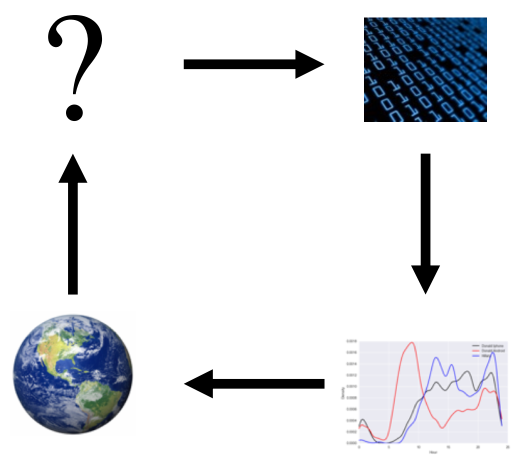
Data Science 100

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

We will give a simple example

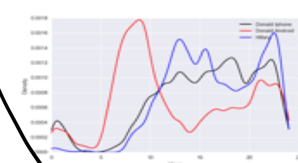
Data Life Cycle

Question Formulation ?

These 2 pieces are crucial

Generalization

Data Design/ Generation



Data Analysis

Most of our focus is on this part of the cycle

START SIMPLE

QUESTION:

What is the typical family size (children only)?

START SIMPLE

DATA:

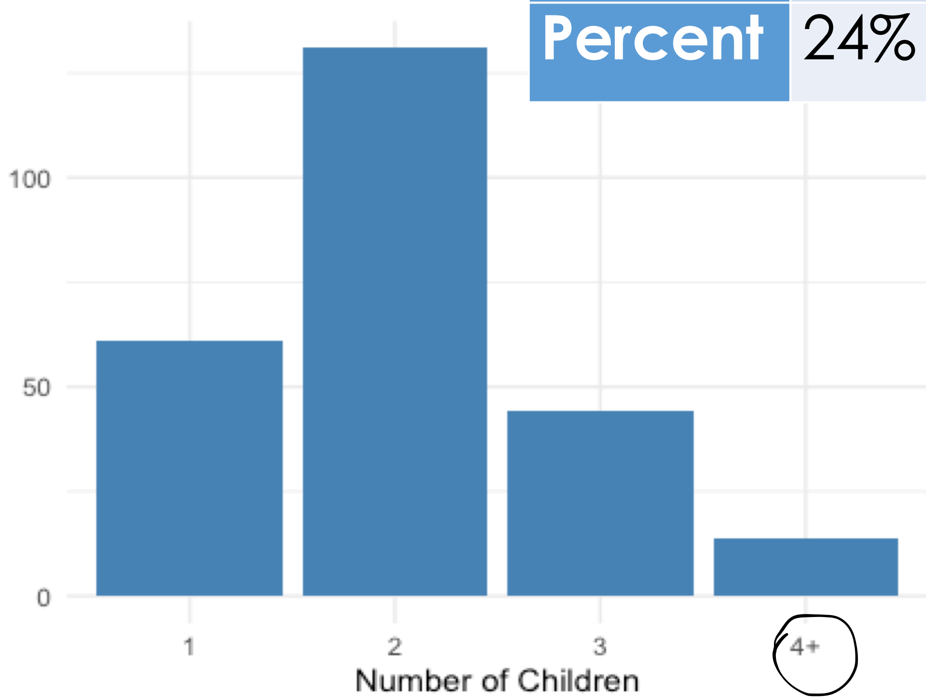
Survey of all 250 students enrolled in Data 100 in Fall 2017 and asked their family size

	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%



Bar Chart is a good visual summary

For now we ignore the + and treat it as 4

Can we provide a summary statistic?

How about the mean?

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

Summarizing the Data

We want a single numeric summary of our data: c

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

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

$\boxed{\text{LOSS: } l: R \to R^+}$

The loss function maps errors to the nonnegative values.
It represents the 'cost' of an error.

We want c to be close to our data.

So, we look at the error between an observation and c

$x_1 - c$

If c is 2 and x_1 is 2 then the error is 0
If x_1 is 4 then it is 2

Summarizing the Data

The Average Loss simply averages the loss $l(x_1 - c), \ldots$ over the data values

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

AKA EMPIRICAL RISK

Minimize the empirical risk

We want to $\min\limits_{c} \frac{1}{n} \sum_{i=1}^{n} l(x_i - c)$

We need to specify the loss function to do this.

Minimize the Average Loss

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

This is l_2 loss. We also call it squared error.

Before we minimize we give a short refresher about sums and averages

It is the most commonly used loss function because it has several useful properties

Refresher

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

Recall the sample mean

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

We will use this property several times

$$= \frac{1}{n}\sum_{i=1}^{n} a x_i + \frac{1}{n}\sum_{i=1}^{n} b$$

$$= a\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n} nb$$

$$= a\bar{x} + b$$

A simple approach that does not involve calculus

Minimize the Average Loss

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

add and subtract

$$= \frac{1}{n}\sum_{i=1}^{n}\left[(x_i-\bar{x})^2 + 2(\bar{x}-c)(x_i-\bar{x}) + (\bar{x}-c)^2\right]$$

$$= \boxed{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2} + 2(\bar{x}-c)\underbrace{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})}_{=0} + \boxed{(\bar{x}-c)^2}$$

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

$$= \bar{x} - \bar{x}$$

We have

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

To minimize wrt c

there is no c here

The minimum is when

$c=\bar{x}$

The Sample Average Minimizes Empirical Risk

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

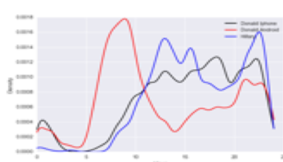
This is the Sample Variance

Let's step back and consider the question

Data Life Cycle

Question Formulation ?

