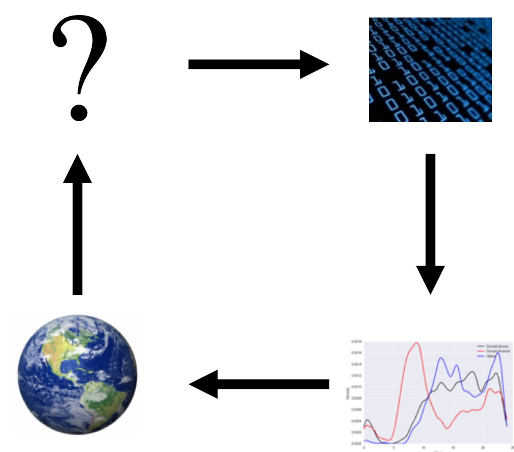
Data Science 100

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

These 4 scenarios are all equivalent to choosing mothers to participate in a survey

Symmetry & Analogy

- Chance the second draw is 1

$$\frac{2}{10}$$

1st draw & 2nd draw have same chance of 1 or Ace

Counting

Name each mother so we can track the combinations

- 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 of these two counts to get the probability

Counting

➢ Chance the second draw is 1

A B C D . . .
1 1 2 2 . . .

comb of 2 moms ?
order matters.

