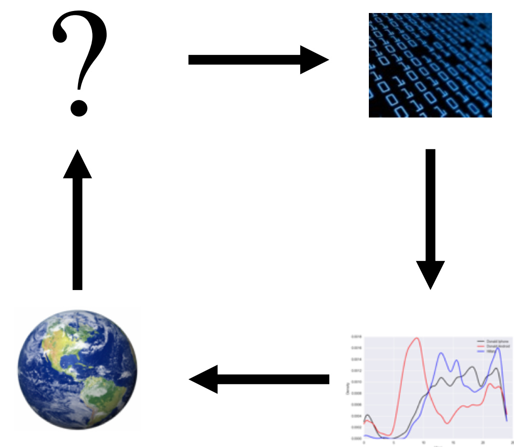
Data Science 100

Principles & Techniques of Data Science



Slides by:

Deborah Nolan

deborah_nolan@berkeley.edu

Announcements for Today

- *The class has been enlarged and the wait list is operating.*
- *If you are a graduate student not on the waitlist, try to get on it ASAP.*
- *Annotated slides are added after class*
- HW 2 will be released tonight and due 11:59 Wednesday Sep 11
- Office hours are found at http://ds100.org/fa19/calendar

Topics for Today

- *How to solve probability problems*
- *Review random variables, probability distribution, expectation and variance*
- *Review Error, Loss, and Risk and the Relationship between the Data and the "World"*
- An Example

How do we solve probability problems?

Basic Approaches

- *Symmetry and Analogy*
- *Counting and equally likely*
- *Trees and conditional probability*

Recall our group of 10 mothers

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

- *Select a mother at random from the 10, record her #kids*
- *Do not replace*
- *Repeat for a total of 3 samples*

Recall our group of 10 mothers

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

- *What is the chance the second mom selected has 1 child?*

Symmetry & Analogy

- Urn with 10 marble one for each mother, indistinguishable except for the # written on it
- Box with 10 indistinguishable tickets, except for the # on it
- Deck of 10 indistinguishable cards, except for the # on the flip side

Symmetry & Analogy

- Draw marbles from well mixed urn
- Select tickets from well mixed box
- Deal cards from top of well shuffled deck
- Deal cards from bottom of well shuffled deck

Symmetry & Analogy

- Chance the second draw is 1

Counting

- 10 people named A, B, C, D, E, F, G, H, I, J
- With values 1, 1, 2, 2, 2, 2, 3, 3, 3, 4
- Number of Combinations of first and second draws
- Number of Combinations where the second draw is 1
- Since each combination is equally likely, we take the ratio

Counting

- Chance the second draw is 1

Tree and Conditioning

Two step process.

Only need to track whether card is 1 or not.

If you know the result of the first draw, compute the conditional chance of the second draw.

Tree and Conditioning

- Chance the second draw is 1

Many approaches to figuring out probabilities

- Get good at one
- But be flexible and try multiple approaches

FUN PROBLEM: There are 3 cards, one has a circle on both sides, one has a square on both sides, and the third has a circle on one side and square on the other. Mix them up and place one card on the table. It displays a circle. What's the chance there is a circle on the reverse side?

Working formally with Random Variables

0-1 Random Variables

- In discussion yesterday, you worked with random variables that take on the 0 or 1 values
- We will start with it as an example

$$X = 0 \text{ with prob } 1 - p$$

$$= 1 \text{ with prob } p$$

Examples?

Probability Distribution

Expected Value and Variance

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

$$\mathbb{V}ar(X) =$$

More Generally, Expected Value and Variance of a Discrete RV

Probability Distribution

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

$$\mathbb{V}ar(X) =$$

More Generally

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

$$\mathbb{V}ar(aX + b) =$$

Sums of 0-1 Random Variables

$$X_i = 0 \text{ with prob } 1-p$$

$$= 1 \text{ with prob } p \qquad \text{for } i = 1, \ldots, n$$

Examples?