Data Design/ Generation



Data Analysis

~~In some cases we do not want/need to Generalize~~

When we have data about all of the individuals that interest us, we don't need to generalize further

Consider the Question Carefully

What is the typical family size (children only)?

How well can we measure this?

What are we trying to measure?

Families come in all different shapes and sizes.

Suppose we are most interested in the # children a woman gives birth to

Focus the Question

We may be interested in comparing # children a woman has today to 20 or 50 years ago

From Female Fertility Perspective:

Some Questions to Help us Focus Our Question

- WHERE US
- WHEN 2016
- WHO Females 40-44 (have had all the children they are going to by now)
- WHAT #births

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

In order to generalize from data to the population of interest, our sample needs to look like the population

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

Tend to be more highly educated than the population which would bias down

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

our sample should be OK here

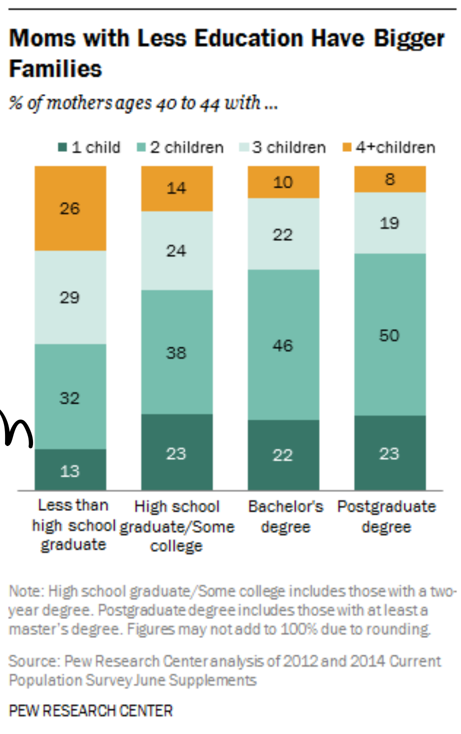
How might these characteristics impact the estimate of the number of children a US woman bears in her lifetime in 2014?

Bias up, Bias Down, Not impact

A mother w/ 4 children has more chances of getting into the sample than a mother w/ 1 child.

Bias Up   This is called Size Biased Sampling

According to Pew Research Center



Shows the education bias,

How might this impact the Data 100 average?

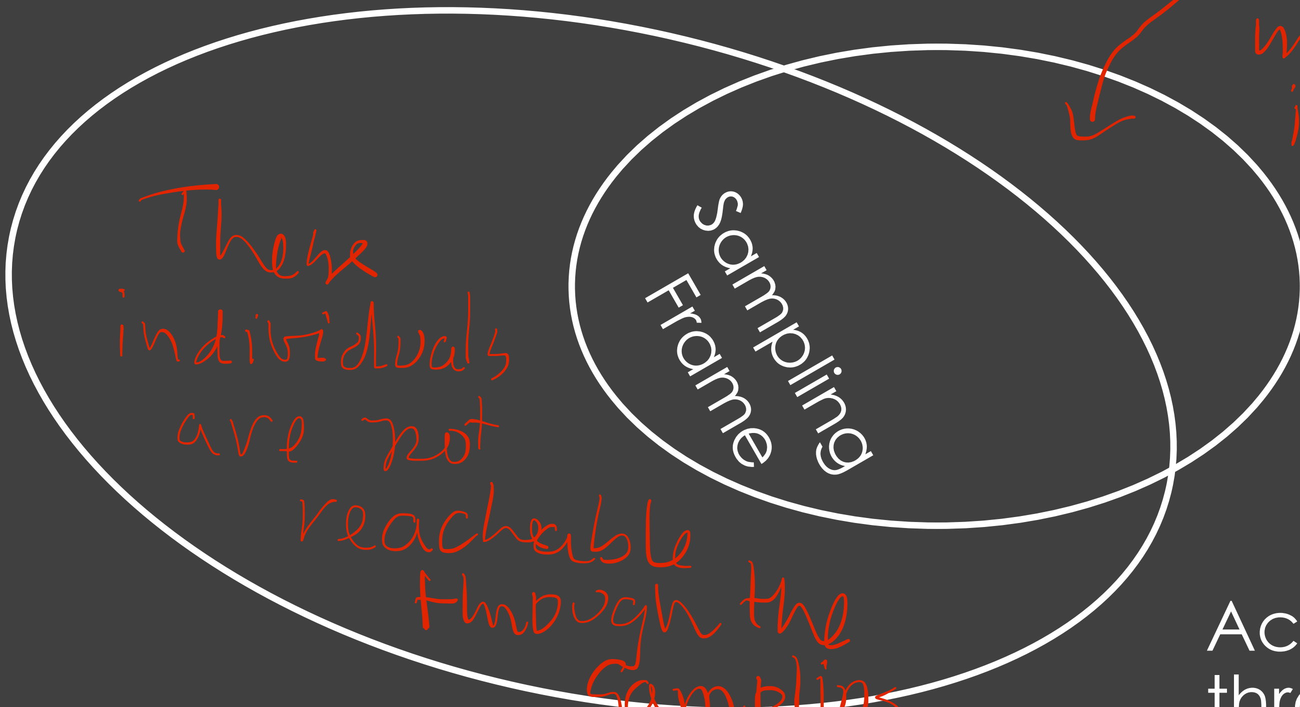
The data used in these analyses are designed to assess women's fertility, and as such a "mother" is here defined as any woman who has given birth. However, many women who do not bear their own children are indeed mothers.