$10 \times 9 = 90$

A,B B,A 2

comb w/ 2nd mom
has 1 child

C,A C,B
⋮
J,A J,B

8×2

18 total

$18/90 = 2/10$

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

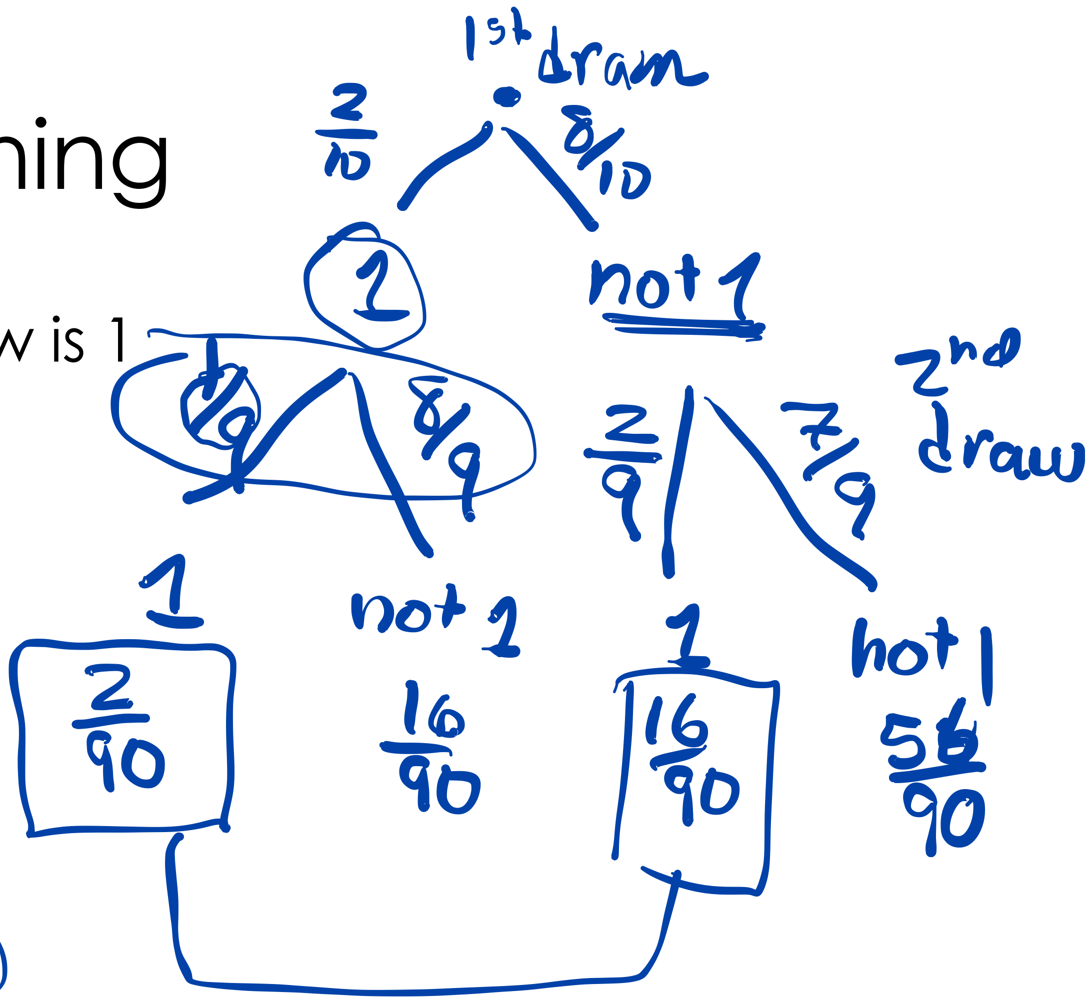
- Chance the second draw is 1

Probabilities go on branches

Probabilities from the same node sum to 1

Probabilities are conditional on the information above (the path on the tree that was traveled to get there)

Multiply down the tree to get the chance of the final result



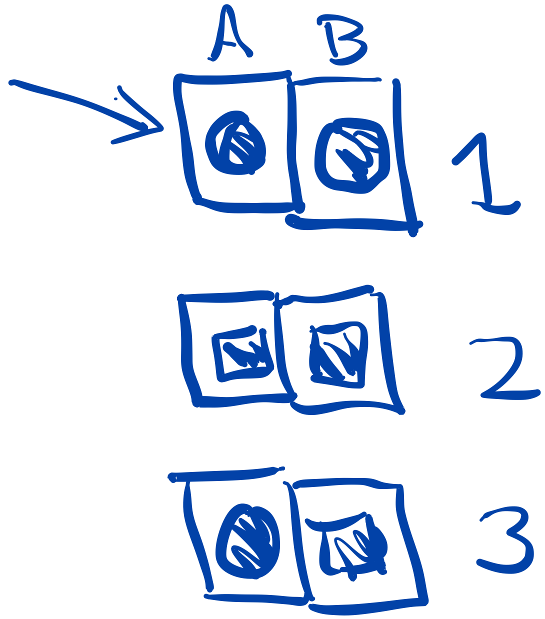
$$\frac{2}{90} + \frac{16}{90} = \frac{18}{90}$$

Add across the bottom

Many approaches to figuring out probabilities

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

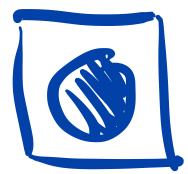
FUN PROBLEM: There are 3 cards, one has a circle on both sides, one has a square on both sides, and the third has a circle on one side and square on the other. Mix them up and place one card on the table. It displays a circle. What's the chance there is a circle on the reverse side?



Mix Up

Pick 1

Put on Table

We See 

What's the chance circle on the other side?