Expected Value

$$\mathbb{E}(X_1 + \cdots + X_n) =$$

Variance

If Independent

$$\mathbb{V}ar(X_1 + \cdots + X_n) =$$

If From a Simple Random Sample

$$\mathbb{V}ar(X_1 + \cdots + X_n) =$$

Probability Distribution

Concrete: $n = 4$ and $Y = X_1 + X_2 + X_3 + X_4$ and the Xs are independent with same chance of 0 or 1 (knowing the value of X_1 doesn't change X_2 distribution).

$P(Y = 2) =$

Probability Distribution

n independent 0-1 variables

$Y = X_1 + \ldots + X_n$

$P(X_i = 1) = p$

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 1, \ldots, n$$

Fun Problem Related to HW:

- Roll a fair die 5 times.
- N_E = number of evens,
- N_P = number of primes
- N_1 = number of 1s

$$P(N_1 = 1, N_P = 2, N_E = 2) =$$

Summary Statistics as Estimators of Population Parameters

Data Life Cycle

Data
Design/
Generation

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

Generalization

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

Empirical (Data)

DATA: $x_1, x_2, \ldots, x_n$

The sample that we have to work with

Model (World)

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

Probability distribution from, e.g., a SRS from the population

Empirical (Data)

DATA: $x_1, x_2, \ldots, x_n$

Summary statistic that minimizes the empirical risk

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

Model (World)

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

Probability parameter that minimizes the Risk

$$\mathbb{E} l(X - c)$$

Empirical (Data)

DATA: $x_1, x_2, \ldots, x_n$

Summary statistic that minimizes the empirical risk

For l_2 loss, $\bar{x}$ minimizes the average loss

Model (World)

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

Probability parameter that minimizes the Risk

For l_2 loss, $\mathbb{E}(X)$ minimizes the average loss

Empirical (Data) **Model (World)**

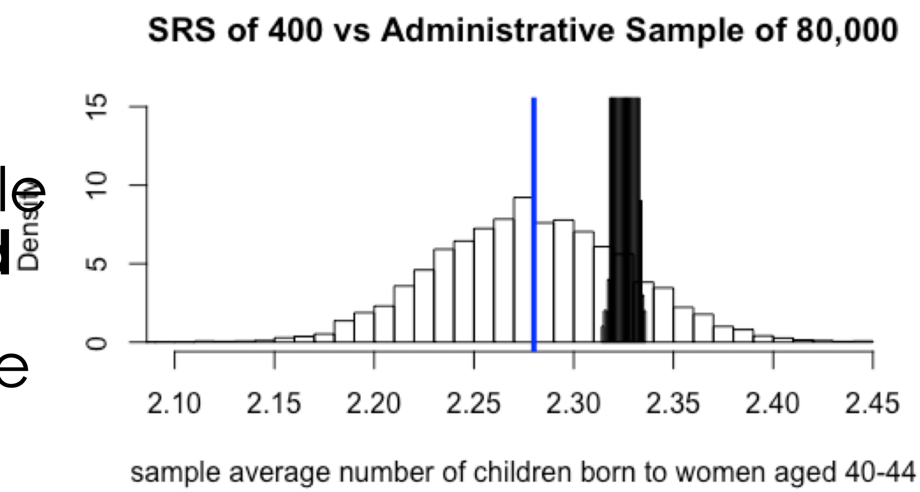
Connect the sample average and expected value:

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

$\bar{X}$ is a random variable

The expected value of a sample average from a SRS is **unbiased**

Its variability is quantifiable – the **sampling error**



Data Life Cycle

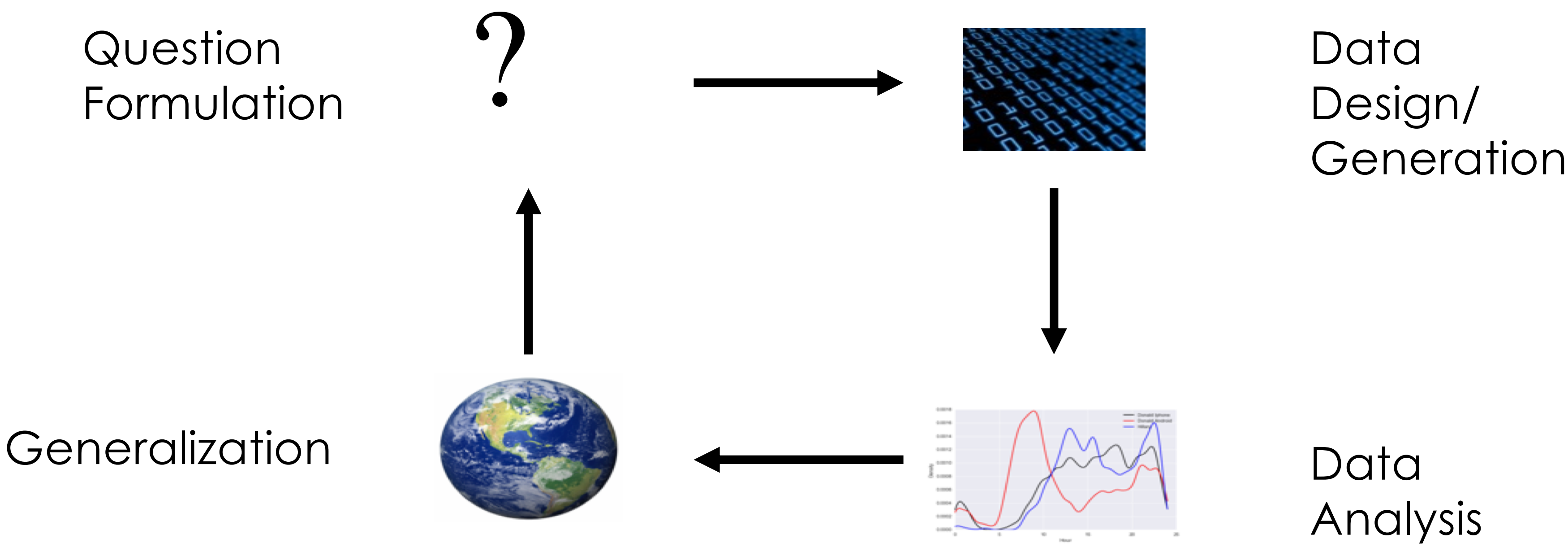
Data
Design/
Generation

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

Generalization

An Example: Wait Time for a Repair

Data Life Cycle



Question

What is the typical wait time for a PG&E repair?