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

Population of Interest

The individuals
we want to
study

Population of Interest

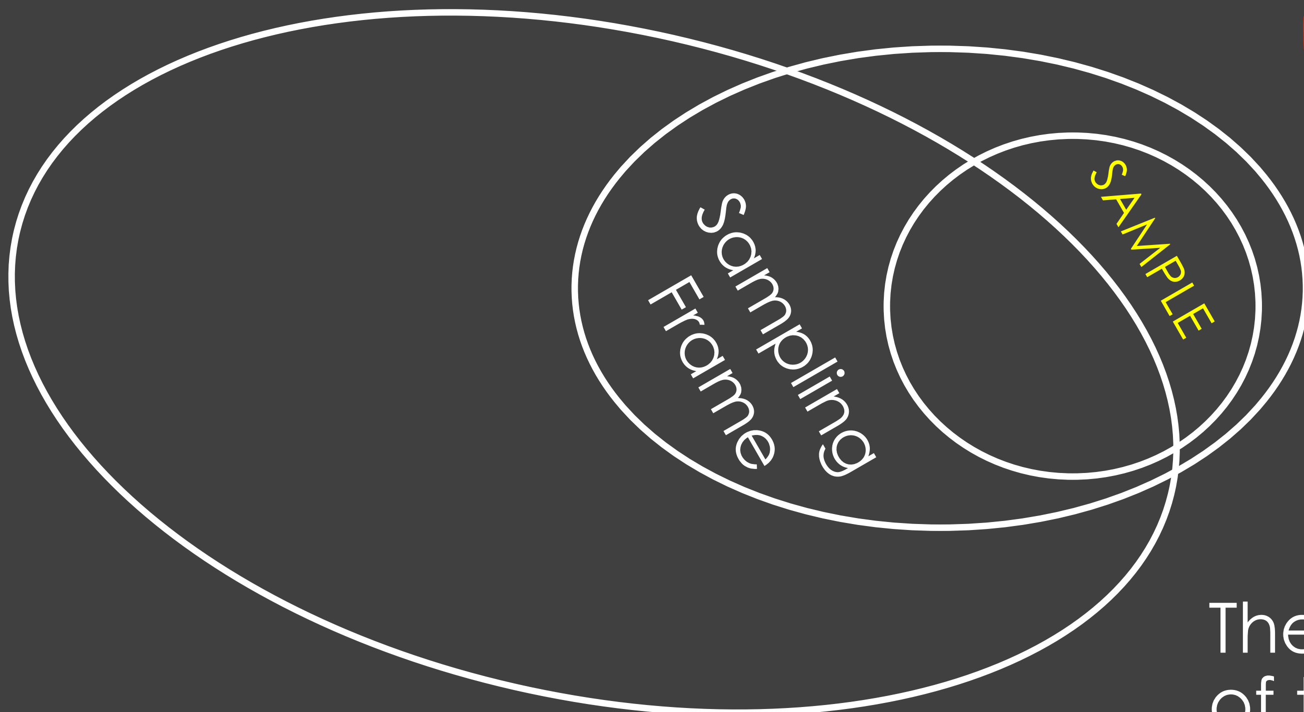


some individuals might not even be in the population — see e.g., the Hite Report Discussion