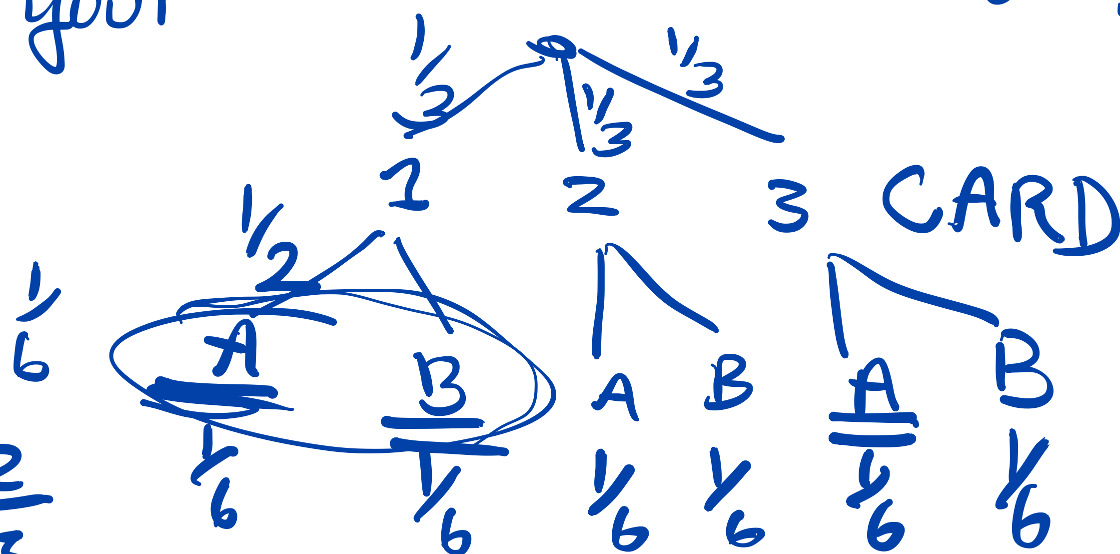
1A or 1B or 3A

$\frac{2}{3}$

Go To

www.yellkey.com/easy

To Register your Answer



$$\frac{1/6+1/6}{1/6+1/6+1/6} = \frac{2}{3}$$

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

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

Chance Process

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

w/ Random Unknown value

Examples?

Medical Trials

Games of Chance

Survey Results

Random occurrences in Nature

Probability Distribution

x	0	1
$P(x)$	$1-p$	p

$$E(X) = 0 \times (1-p) + 1 \times p = p$$

$$Var(X) = E(X-p)^2 = (0-p)^2(1-p) + (1-p)^2 p$$

$$= p^2(1-p) + p(1-p)^2 = p(1-p)\underbrace{\cancel{[p+1-p]}}_{1}$$

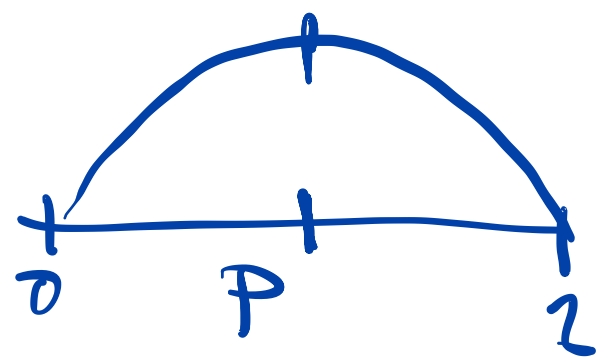
$$= p(1-p)$$

Expected Value and Variance

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

For what value of p is Var(X) largest

$$\mathbb{V}ar(X) = p(1-p)$$



More Generally, Expected Value and Variance of a Discrete RV

Probability Distribution

X	x_1,	x_2	$\cdots$	x_m
$P(X=x)$	p_1	p_2	$\cdots$	p_m

$P(X=x_m) = p_m$

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

$$\mathbb{V}ar(X) = \mathbb{E}(X-\mu)^2 = \sum_{j=1}^{m} (x_j-\mu)^2 p_j = \sigma^2$$

$$\sum_{j=1}^{m}(ax_j+b)p_j = a\sum_{j=1}^{m}x_jp_j + b\sum_{j=1}^{m}p_j$$

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

More Generally

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

$$\mathbb{V}ar(aX+b) = |a|^2\mathbb{V}ar(X)$$

$$\mathbb{E}[aX+b-(a\mathbb{E}(X)+b)]^2$$

$$= \mathbb{E}[aX-a\mathbb{E}(X)]^2$$

$$= \sum_j (ax_j - a\mathbb{E}(X))^2 p_j$$

$$= a^2\sum_j (x_j-\mathbb{E}(X))^2 p_j$$

$$= a^2\,\mathbb{V}ar(X)$$

Sums of 0-1 Random Variables

draw w/o rep

? p

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

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

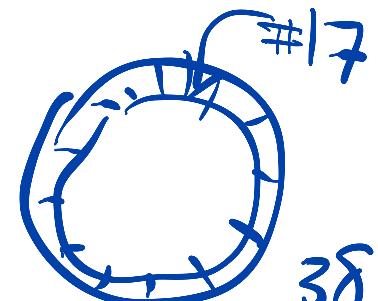
same p for each X_i

# of trials

Examples?

