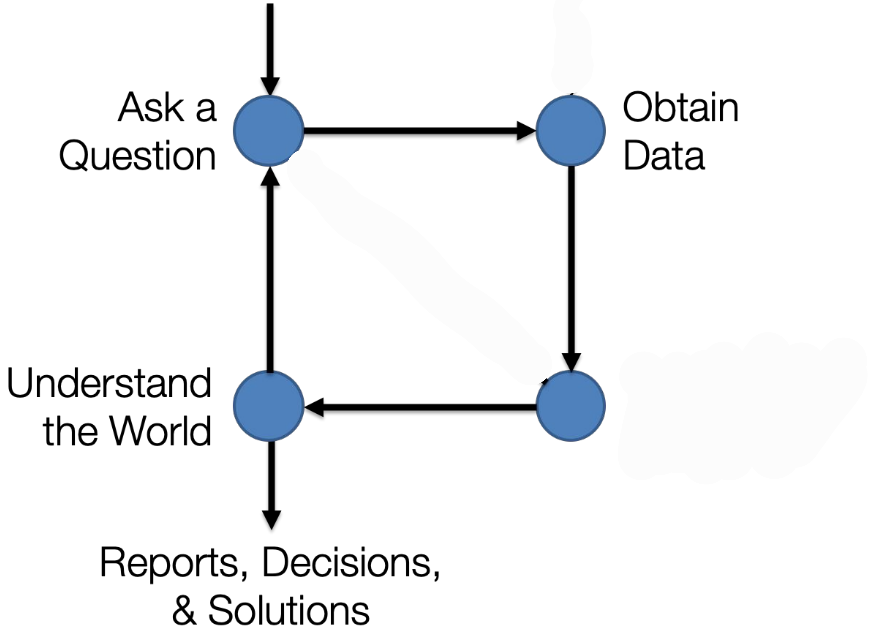
LECTURE 3

Estimation and Bias

Random variables, expected value, parameters, statistics and bias.

Data 100/Data 200, Spring 2021 @ UC Berkeley
Andrew Bray and Joseph Gonzalez
(content by Anthony D. Joseph, Suraj Rampure, Ani Adhikari)

Understanding the world through data



Lectures 2 - 3

- Simple data
  - One variable
  - One unit of observation
  - Few values

Lectures 4 - 10

- Complex data
  - Many variables
  - Many units of observation
  - Messy!

Where we're headed today

What is **statistical bias**?

The difference between your estimate and the truth.

Recap: Data Sampling and Probability

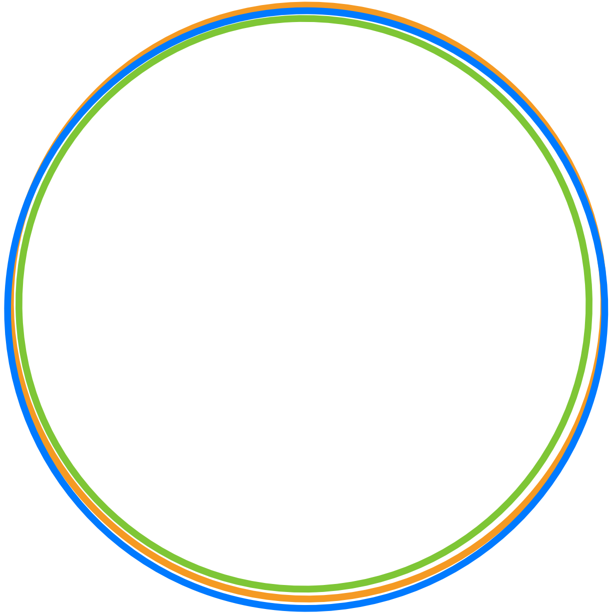
Key Concepts in Sampling

Population: the set of all units of interest, size N.

Sampling frame: the set of all possible units that can be drawn into the sample

Sample: a subset of the sampling frame, size n.