These individuals are not reachable through the sampling frame

Access the Population through the Frame

Population of Interest



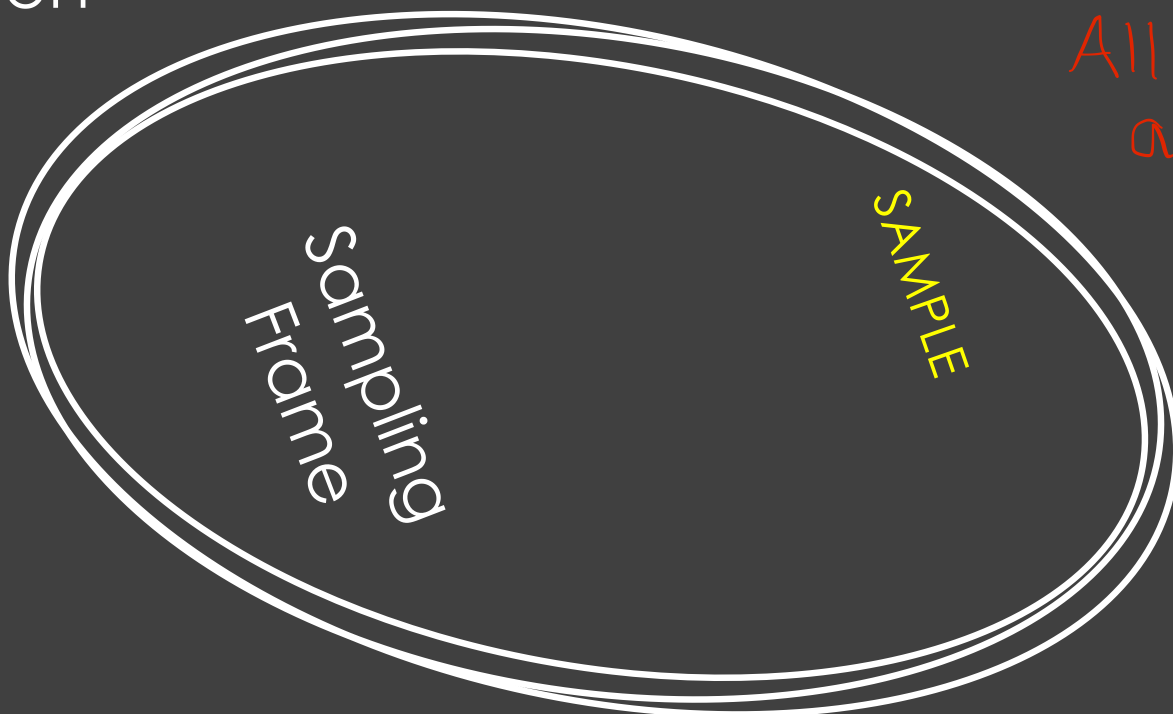
How we select the individuals from the sampling frame matters