SRS moms 1 child or not
dependent 0 more thn 1

Spin roulette wheel 1 lands 17
independen 0 other #



38 slots

$p = \frac{1}{38}$

Expected Value

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

$$= np$$

for both
ind
& dep X_is

Variance

If Independent

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

$$= np(1-p)$$

If From a Simple Random Sample

$$\mathbb{V}ar(X_1 + \cdots + X_n) = \boxed{}\, np(1-p)$$

N = pop size

6000000

$\frac{N-n}{N-1}$

finite pop correction factor

With large populations there is little difference between SRS & independent draws

Probability Distribution

Roulette wheel

$p = \frac{1}{38}$

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

$X_1 = 1$ if 17

0 other

P(Y = 2) =

1100
1010
1001
0110
0101
0011

6 ways

ways to arrange 4 unique things

$\frac{4!}{2!\,2!} = \binom{4}{2}$

$\boxed{p^2(1-p)^2}$

Chance any one of the six scen

Recall

$4! = 4 \times 3 \times 2 \times 1$

$0! = 1$

$P(Y=2)$ → $\binom{4}{2} p^2(1-p)^2$

Probability Distribution

n independent 0-1 variables

$Y = X_1 + \ldots + X_n$

$P(X_i = 1) = p$

arrangements

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

prob mass fcn

#1

#0s

Binomial Dist(n, p)

Possible outcome from a toss

1 2 3 4 5 6

$\underbrace{1}_{1} \quad \underbrace{2, 4, 6}_{E} \quad \underbrace{3, 5}_{T}$

Fun Problem Related to HW:

- Roll a fair die 5 times.
- N_E = number of evens,
- $N_{\cancel{P}T}$ = number of ~~primes~~ 3 or 5
- N_1 = number of 1s

$P(1) = \frac{1}{6}$

$P(\text{Even}) = \frac{3}{6}$

$P(3 \text{ or } 5) = \frac{2}{6}$

$$P(N_1 = 1, N_{\cancel{P}T} = 2, N_E = 2) =$$

1 E E T T

E 1 E T T

⋮

Chance $\frac{1}{6} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{3} \times \frac{1}{3}$

$= \left(\frac{1}{6}\right)^1 \left(\frac{1}{2}\right)^2 \left(\frac{1}{3}\right)^2$ ·

Count # arrangements

$\Longrightarrow$

$$\frac{5!}{1!\,2!\,2!}$$ # arrangements of 5 unique things

The Es are interchangeable

The Ts are interchangeable

$$\mathbb{P}(N_1=1, N_T=2, N_E=2) = \frac{5!}{1!\,2!\,2!}\left(\frac{1}{6}\right)^1\left(\frac{1}{3}\right)^2\left(\frac{1}{2}\right)^2$$

This is a trinomial distribution

The parameters are: $n=5$ (p_1, p_2, p_3)

$$p_1 = \frac{1}{6}, \quad p_2 = \frac{1}{3}, \quad p_3 = \frac{1}{2}$$

Skip Ahead to Slide 53

The next 7 slides are a review of

Error, Loss, & Risk

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: $x_1, x_2, \dots, x_n$

The sample that we have to work with

Model (World)

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

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)

Connect the sample average and expected value:

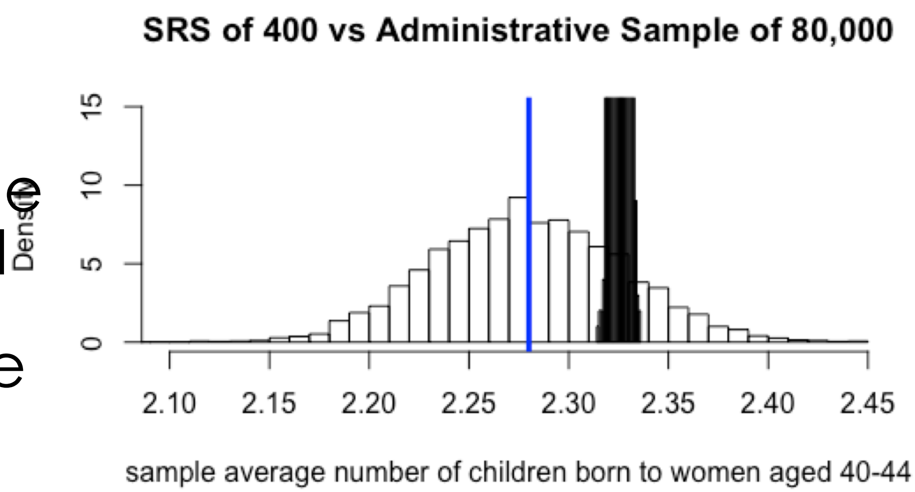
$$\mathbb{E}(\bar{X}) = \mathbb{E}(X)$$

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

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

Model (World)

$\bar{X}$ is a random variable



Data Life Cycle

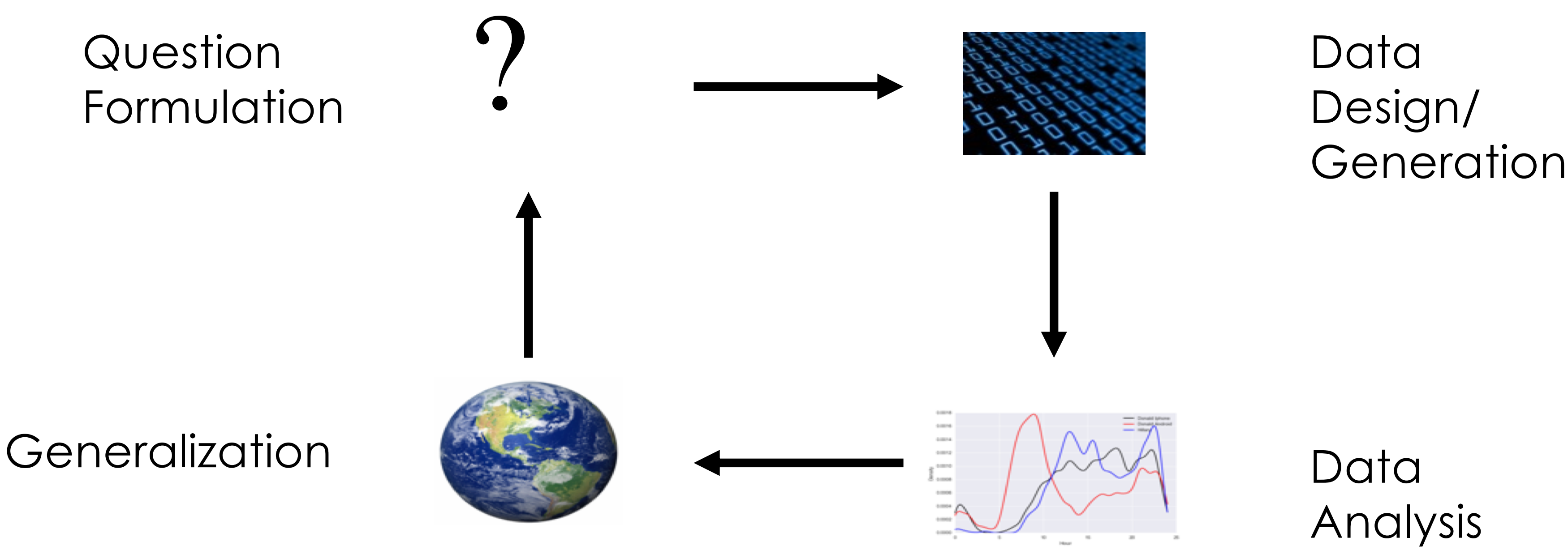
Data
Design/
Generation

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.