Scenario 1: A census

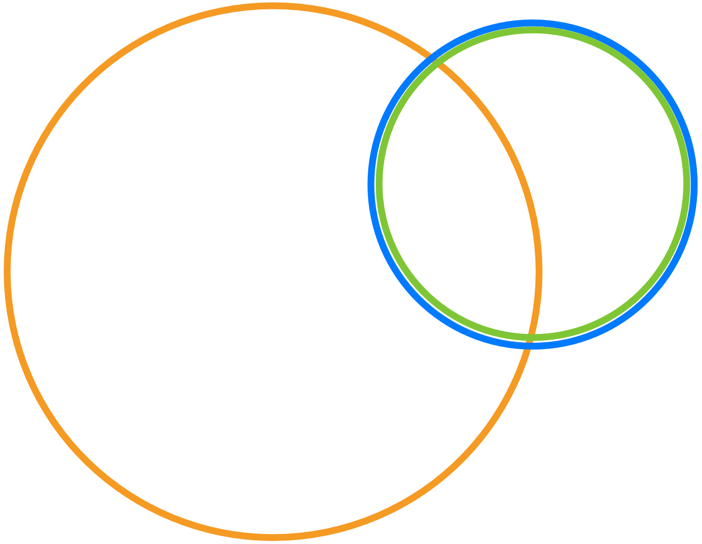


Population
Sampling frame
Sample

Key Features

- population ↔ sampling frame ↔ sample
- Pros: Lots of data
  - No selection bias
  - Easy inference
- Cons:
  - Expensive (time, money)
  - Often impossible

Scenario 2: Administrative Data



Key Features

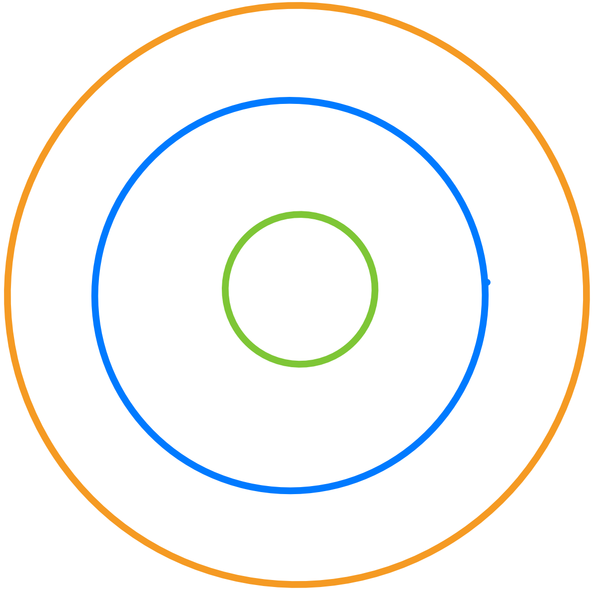
- Sampling frame contains contains a lot not in population.
- Have access to entire frame.

Population

Sampling frame

Sample

Scenario 3: What we like to think we have

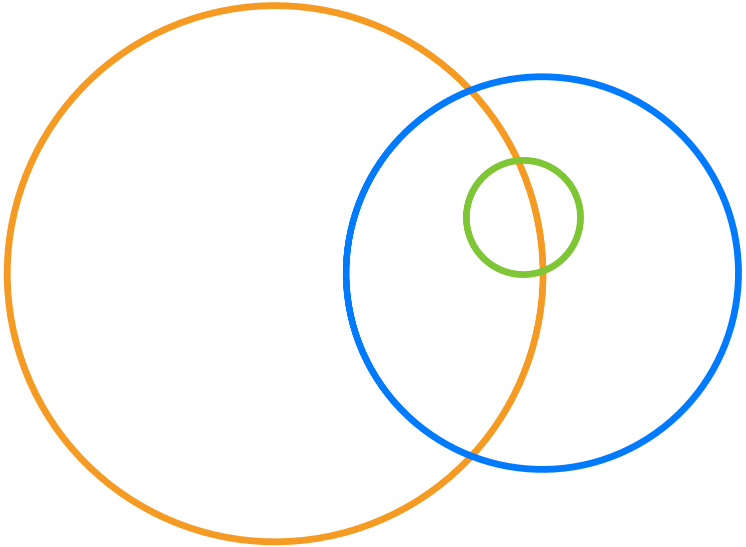


Key Feature

- Optimistic sense that sample is representative of population.