The Sample is a subset of the Frame

Sample =
Sampling Frame =
Population

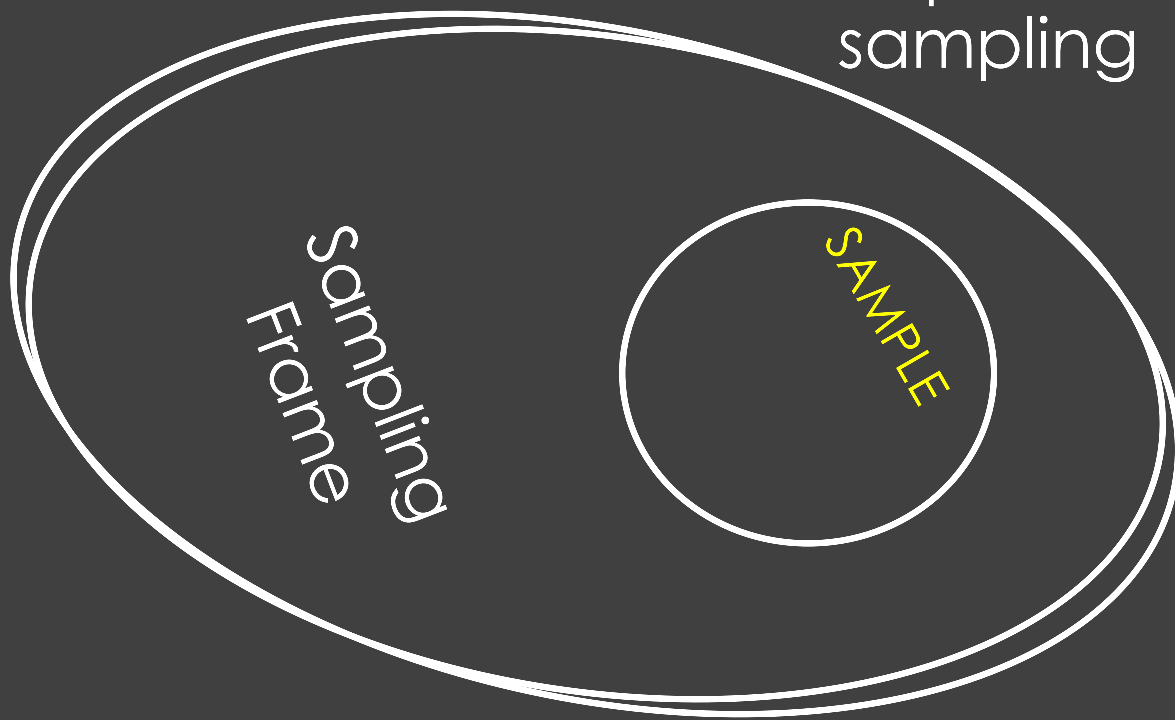
Scenario: Census

All individuals are studied



Sampling Frame = Population

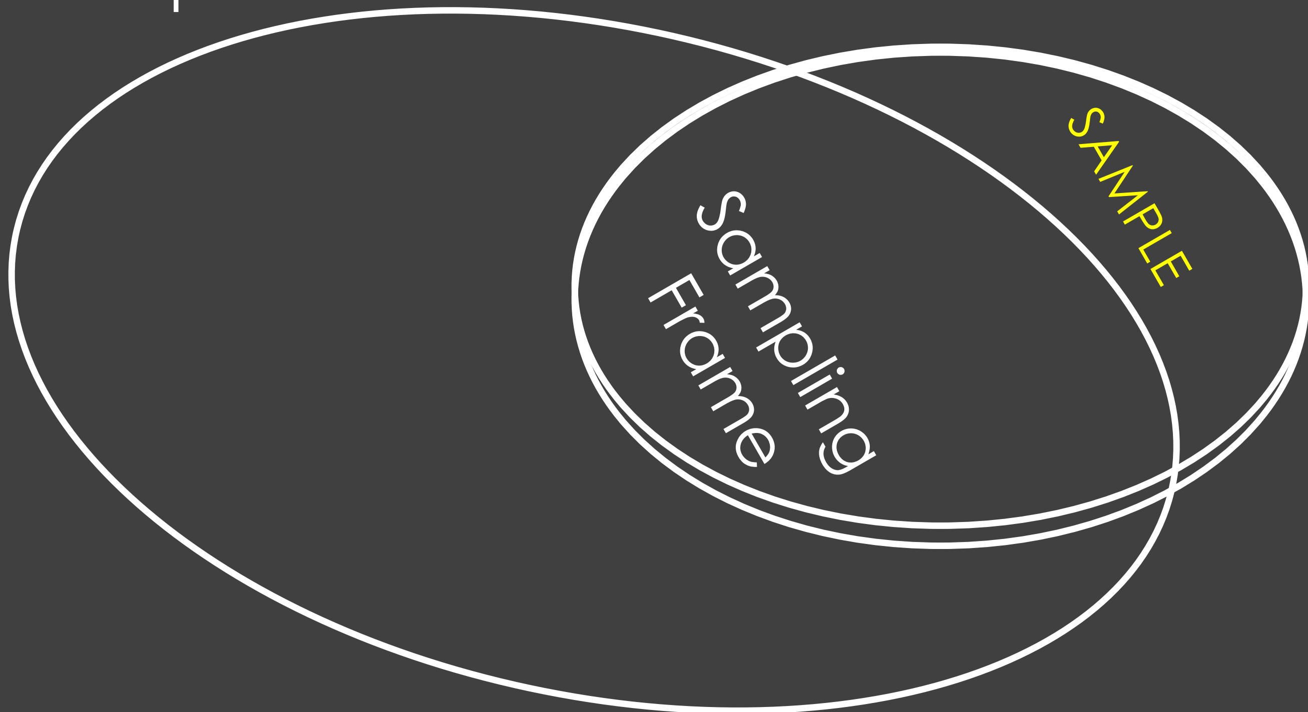
Scenario: Access to all members of the Population when sampling



We often assume away the difference between the frame & population

Sampling Frame = Sample

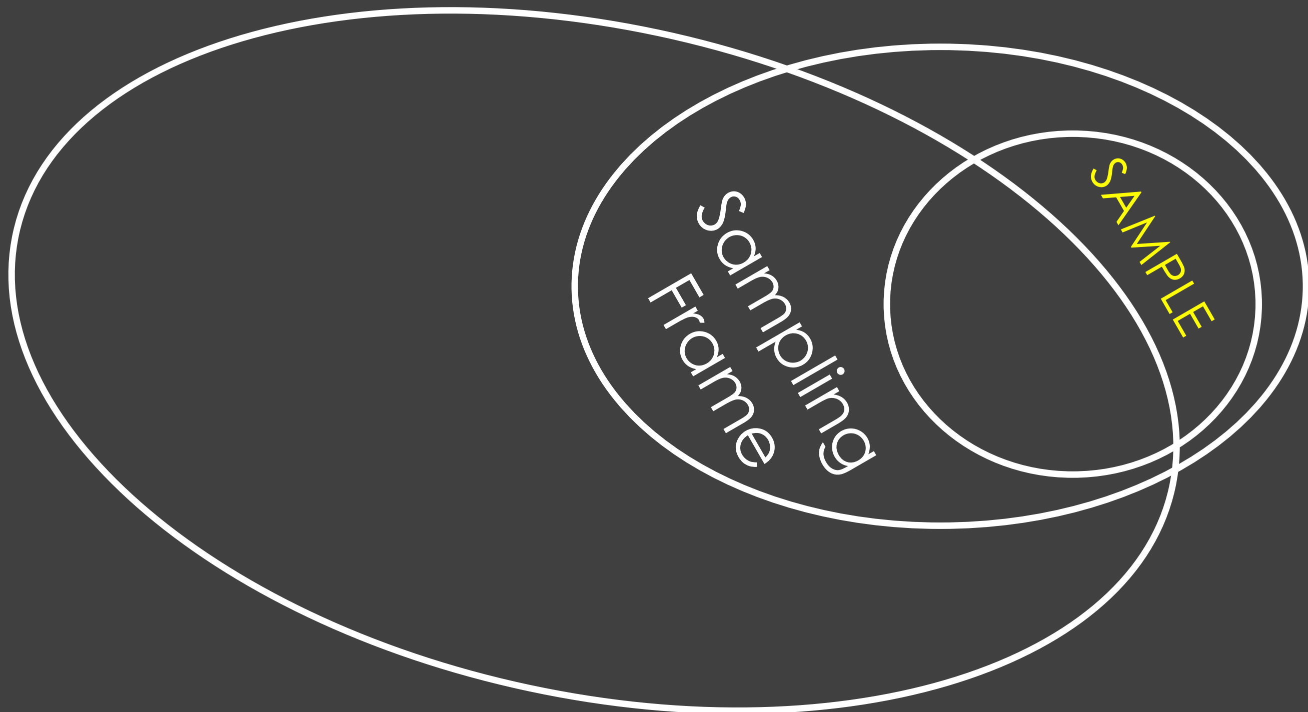
Scenario: Administrative Data



With Admin Data we have access to all in our Frame.

