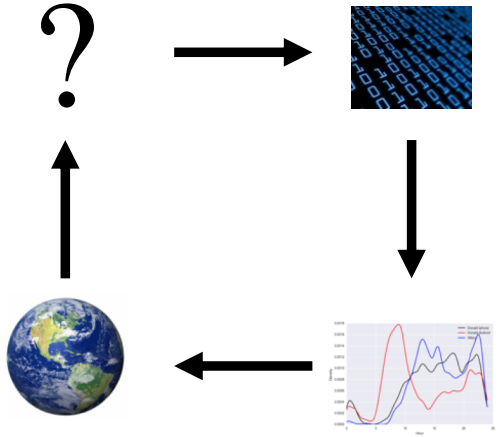
Data Science 100

Principles & Techniques of Data Science

Slides by:

Deborah Nolan

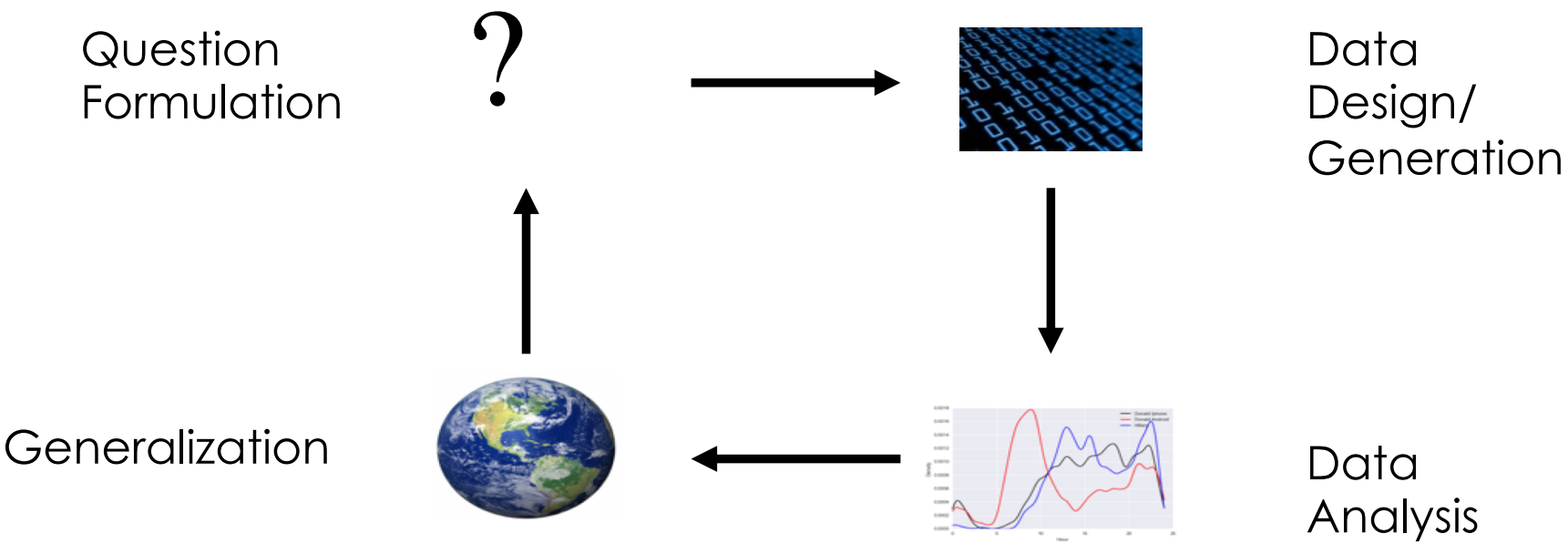
deborah_nolan@berkeley.edu



Announcements for Today

- *We have asked for permission to increase the class size to enroll about 100 people from the wait list…. Stay tuned*
- We will try using Google forms today
- Slides and notes from lecture available online at http://ds100.org/fa19
- HW 1 is due 11:59 Wednesday Sep 4
- Office hours are found at http://ds100.org/fa19/calendar

Data Life Cycle



START SIMPLE

QUESTION:

What is the typical family size (children only)?

START SIMPLE

DATA:

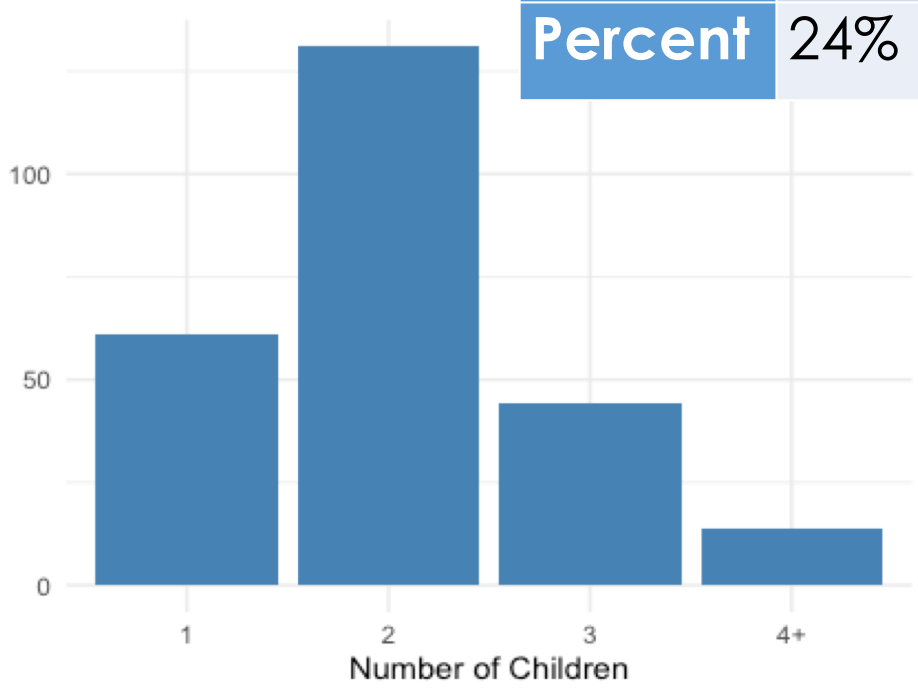
Survey of all 250 students enrolled in Data 100 in Fall 2017 and asked their family size

