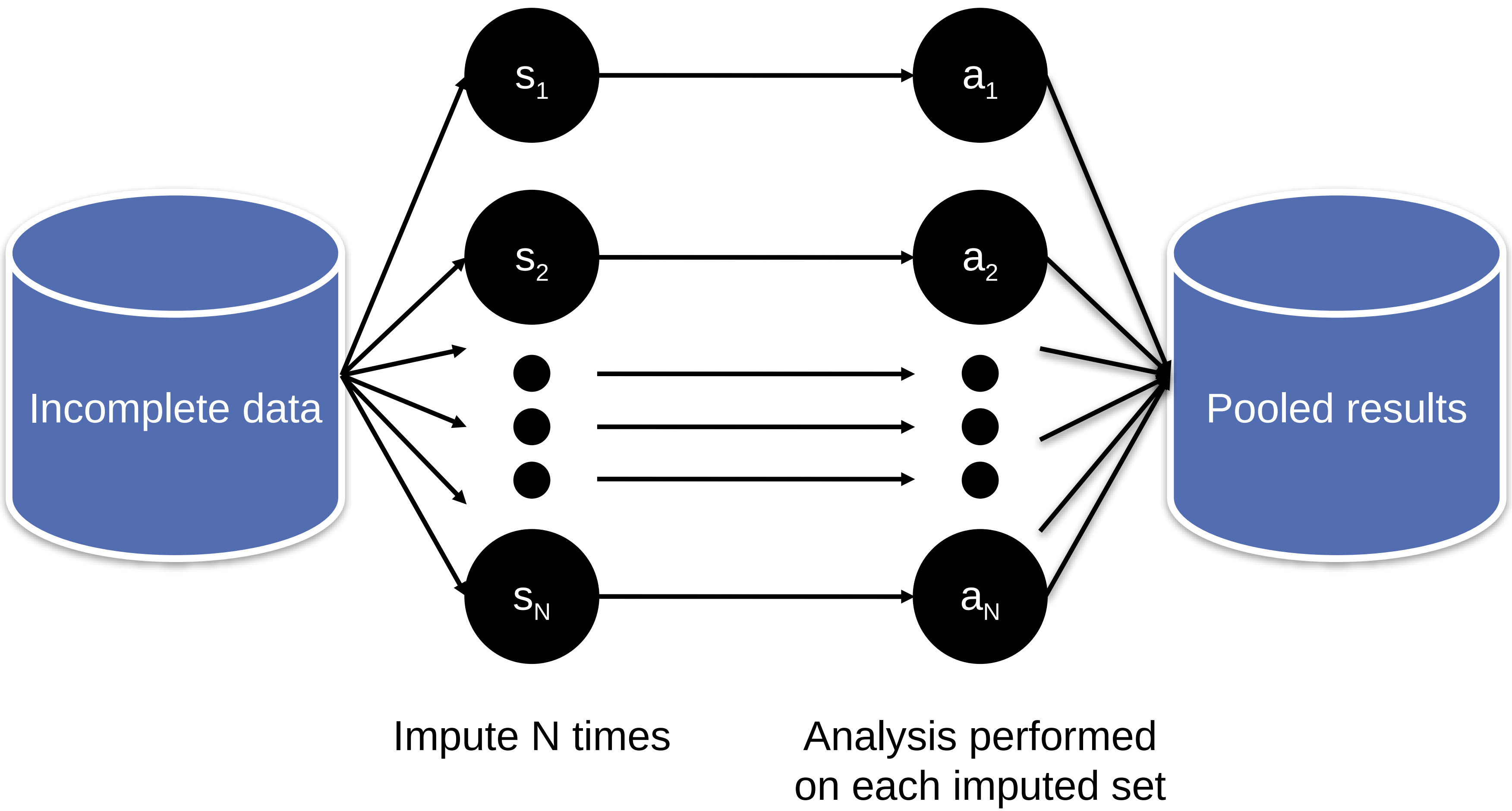
Developed to deal with noise during imputation

- Impute once treats imputed value as observed

We have uncertainty over what the observed value would have been

Multiple imputation: generate several random values for each missing data point during imputation

IMPUTATION PROCESS



TINY EXAMPLE

X	Y
32	2
43	?
56	6
25	?
84	5

Independent variable: X
Dependent variable: Y
We assume Y has a linear relationship with X

LET’S IMPUTE SOME DATA!

Use a predictive distribution of the missing values:

- Given the observed values, make random draws of the observed values and fill them in.
- Do this N times and make N imputed datasets

X	Y
32	2
43	5.5
56	6
25	8
84	5

X	Y
32	2
43	7.2
56	6
25	1.1
84	5

For very large values of N=2 …

INFERENCE WITH MULTIPLE IMPUTATION

Now that we have our imputed data sets, how do we make use of them?   ???????????

- Analyze each of the separately

X	Y
32	2
43	5.5
56	6
25	8
84	5

Slope	-0.8245
Standard error	6.1845

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

X	Y
32	2
43	7.2
56	6
25	1.1
84	5

Slope	4.932
Standard error	4.287

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

POOLING ANALYSES

Pooled slope estimate is the average of the N imputed estimates

Our example, β_{1p} = (4.932-.8245) x 0.5 = 2.0538

$$s = \frac{\sum Zi}{m} + \left(1 + \frac{1}{m}\right) \mathbf{x} \frac{1}{m-1} * \sum(\beta 1i - \boldsymbol{\beta}_{1p})^2$$

Where Z_i is the standard error of the imputed slopes

Our example: (4.287 + 6.1845)/2 + (3/2)*(16.569) = 30.08925

Standard error: take the square root, and we get 5.485

BAYESIAN IMPUTATION

Establish a prior distribution:

- **Some distribution of parameters of interest θ before considering the data, $P(\theta)$**
- **We want to estimate θ**

Given θ, can establish a distribution $P(X_{obs}|\theta)$

Use Bayes Theorem to establish $P(\theta|X_{obs})$ …

- **Make random draws for θ**
- **Use these draws to make predictions of Y_{miss}**

HOW BIG SHOULD N BE?

Number of imputations N depends on:

- Size of dataset
- Amount of missing data in the dataset

Some previous research indicated that a small N is sufficient for efficiency of the estimates, based on:

- (1 +)-1
- N is the number of imputations and λ is the fraction of missing information for the term being estimated [Schaffer 1999]

More recent research claims that a good N is actually higher in order to achieve higher power **[Graham et al. 2007]**

MORE ADVANCED METHODS

Interested? Further reading:

- Regression-based MI methods
- Multiple Imputation Chained Equations (MICE) or Fully Conditional Specification (FCS)
  - Readable summary from JHU School of Public Health: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3074241/
- Markov Chain Monte Carlo (MCMC)
  - We'll cover this a bit, but also check out CMSC422!