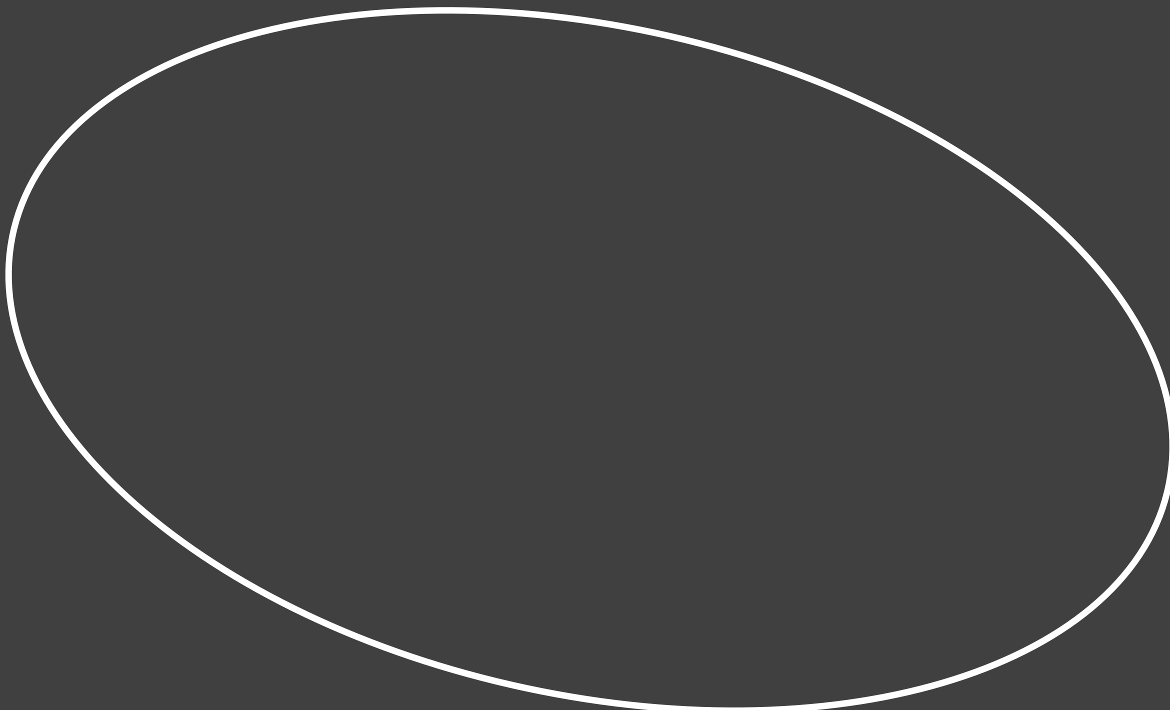
Context

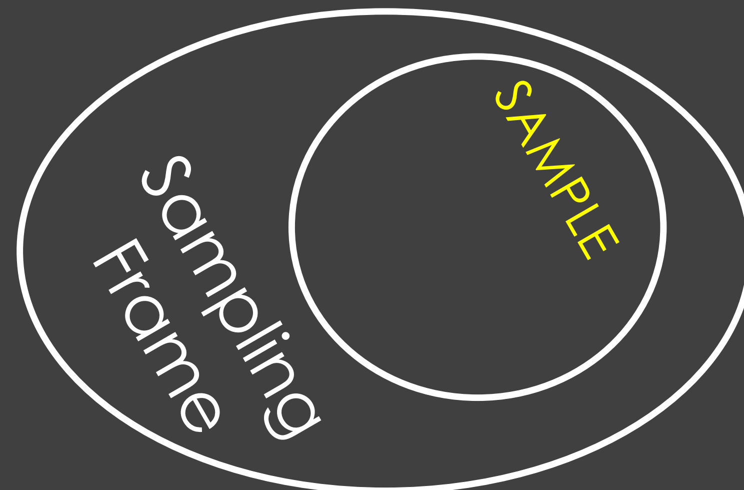
PG&E must report to a utilities commission about its service record.

How might we/they focus this question?

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

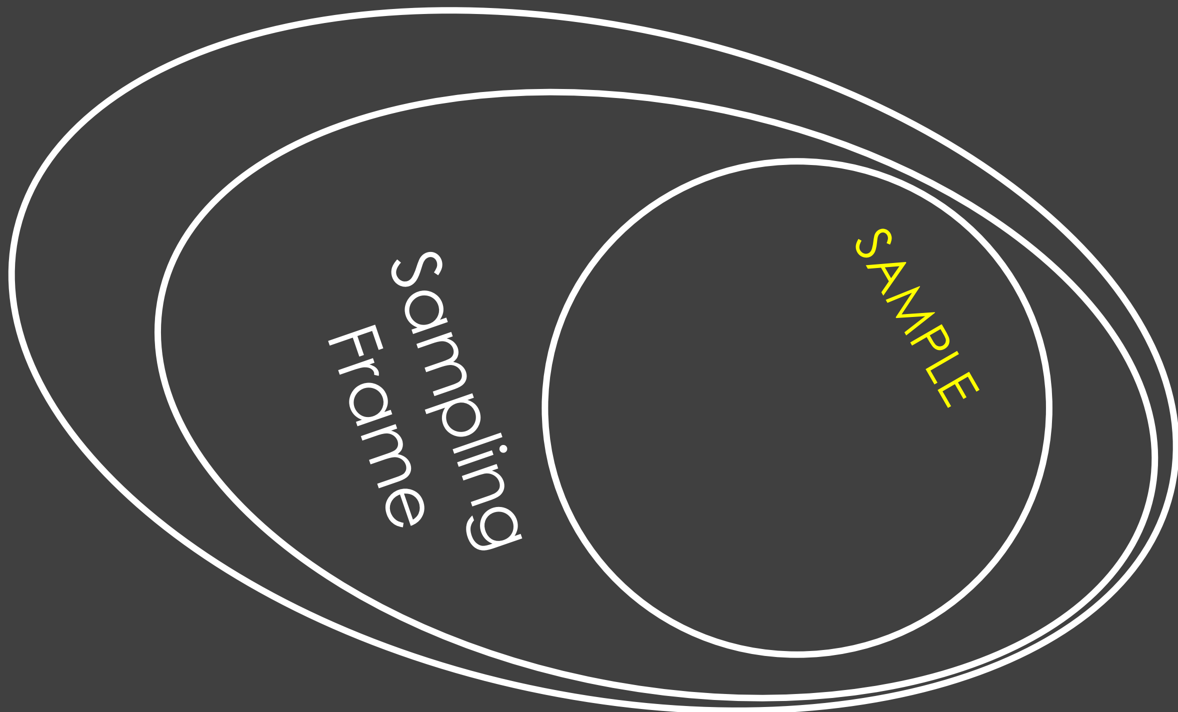
What is the Population of Interest?