	1	2	3	4+
Counts	61	131	44	14
Percent	24%	52%	18%	6%

START SIMPLE

ANALYSIS:

	1	2	3	4+
Counts	61	131	44	14
Percent	24%	52%	18%	6%



Can we provide a summary statistic?

DETOUR:
Why is the sample mean such a desirable summary?

Summarizing the Data

DATA: $x_1, x_2, \ldots, x_n$ where n is 250 in our example

ERROR: $x_1 - c, x_2 - c, \ldots, x_n - c$

LOSS: $l: R \rightarrow R^+$

Summarizing the Data

AVERAGE LOSS: $\frac{1}{n}\sum_{i=1}^{n} l(x_i - c)$

AKA EMPIRICAL RISK

Minimize the empirical risk

Minimize the Average Loss

$$\frac{1}{n}\sum_{i=1}^{n} l(x_i - c) = \frac{1}{n}\sum_{i=1} (x_i - c)^2$$

Refresher

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\frac{1}{n} \sum_{i=1}^{n} (ax_i + b) = a\bar{x} + b$$

Minimize the Average Loss

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

The Sample Average Minimizes Empirical Risk

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 <= \frac{1}{n}\sum_{i=1}^{n}(x_i - c)^2$$

This is the Sample Variance

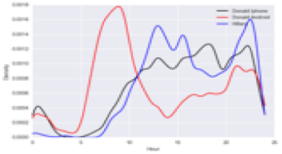
Data Life Cycle

Question Formulation

?

Data Design/ Generation

In some cases we do not want/need to Generalize



Data Analysis

Consider the Question Carefully

What is the typical family size (children only)?

How well can we measure this?

What are we trying to measure?

Focus the Question

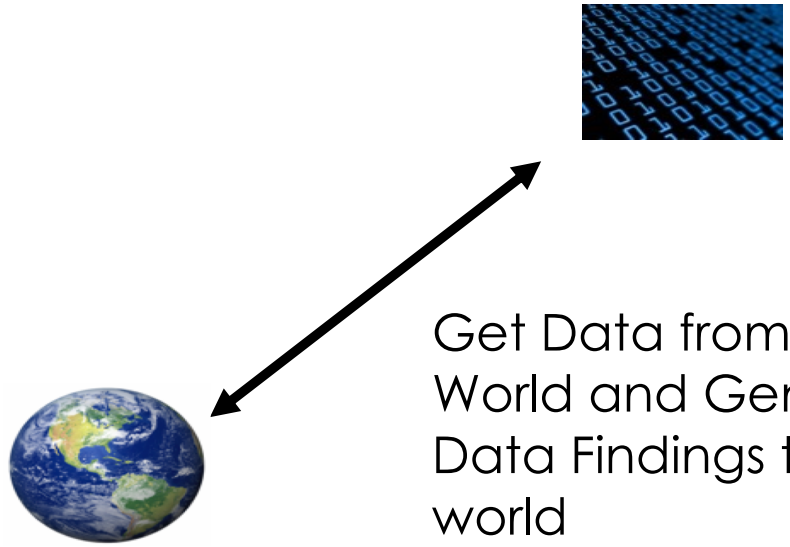
From Female Fertility Perspective:

- WHERE
- WHEN
- WHO
- WHAT

Focus the Question

The Question gives focus to the Population that we want to study

Data Life Cycle



Data
Design/
Generation

Get Data from the
World and Generalize
Data Findings to the
world

Generalization

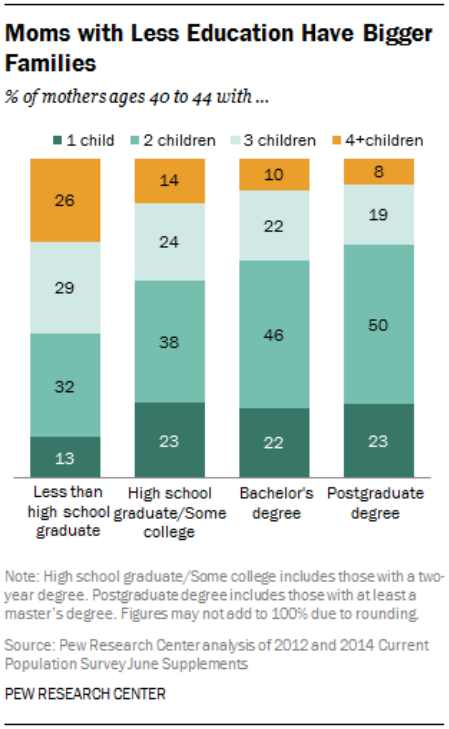
How Well Does our Data 100 class represent the group of interest?

- Mothers of children at UC Berkeley
- Measure the mothers via the children
- Mothers who are 40-44 in 2014

How might these characteristics impact the estimate of the number of children a US woman bears in her lifetime in 2014?