Most Common Scenario

Population of Interest



How are the data generated?

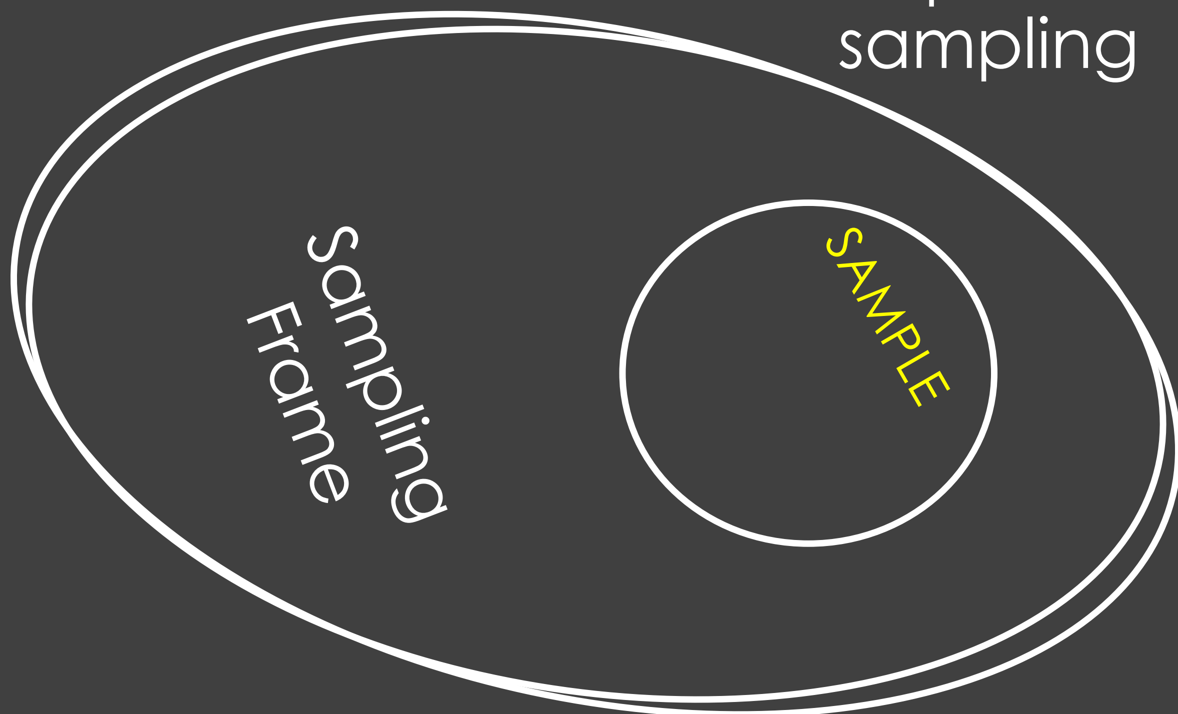
- What is the population of interest?
- What is the sampling frame?
- How are the data generated?

We will turn our focus to this question

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

There are $\binom{N}{n}$ possible samples of size n from N

N choose n

Recall that

$$\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

Convince yourself of this with a simple example →

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…*

$N = 4$ A, B, C, D are the individuals

$n = 2$ Possible samples of size 2

(A,B) (A,C) (A,D)

(B,C) (B,D)

(C,D)

$$\binom{4}{2} = \frac{4!}{2!\,2!} = \frac{4\times 3}{2\times 1} = 6$$

6 samples of size 2

Start Simple

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

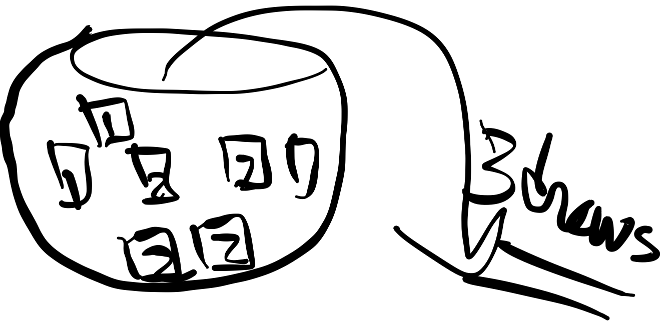
	Number of Children			
	1	2	3	4+
Count	2	4	3	1
Proportion	20%	40%	30%	10%

There are $\binom{10}{3}$ possible samples $\frac{10!}{3!\,7!} = \frac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120$

One way to think about taking the sample:

Write each mother's value on a ticket;
Put the tickets in an urn; Mix; Draw one at a time
Without Replacement

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

Random Variables

We don't know what we will get

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+
$P(X_1 = x)$	20%	40%	30%	10%

$P(X_1 = 1)$ = Chance draw a 1 from the urn

$= \frac{2}{10}$ ← # 1s ← # tickets

	Probability Distribution			
x	1	2	3	4+
$P(X_1 = x)$	2/10	4/10	3/10	1/10

X_1 The number of children for the first mother chosen

What is the expected value of X_1?