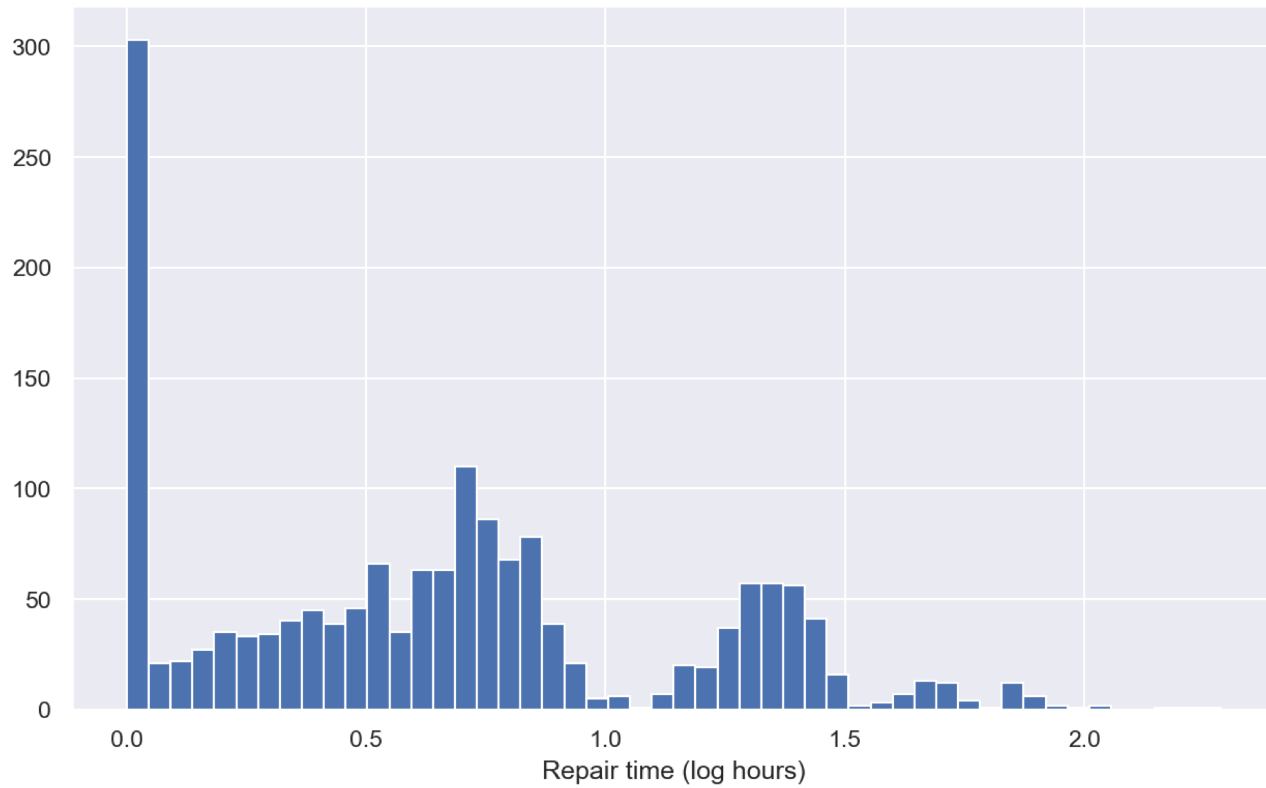
What is the Sampling Frame?