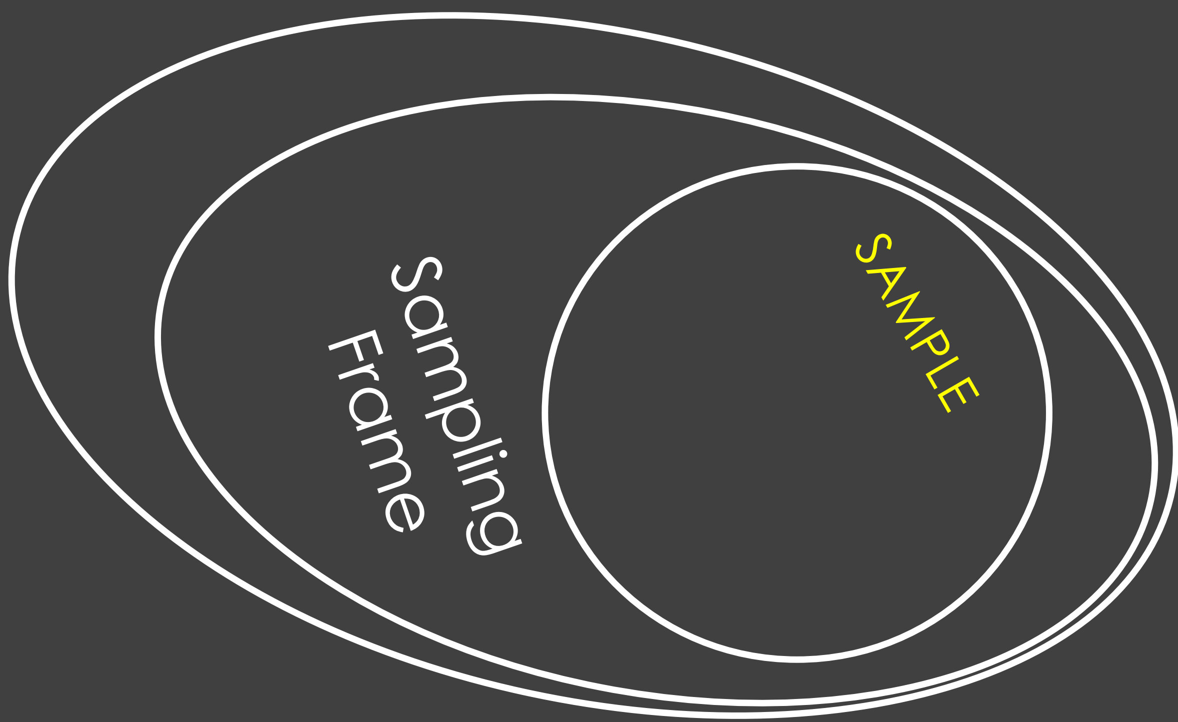
How might we/they focus this question?

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

What is the Population of Interest?

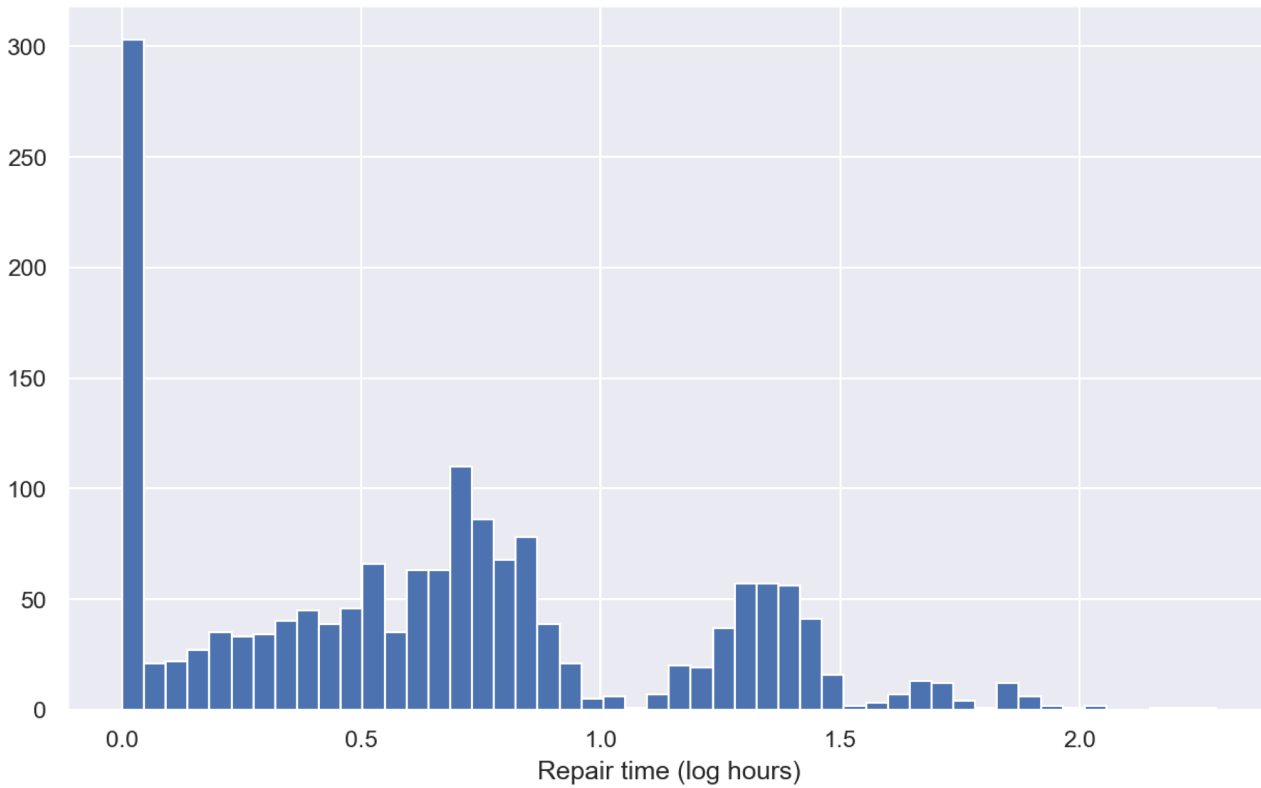
What is the Sampling Frame?