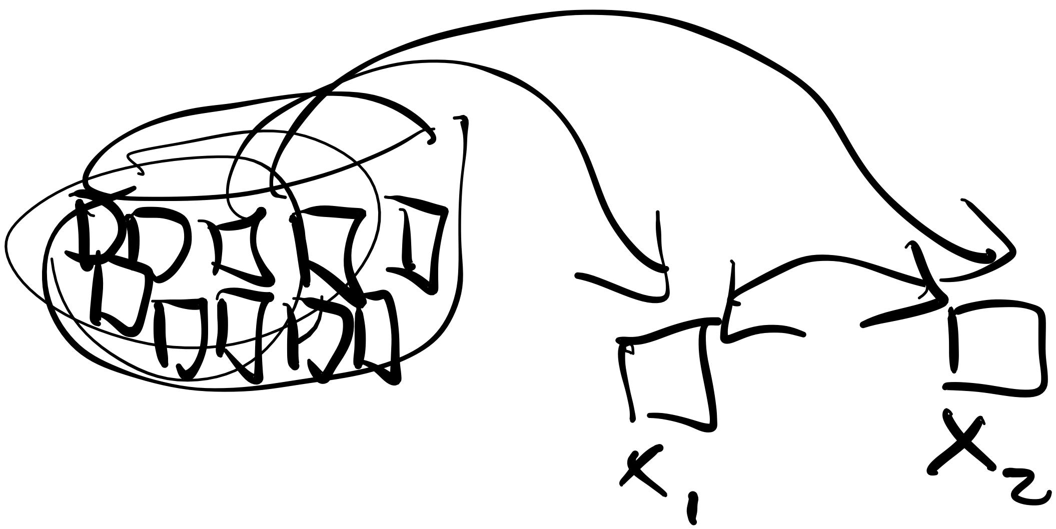
$$\mathbb{E}(X_1) = \sum_{j=1}^{4} x_j P(x_j)$$

$$= 1 \cdot \frac{2}{10} + 2 \times \frac{4}{10} + 3 \times \frac{3}{10} + 4 \times \frac{1}{10}$$

$$= 2.3$$

X_2 number of children for the 2nd mother chosen

	Probability Distribution			
x	1	2	3	4+
$P(X_2 = x)$	2/10	4/10	3/10	1/10



For example, we could draw the two tickets and then swap them

By Symmetry $P(X_1=1)=P(X_2=1)$

X_2 number of children for the 2nd mother chosen

Counting Way

A	B	C	D	E	F	G	H	I	J	moms
1	1	2	2	2	2	3	3	3	4	values

\# pairs with 1 for 2nd mom

(A, B) (B, A) 2

(C, A) (C, B) 2

⋮

(J, A) (J, B) 2

$\Big]\, 8$

$2 + 2 \times 8 = 18$

\# pairs (order matters)

10×9

10 ways to pick 1st mom

9 ways to choose 2nd mom

$$\frac{18}{90} = \frac{2}{10} \; !$$

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

Random Variables

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

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

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

Use $\mathcal{L}_2$ loss again

$(X - c)^2$

In General

$X - c$

is the error

It is a random variable

Now find the Expected Value of the loss

AKA RISK $E(X-c)^2$

Summarizing the Probability Distribution

EXPECTED LOSS:

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

Minimize the risk $$\mathbb{E}(X-\mathbb{E}(X)+\mathbb{E}(X)-c)^2$$

Like before we add and subtract $\mathbb{E}(X)$

Properties of Expected Value

$$\mathbb{E}(X) = \sum_{j=1}^{m} x_j P(X = \overset{x}{\cancel{m}}_j)$$

There are j distinct values X can take on

$$\mathbb{E}(aX + b) = \sum_{j=1}^{m} (a x_j + b) P(X = x_j)$$

$$= a \sum_{j=1}^{m} x_j P(X = x_j) + b \underbrace{\sum_{j=1}^{m} P(X = x_j)}_{\text{sum to } 1}$$

$$= a\,\mathbb{E}(X) + b$$

Each with Probability

$P(x_j) = p_j$ for short

To simplify the writing

let $\mathbb{E}(X) = \mu$

Minimize the Risk

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

$$= \sum_{j=1}^{m} (x_j - \mu + \mu - c)^2 p_j$$

$P(X = x_j)$

$$= \underbrace{\sum_{j=1}^{m} (x_j-\mu)^2 p_j}_{\mathbb{E}(X-\mu)^2} + 2(\mu-c)\underbrace{\sum_{j=1}^{m}(x_j-\mu)p_j}_{0} + \underbrace{\sum_{j=1}^{m}(\mu-c)^2 p_j}_{(\mu-c)^2}$$