Scenario:
Administrative Data



The Data

$x_1, x_2, \ldots, x_n$ every wait time over a 3 month period



Can we provide a summary statistic?

Why is the sample median such a desirable summary?

Summarizing the Data

DATA: $x_1, x_2, \ldots, x_n$ where n is 1665 for our data

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

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

Minimize the Average L_1 Loss

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

Minimize the Average Absolute Error

$$\frac{1}{n}\sum_{i=1}^{n}|x_i - c|$$

Data Life Cycle

Data
Design/
Generation

Generalization

Probability Samples give us Representative Data where the sample median is a good estimate of the population

Where does Probability Sampling Come into this Problem?

Probabilistic Behavior of the Median

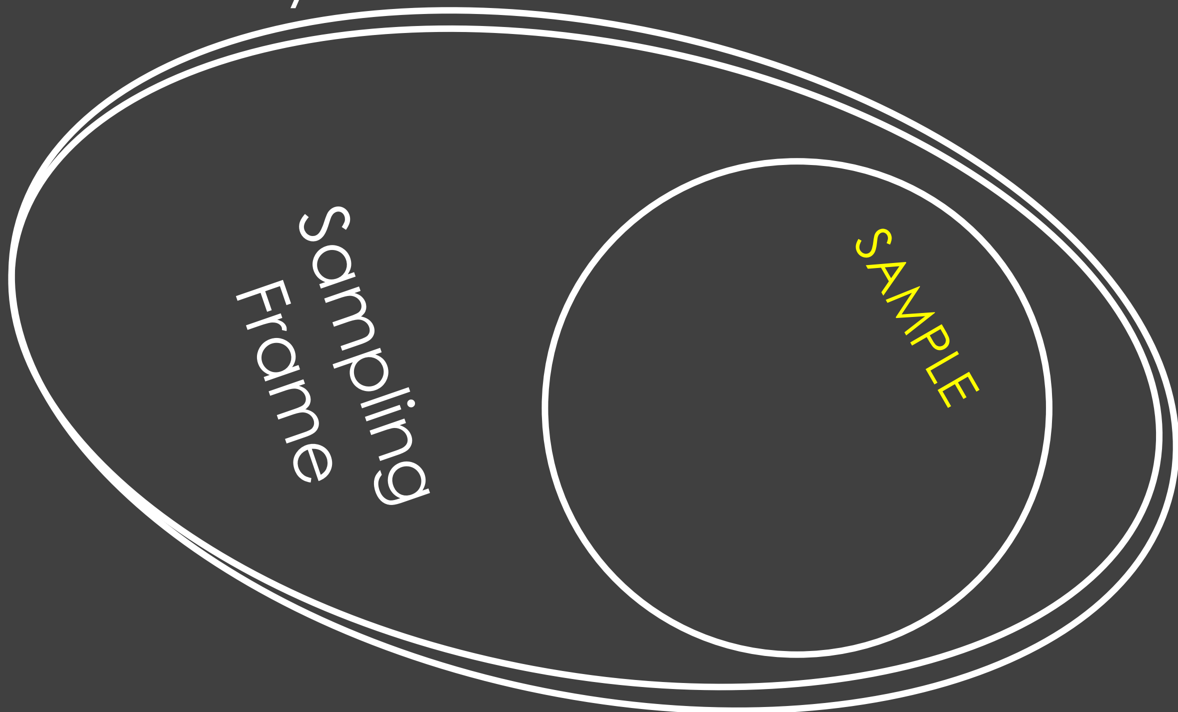
- Not as simple to work with as the mean
- We need to make more assumptions about the underlying probability distribution of X
- In many circumstances the sample median is well-behave and close to the median(X)