Bias up, Bias Down, Not impact

According to Pew Research Center

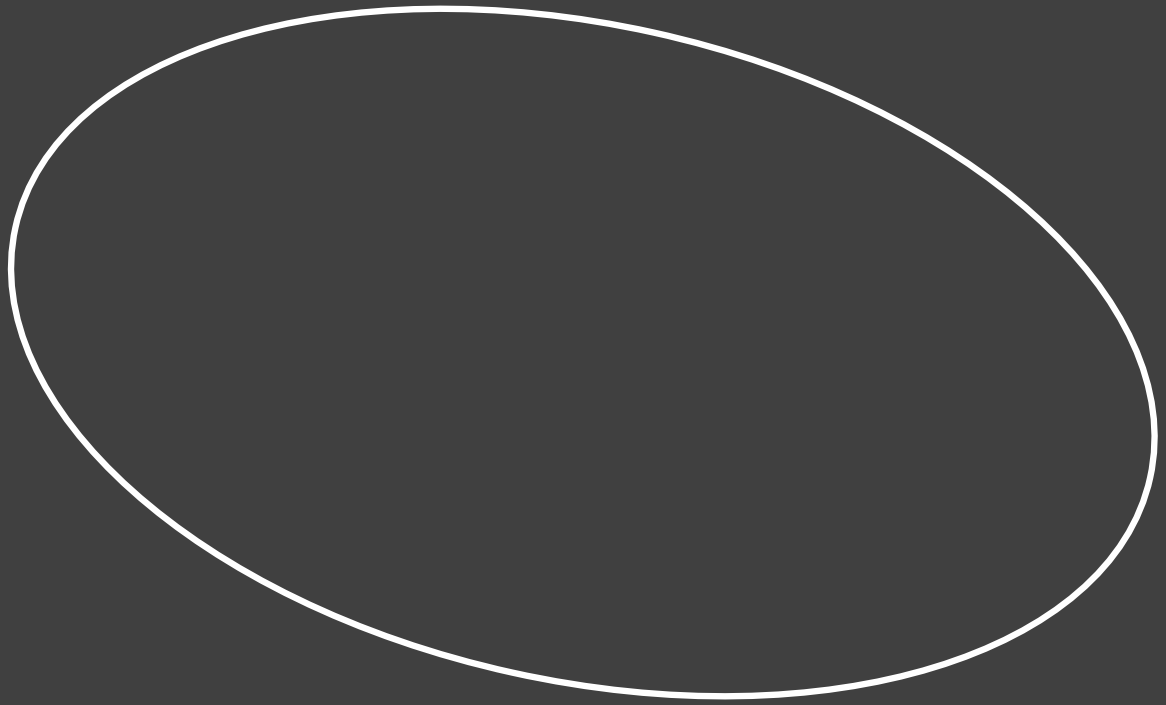


How might this impact the Data 100 average?

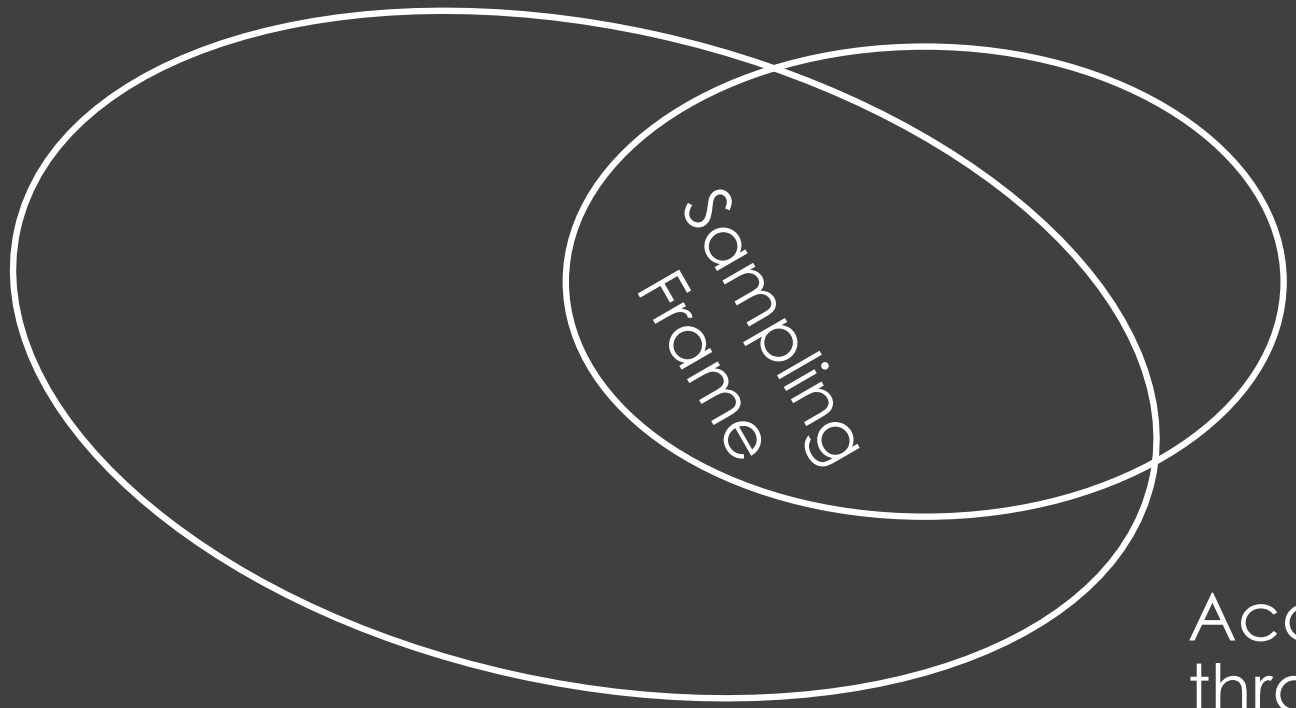
The data used in these analyses are designed to assess women's fertility, and as such a "mother" is here defined as any woman who has given birth. However, many women who do not bear their own children are indeed mothers.