$$= \mathbb{E}(X-\mu)^2 + (\mu-c)^2$$

Min for $c = \mu$

The Expected Value Minimizes Risk

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

$$\mathbb{E}(X-\mu)^2 = \sum_{j=1}^{m} (x_j - \mu)^2 P(x_j)$$

This side is the Variance

Data Life Cycle

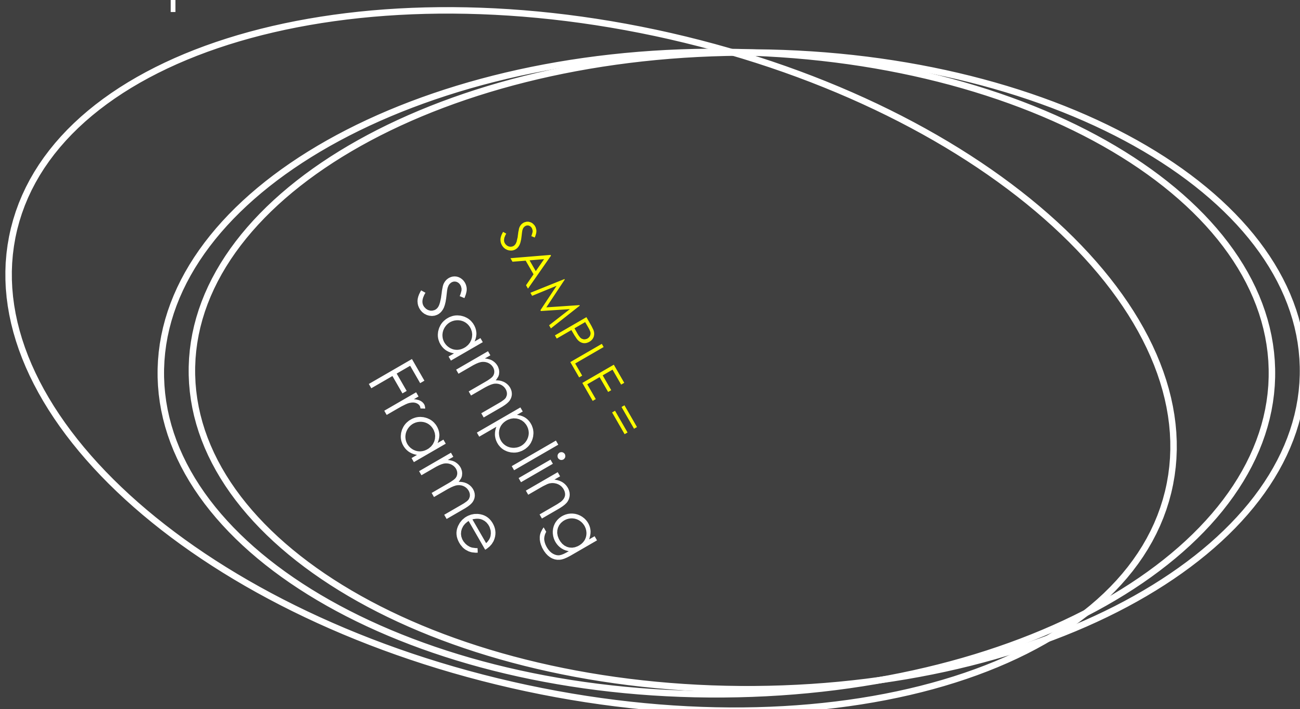
Data
Design/
Generation

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.

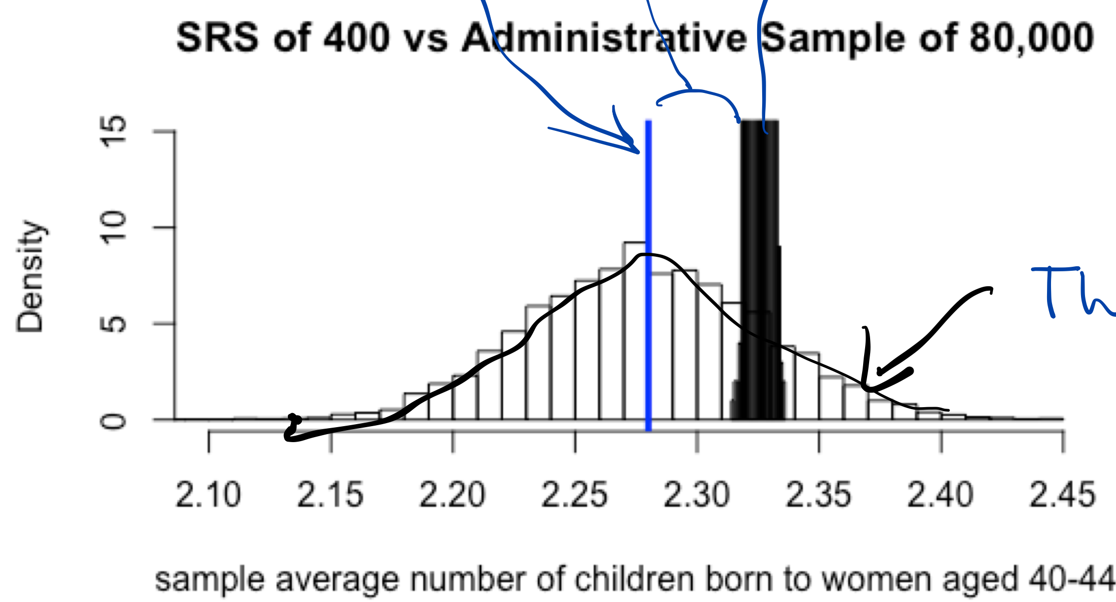
Active Area of Research Area

Large Administrative Data vs Small SRS

Truth

Bias

The admin data set is very accurate but biased



The possible values the sample avg could be for a SRS of n=400

The bias may be small enough to not matter. If it isn't, it's a problem

Data Life Cycle

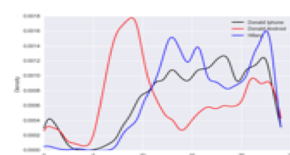
Precise

Question Formulation ?

$\bar{x}$ in our sample

Data Design/ Generation

Probability



Data Analysis

Generalization

$$E(\bar{X}) = E(X)$$

μ in the world

$$= \mu$$