Population
Sampling frame
Sample

Scenario 4: What we usually have

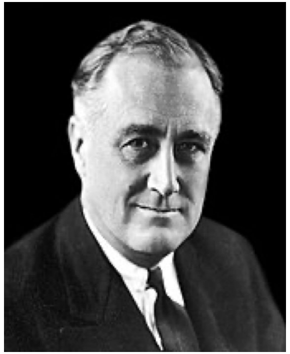


Key Feature

- Sample may be drawn from a skewed frame and may not be representative of population.

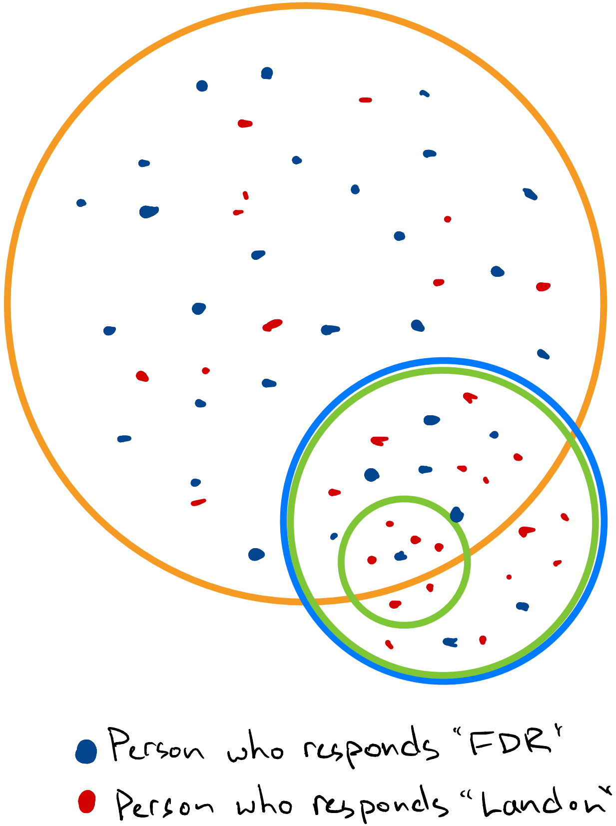
Population
Sampling frame
Sample

Case study – 1936 Presidential Election



Roosevelt (D)

Landon (R)