http://www.pewsocialtrends.org/2015/05/07/family-size-among-mothers/

Population of Interest

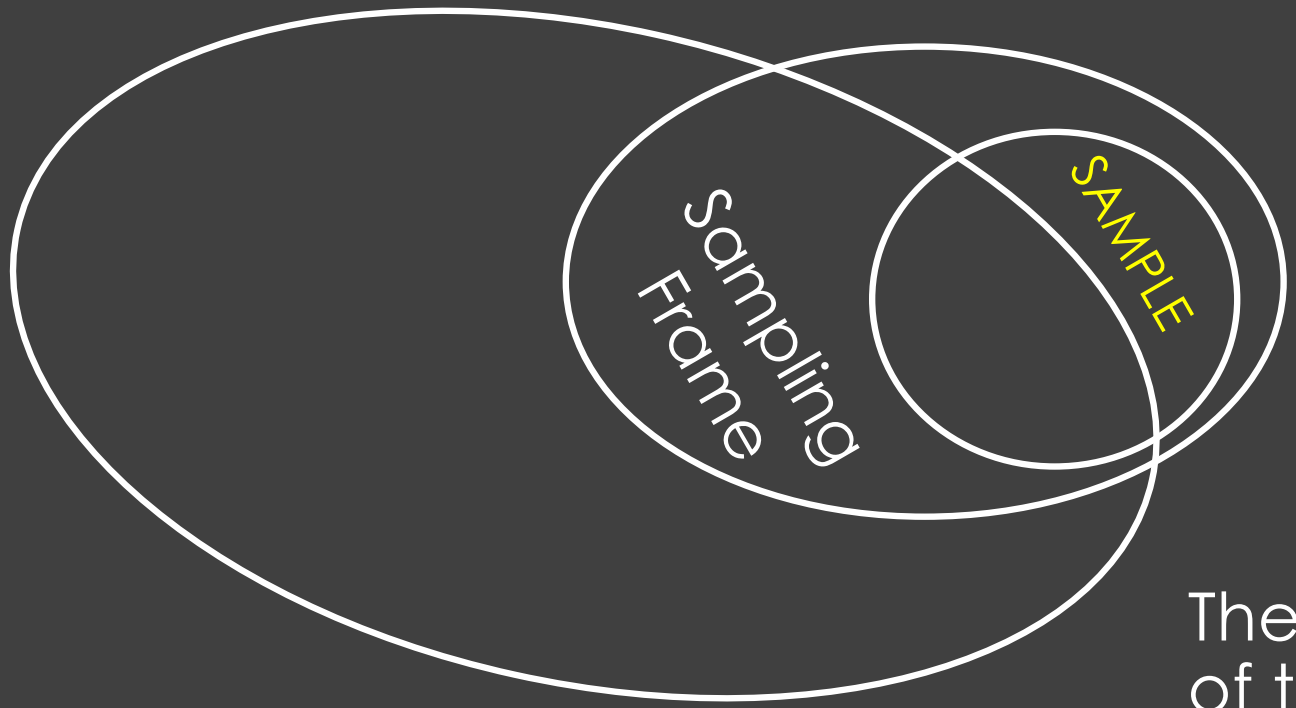


Population of Interest



Access the Population
through the Frame

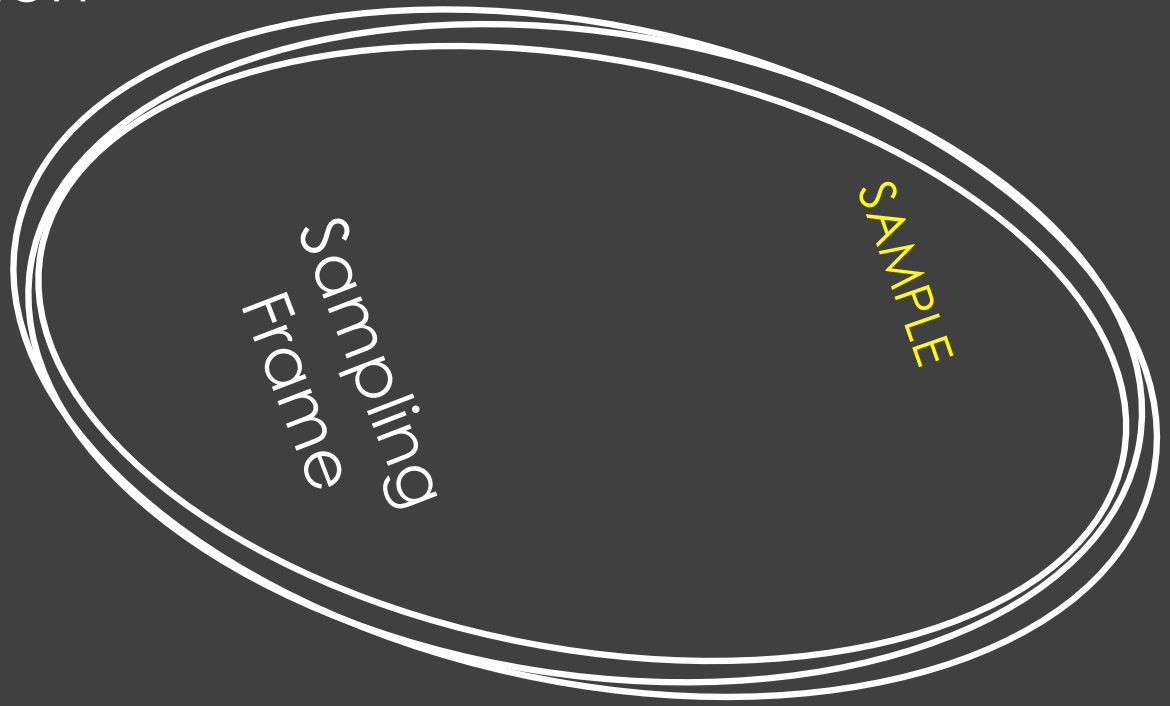
Population of Interest



The Sample is a subset of the Frame

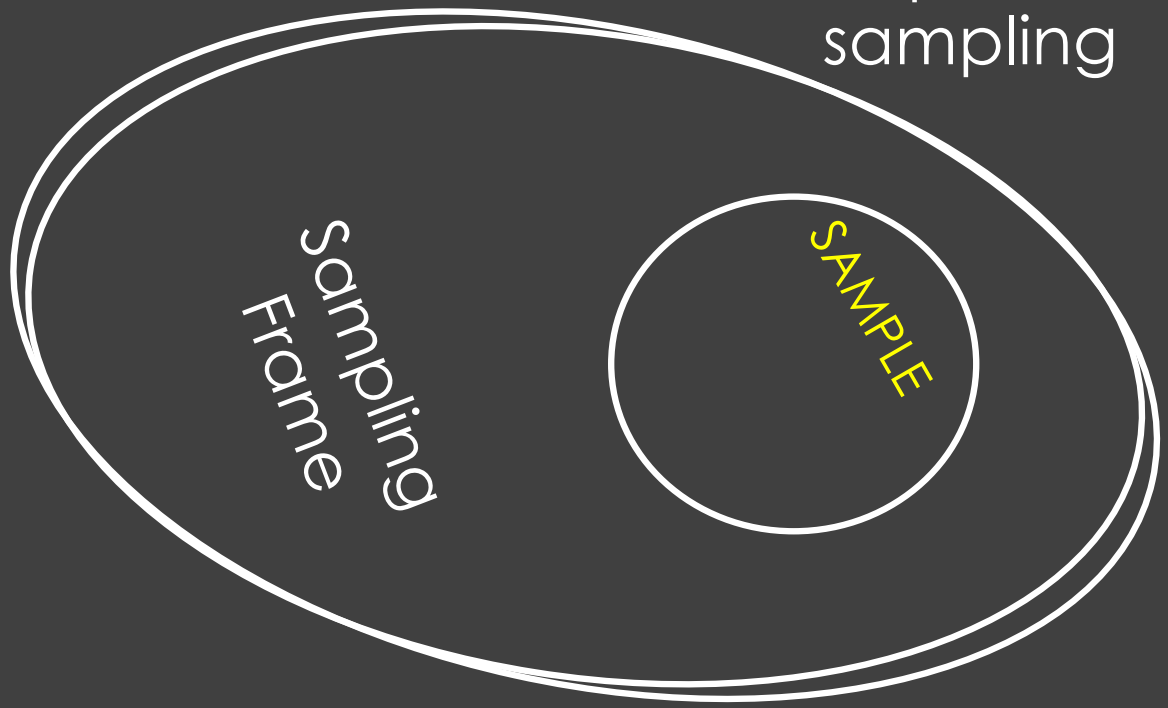
Sample =
Sampling Frame =
Population

Scenario: Census



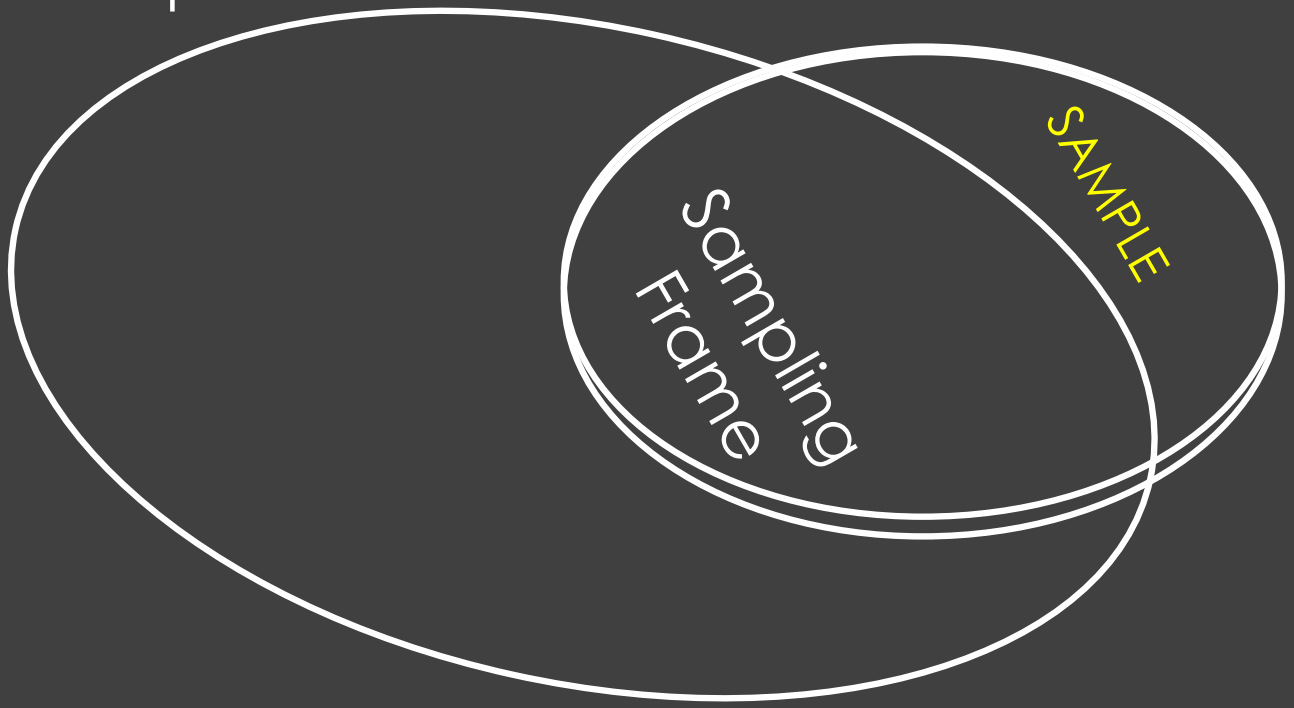
Sampling Frame = Population

Scenario: Access to all members of the Population when sampling

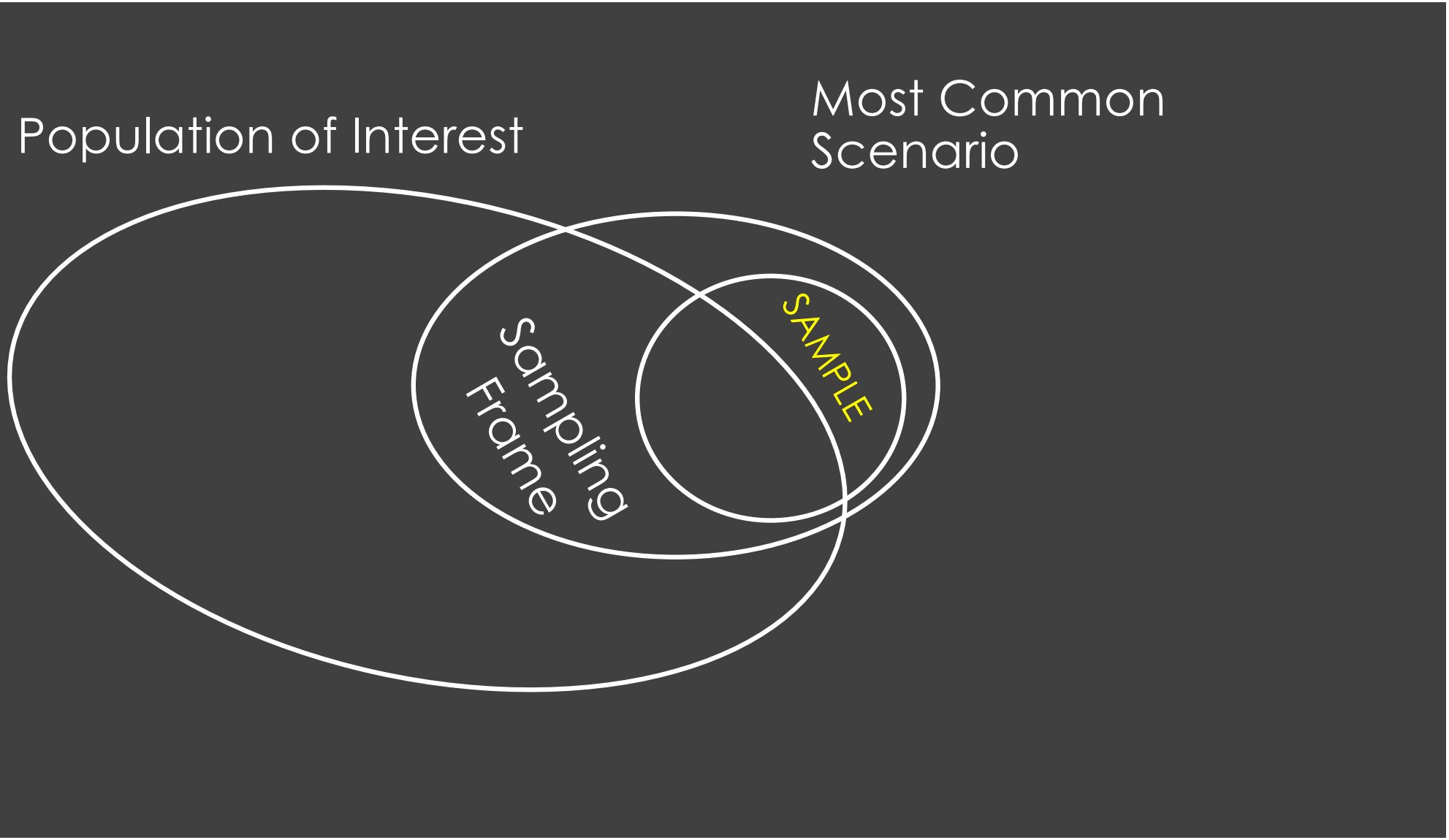


Sampling Frame =
Sample

Scenario:
Administrative Data



Most Common Scenario



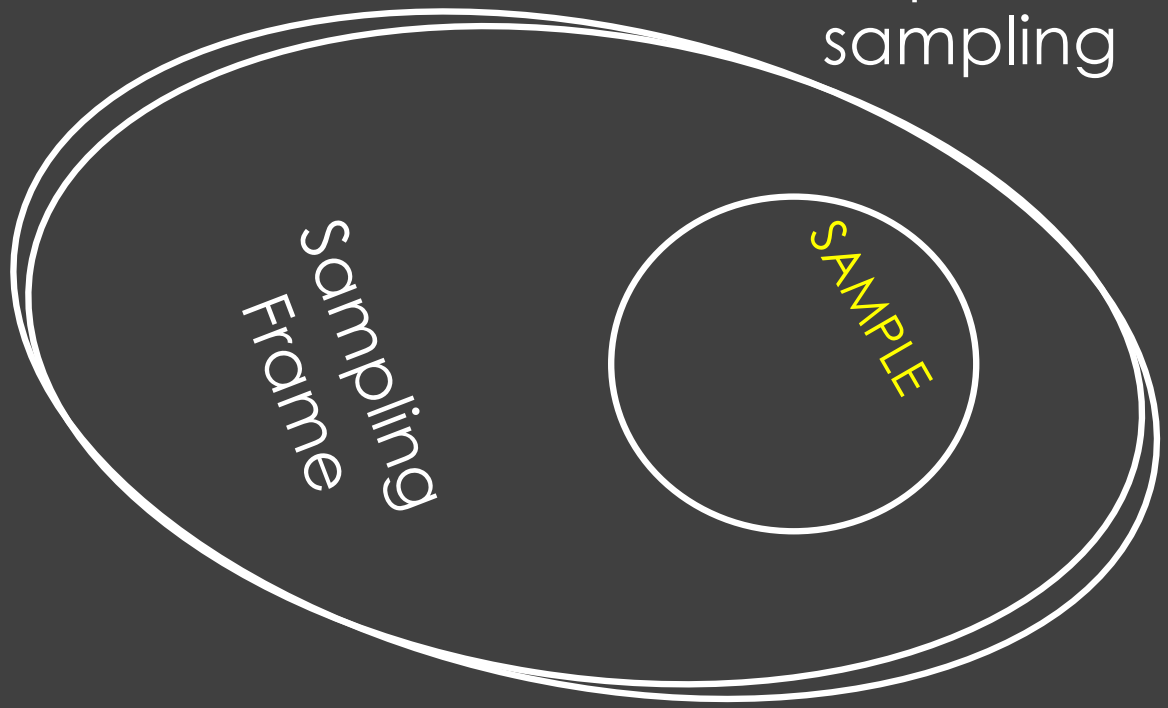
How are the data generated?