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

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

Generalization

Where does Probability Sampling Come into this Problem?

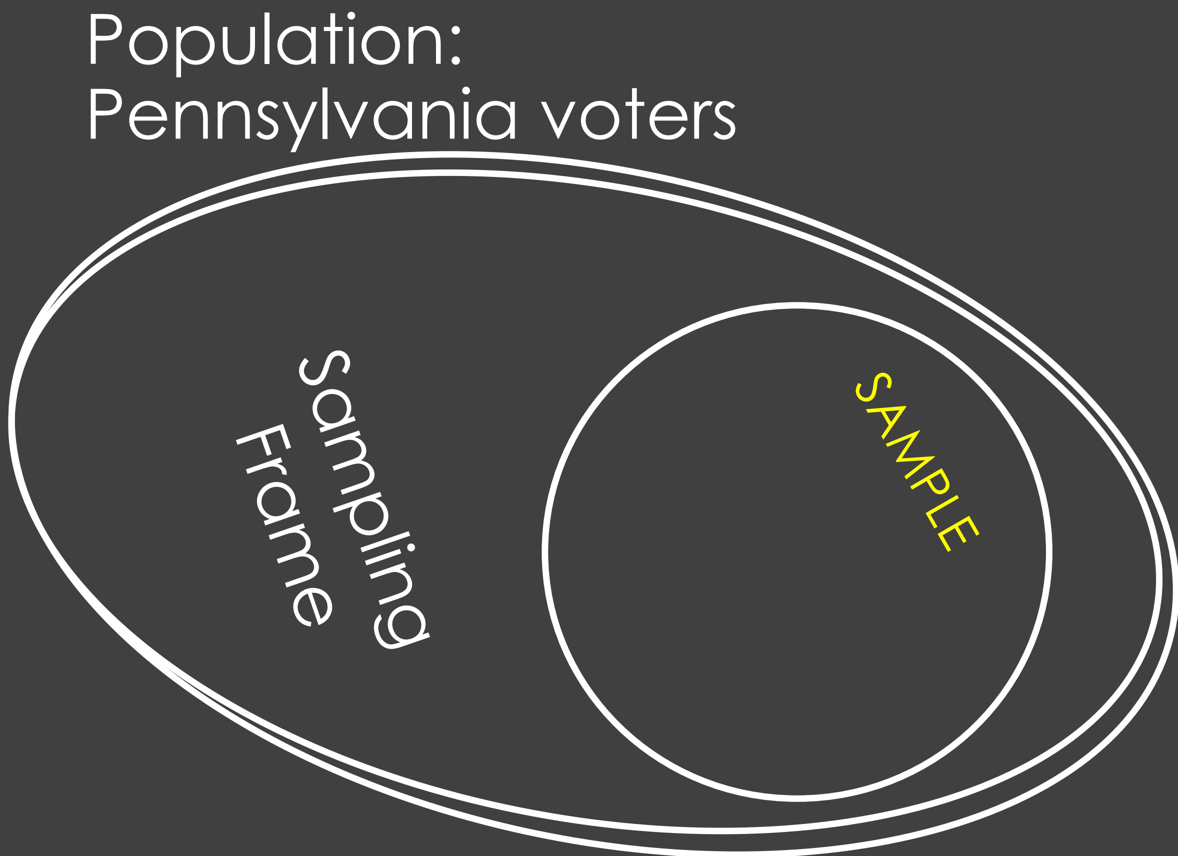
Probabilistic Behavior of the Median

- 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

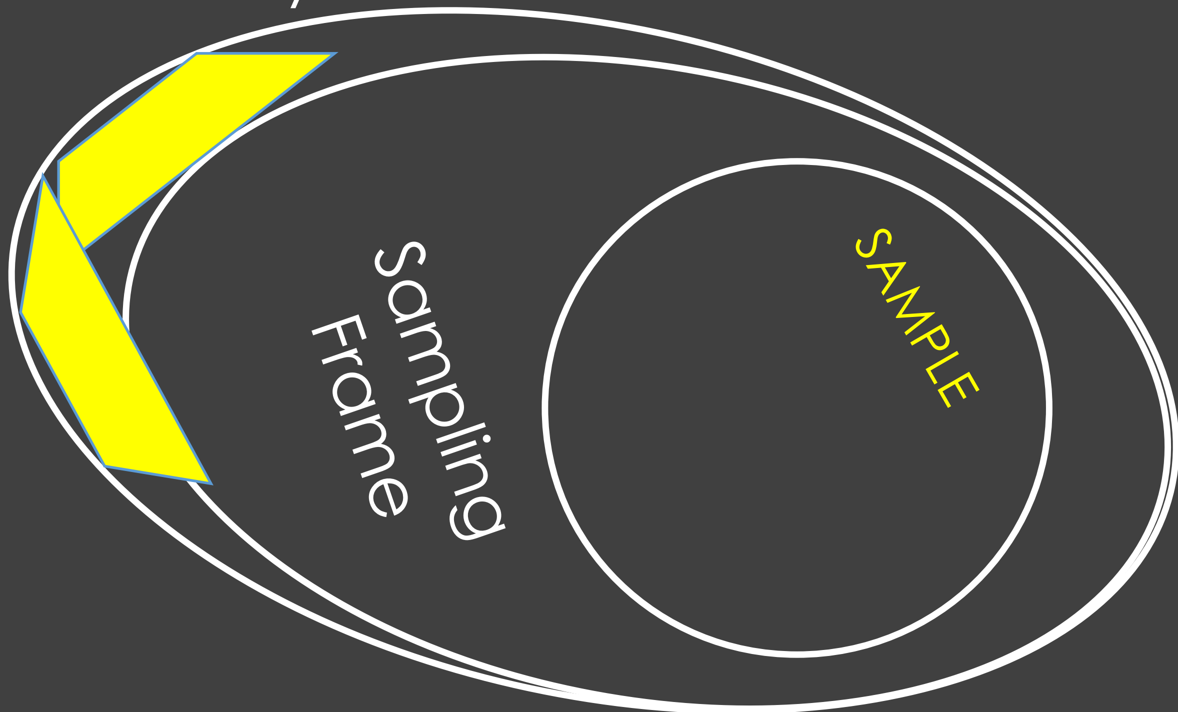
- Outcome took many by surprise
- Most polls were predicting Clinton victory was 90%
- FiveThirtyEight said 70% and a couple of days before indicated that Trump had a chance to win
- Now that the election has passed, we have the opportunity to see the world (voters who voted in the election)



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

Population:
Pennsylvania voters



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