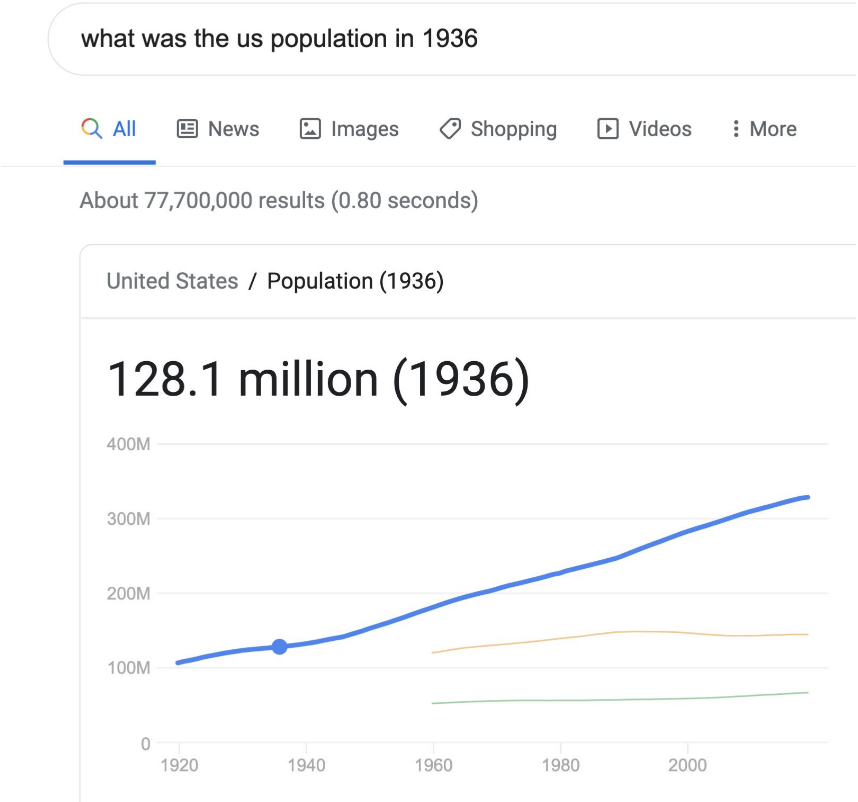
Q: What was the population?

A: Population All people who will cast votes in the 1936 Presidential election.

Selection bias: systematically favoring (or excluding) certain groups for inclusion in the sample.

Non-response bias: when people who don't respond are non-representative of the population.

Data Quality vs. Data Quantity.



Literary Digest 1936 Poll: $n = 10$ million

US population 1936: $N = 128$ million.

$\rightarrow$ 8%!

Gallup 1936 Poll: $n = 50{,}000$

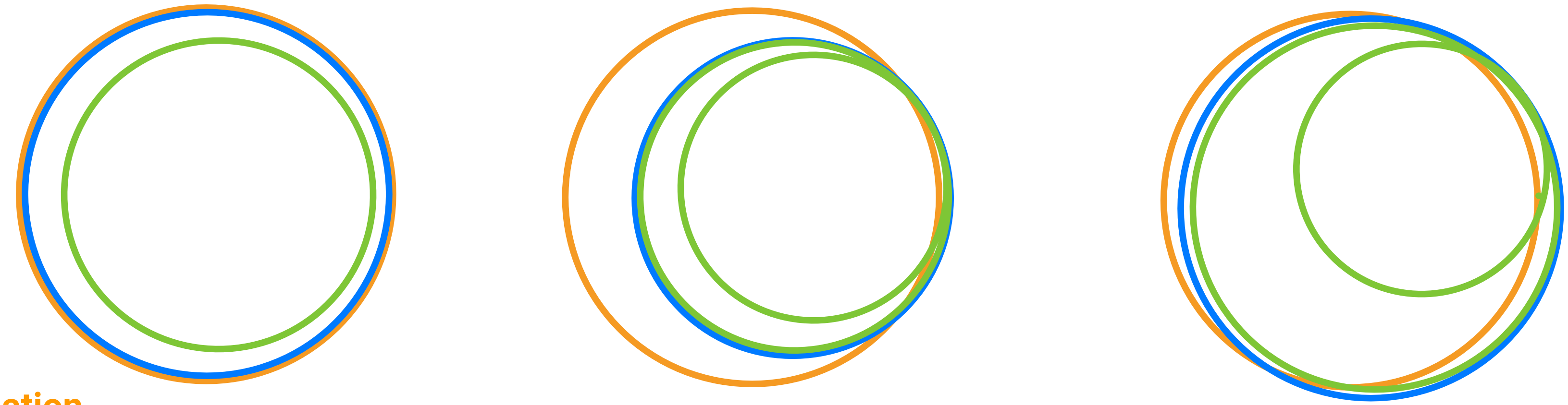
Gallup 2021 : $n = 1000$

into the sample. The typical sample size for a Gallup poll, either a traditional stand-alone poll or one night's interviewing from Gallup's Daily tracking, is 1,000

Case Study - Gender diversity in Data Science

Question: What proportion of Data 100 students identify as female?