- What is the population of interest?
- What is the sampling frame?
- How are the data generated?

DETOUR:
1. The simple random sample
2. Why is a probability sample so desirable?

Sampling Frame = Population

Scenario: Access to all members of the Population when sampling



HOW IS THE SAMPLE TAKEN?

The Simple Random Sample

- *Suppose we have a population with N subjects*
- *We want to sample **n** of them*
- ***The SRS is a random sample where every unique subset of n subjects has the same chance of appearing in the sample***
- This means each person is equally likely to be in the sample

The Advantages of a SRS

- *Representative: The sample tends to look like the population*
- *Statistics based on the sample tend to be close to statistics based on the population*
- *We can provide typical deviations of sample statistics from population values.*
- *AND MORE…*

Start Simple

- *Suppose our population contains only 10 mothers and we take a **Simple Random Sample** of 3 for our survey.*

	Number of Children			
	1	2	3	4+
Count	2	4	3	1
Proportion	20%	40%	30%	10%

Formal Set Up

	Number of Children			
	1	2	3	4+
Count	2	4	3	1
Proportion	20%	40%	30%	10%

X_1 The number of children for the first mother chosen

X_2 The number of children for the second mother chosen

X_3 The number of children for the third mother chosen

Formal Set Up

	Number of Children			
	1	2	3	4+
Count	2	4	3	1
Percent	20%	40%	30%	10%

X_1 The number of children for the first mother chosen

	Probability Distribution			
x	1	2	3	4+
$P(X_1 = x)$				

	Probability Distribution			
x	1	2	3	4+
$P(X_1 = x)$				

X_1 The number of children for the first mother chosen

What is the expected value of X_1?

$\mathbb{E}(X_1)$

X_2 number of children for the 2nd mother chosen

	Probability Distribution			
x	1	2	3	4+
$P(X_2 = x)$				

X_2 number of children for the 2nd mother chosen

DETOUR CONTINUED: Why is the expected value a desirable summary of a probability distribution?

Random Variables

Random Variables: $X_1, X_2, \ldots, X_n$

Random ERROR: $X_1 - c, X_2 - c, \ldots, X_n - c$

LOSS: $l: R \to R^+$

Summarizing the Probability Distribution

EXPECTED LOSS:

AKA RISK $\mathbb{E}[l(X-c)] = \mathbb{E}[(X-c)^2]$

Minimize the risk

Properties of Expected Value

$$\mathbb{E}(X) = \sum_{j=1}^{m} x_j P(X = m_j)$$

$$\mathbb{E}(aX + b)$$

Minimize the Risk

$$\mathbb{E}[(X - c)^2]$$

The Expected Value Minimizes Risk

$$\mathbb{E}[X - \mathbb{E}(X)]^2 \leq \mathbb{E}[(X - c)^2]$$

This side is the Variance

Data Life Cycle

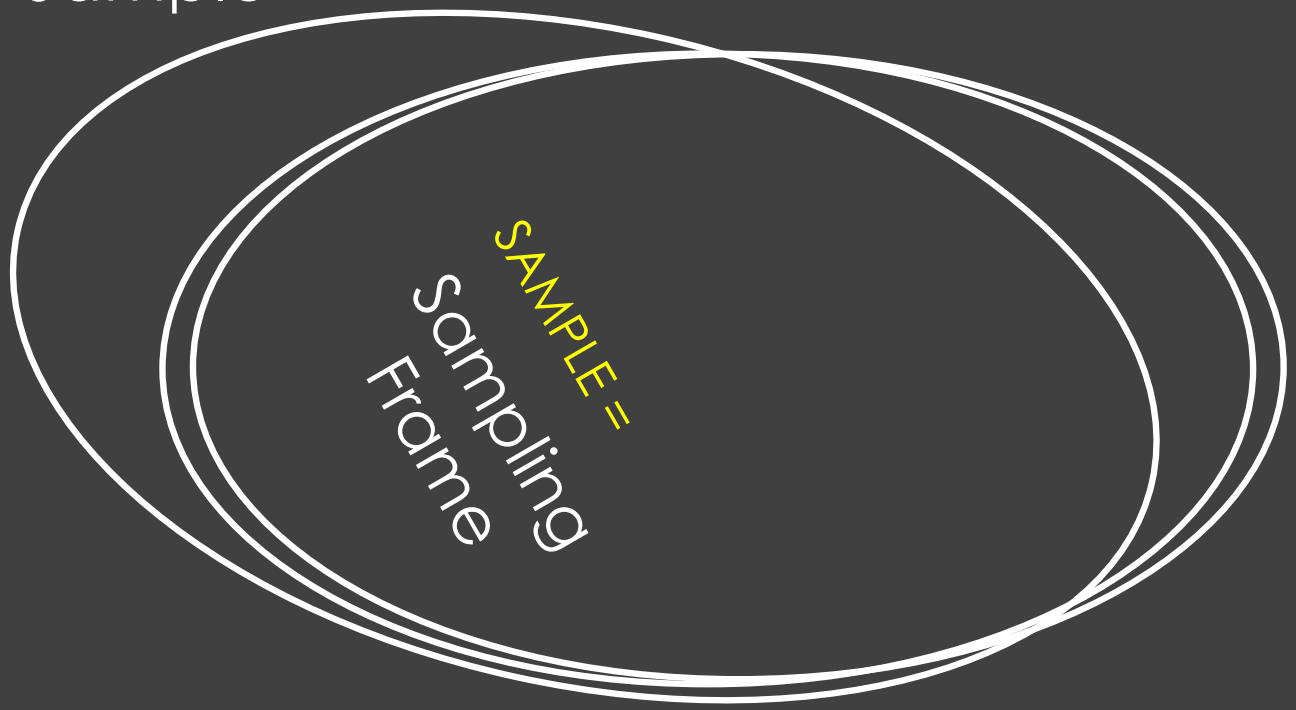
Data
Design/
Generation

Probability Samples give us Representative Data where the sample average is well behaved and an accurate estimate of the population average

Generalization

Sampling Frame = Sample

Scenario: Administrative Data



Can we make up for no Probability Sample with Big Data?

Sample and Population Averages

The gap between these is based on three things:

- Data **quality** measure (the correlation between the sampling technique and the response)
- Data **quantity** measure (how big is the sample relative to the population)
- **Problem difficulty** measure (how variable is the response)

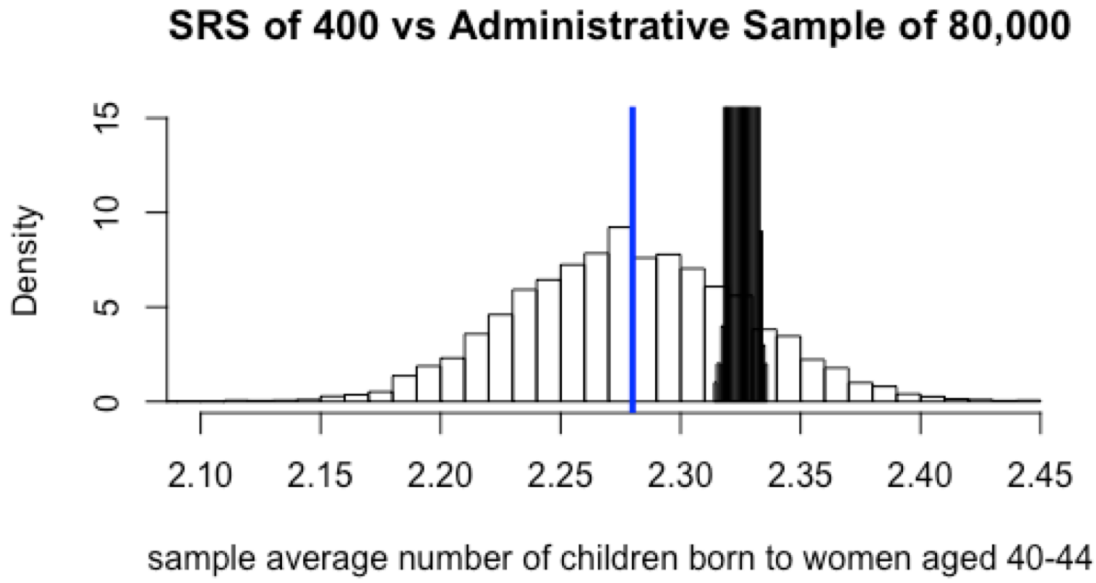
Meng 2018, Annals Applied Probability

Sample and Population Averages

- Probabilistic sampling ensures high data quality by eliminating selection bias and confounding
- When combining data sources for population inferences, those relatively tiny but higher quality sources should be given far more weights than suggested by their sizes.

Active Area of Research Area

Large Administrative Data vs Small SRS



Data Life Cycle