HW 2 Introduction

2016 Presidential Election

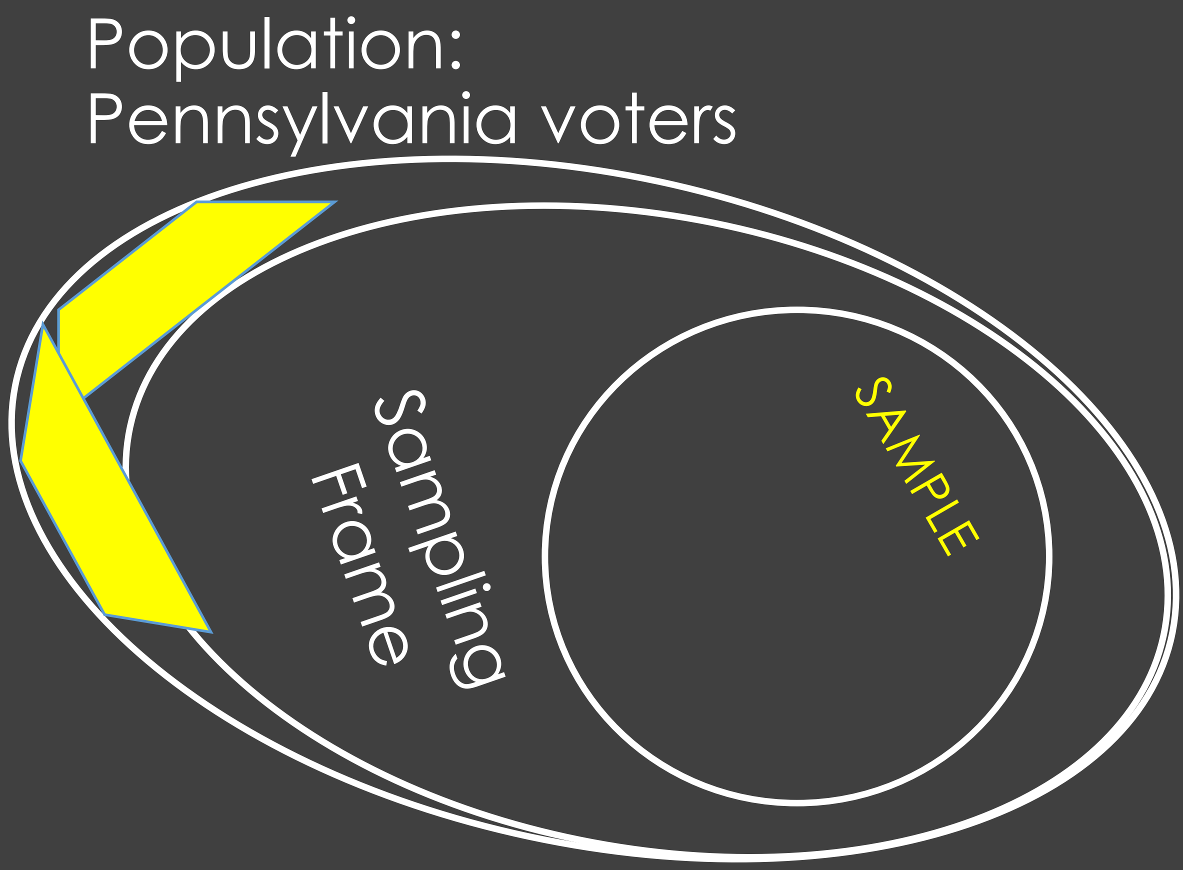
- Outcome took many by surprise
- Most polls were predicting Clinton victory was 90%
- FiveThirtyEight said 70% and a couple of days before indicated that Trump had a chance to win
- Now that the election has passed, we have the opportunity to see the world (voters who voted in the election)

Population:
Pennsylvania voters



We have a record on the
Trump votes
Clinton votes
Other votes

We can simulate the polls to see the sampling distribution of:
(# T votes - # C votes) / Total Votes Sampled



We can introduce a little bias

Simulate the polls to see the sampling distribution of the biased sampling frame:
(# T votes - # C votes) / Total Votes Sampled