Try 1: Babynames → 43%   **Try 2**: Zoom poll → 49%   **Try 3**: Pre-class survey → 48%

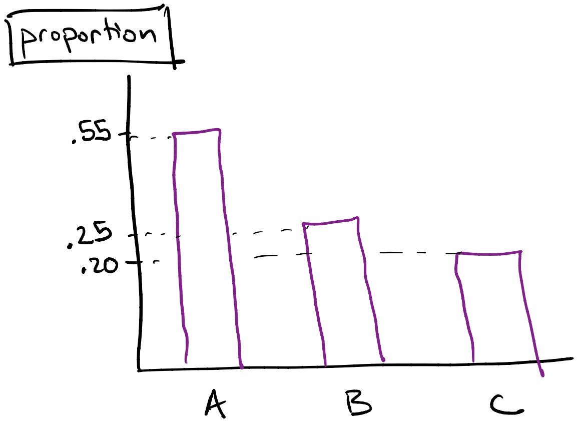


Population
Sampling frame
Sample

Random Variables

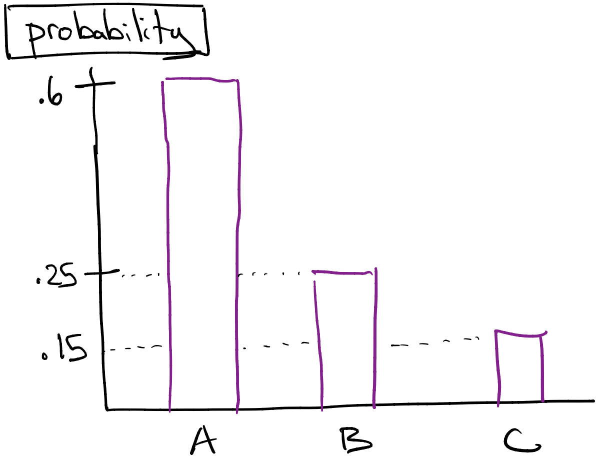
Distributions and Data Generation

Zoom Poll Data



Empirical Distribution: the distribution of your sample (values and proportions)

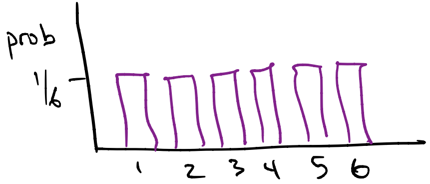
Polling a Student from the full class



Probability Distribution: a model for how the sample is generated (values and probabilities).

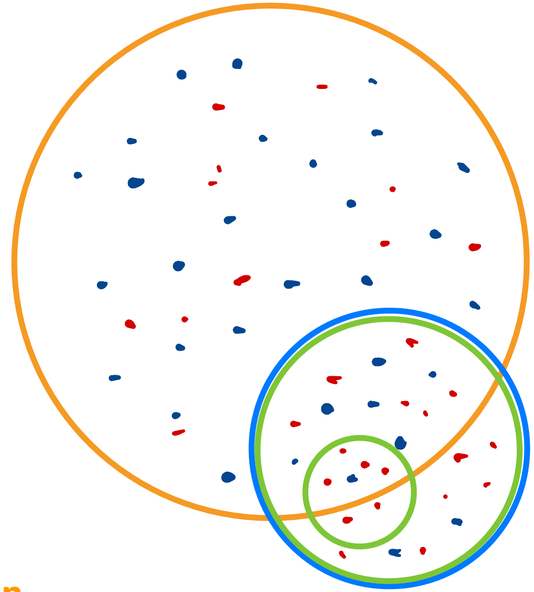
Note: Probability Distributions

- Can describe sampling from a population, but that's not all!
  - → # of pips on a die roll



- Often not known

Generating Data for FDR vs. Langdon

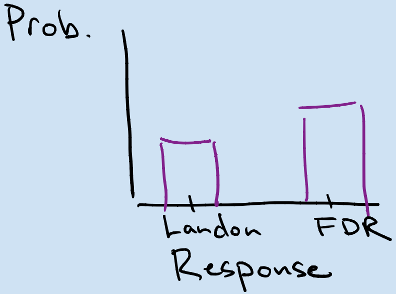


Population
Sampling frame
Sample

Probability Distributions

Sampling process 1: draw n = 10 million

Sampling process 2: draw n = 1 person



Random Variable

A **random variable** is a variable that can takes numerical values with particular probabilities.

Example 1: Let X take the value 1 if FDR, 0 if Landon.

Example 2: Let Y be the # of pips of a roll of a 6-sided die.

Notation:

- Random Variables (RVs) use capital letters: X, Y, Z
- A particular value taken by a RV indicated by a lower case letter. x, y, z
- The (Probability) Distribution of a discrete R.V. can be expressed as a table or graphic.

$$P(X = x)$$

probability

RV. X

particular value x

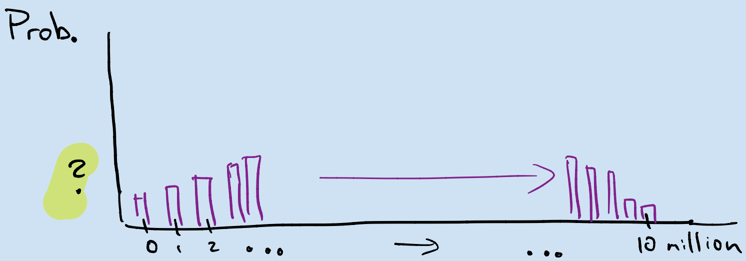
Functions of Random Variables

A function of random variables is *also* a random variable.

Example 1, cont.:

Let S be the total # of voters that say "FDR" in a sample of size 10 million

$$S = \overset{\text{1st response}}{X_1} + \overset{\text{2nd response}}{X_2} + \dots + \overset{\text{last response}}{X_{10M}}$$



Abstracting Random Chance

Q: What do these have in common?

Ask a randomly drawn American who they plan to vote for

The outcome of a coin flip

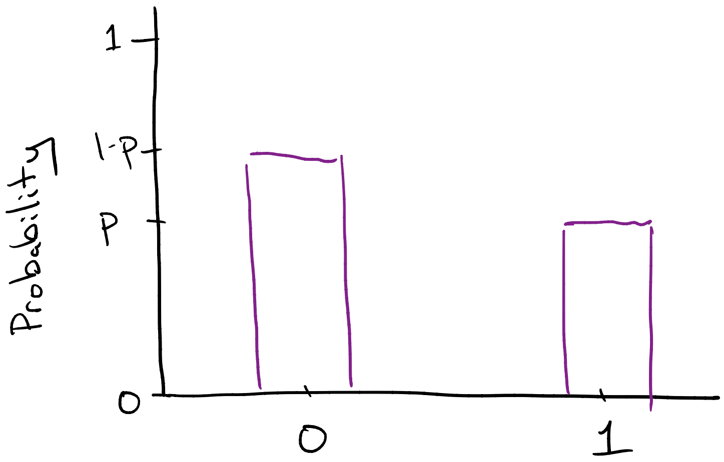
The outcome of a COVID test for a randomly selected Californian

A: Each have only two outcomes, one of which happens w/ a particular probability p.

* note the little p

Bernoulli Distribution

A random variable that takes the value 1 with probability p and 0 otherwise.



X is Bernoulli(p) if:

$$P(X=1) = p$$
$$P(X=0) = 1-p$$

Probability Mass Function (PMF)

Examples:

Ask a randomly drawn American who they plan to vote for	The outcome of a coin flip	The outcome of a COVID test for a randomly selected Californian
Bernoulli(p=.61)	Bernoulli(p=.5)	Bernoulli(p=.02)

Abstracting Random Chance

Q: What do these have in common?

Count the number of people that answered “FDR” in a sample of n = 10

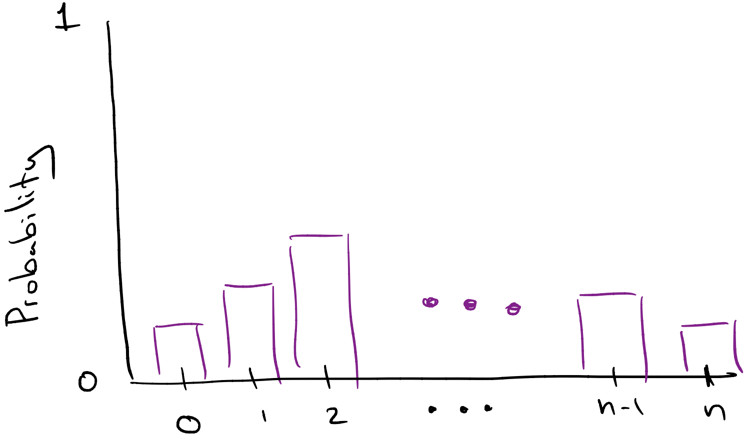
The total number of heads in a series of 5 coin flips.

The total number of Californians that will test positive for COVID in a given month.

A: Each is a sum of Bernoulli RVs.

Binomial Distribution

A random variable that counts the number of “successes” in n independent trials where each succeeds with probability p.

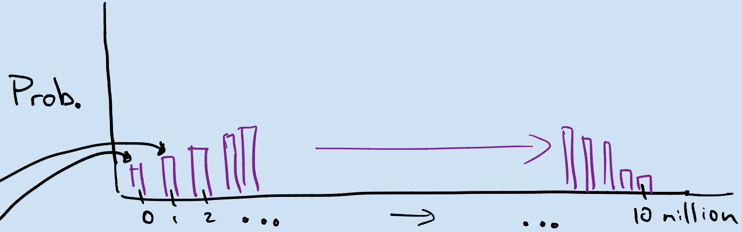


Y is binomial(n, p) if:

$$P(Y=y) = \binom{n}{y} p^{y} (1-p)^{n-y}$$

Recall: $S = X_1 + X_2 + \dots X_{10M}$

S is binomial($n = 10M$, $p = .61$)



$$P(S=0) = \binom{10M}{0} .61^{0} \, .39^{10M}$$

$$P(S=1) = \binom{10M}{0} .61^{1} \, .39^{10M-1}$$

Abstracting Random Chance

A random variable that counts the number of "successes" in n independent trials where each succeeds with probability p.

Q: What do these have in common?

Count the number of people that answered "FDR" in a sample of n = 10 M

binomial(n = 10 M, p = .61)

- each X_i is not quite independent with the same p.

The total number of heads in a series of 5 coin flips.

binomial(n = 5, p = 1/2)

- good fit!

The total number of Californians that will test positive for COVID in a given month.

~~binomial(n = .5M, p = .08)~~

- ~~probably not independent.~~
  ~~→ contagious!~~
- ~~probably not a good fit~~

Types of distributions

Probability distributions largely fall into two main categories.

- **Discrete**.
  - The set of possible values that X can take on is either finite or countably infinite.
  - Values are separated by some fixed amount.
  - For instance, X = 1, 2, 3, 4, …
- **Continuous**.
  - The set of possible values that X can take on is uncountable.
  - Typically, X can be any real number in some interval (not just our counting numbers).

Here, we will focus almost exclusively on discrete distributions. However, it's important to know that continuous distributions exist. They will reappear later on! (bias-variance tradeoff, KDEs).

Common distributions

Discrete

- Bernoulli (p).
  - Takes on the value 1 with probability p, and 0 with probability 1-p.
- Binomial (n, p).
  - Number of 1s in n independent Bernoulli (p) trials.
  - Probabilities given by the binomial formula.
- Uniform on a finite set.
  - Probability of each value is 1 / (size of set). For example, a standard die.

Continuous

- Uniform on the unit interval.
  - U could be any real number in the range [0, 1].
- Normal (μ, σ^2).

Parameters of a distribution are the constants associated with it. These define its shape and the values it takes on. These are the numbers provided in parentheses. (https://ismay.shinyapps.io/ProbApp/)

Poll: How many total heads would you expect to get in 5 flips of a fair coin?

Expected Value

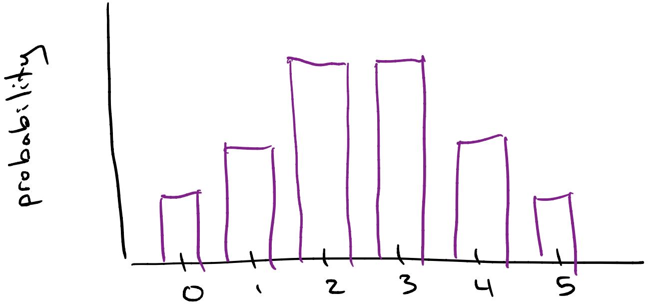
The **expected value** of a random variable X is the weighted average of the values of X, where the weights are the probabilities of the values.

$$E(X) = \sum_{\text{all } x_i} x_i \, P(X = x_i) = \mu$$

- Expected value is a number, not a random variable
- It is analogous to the average.
  - It has the same units as the random variable.
  - It doesn't need to be a possible value of the random variable.
  - It is the center of gravity of the probability histogram.

Ex:

Let Y be the # H in 5 coin tosses.



$$E(Y) = 0 \cdot P(Y=0) + \dots + 5 \cdot P(Y=5)$$

$$\uparrow \binom{5}{0} .5^0 .5^5$$

Properties of Expected Values

Linear transformations

constants

Let

$$Z = aX + b \; ; \quad E(Z) = E(aX+b) = aE(X) + E(b) = aE(X) + b$$

Additivity

Let

$$W = X_1 + X_2 \; ; \quad E(W) = E(X_1 + X_2) = E(X_1) + E(X_2)$$

Linearity of Expectation

Let

$$V = aX_1 + bX_2 \; ; \quad E(V) = E(aX_1 + bX_2) = aE(X_1) + bE(X_2)$$

Calculating Expected Values

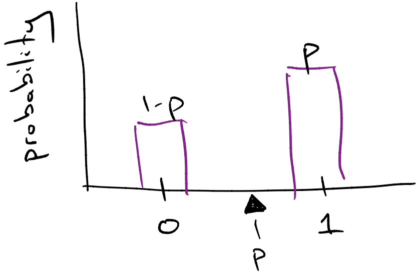
Bernoulli

Let X be Bernoulli(p).

$$E(X) = \sum_{\text{all } x_i} x_i P(X = x_i)$$

$$= 0 \cdot (1-p) + 1 \cdot p$$

$$= p$$



Binomial

Let Y be binomial(n,p).

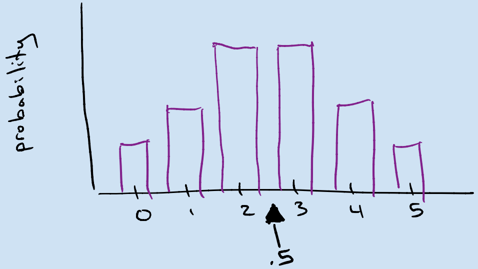
$$Y = X_1 + X_2 + \dots + X_n$$

$$E(Y) = E(X_1 + X_2 + \dots + X_n)$$

$$= E(X_1) + E(X_2) + \dots E(X_n)$$

$$= p + p + \dots + p$$

$$= np$$



$n = 5,\ p = .5$

$np = 2.5$

Random Variables: Summary

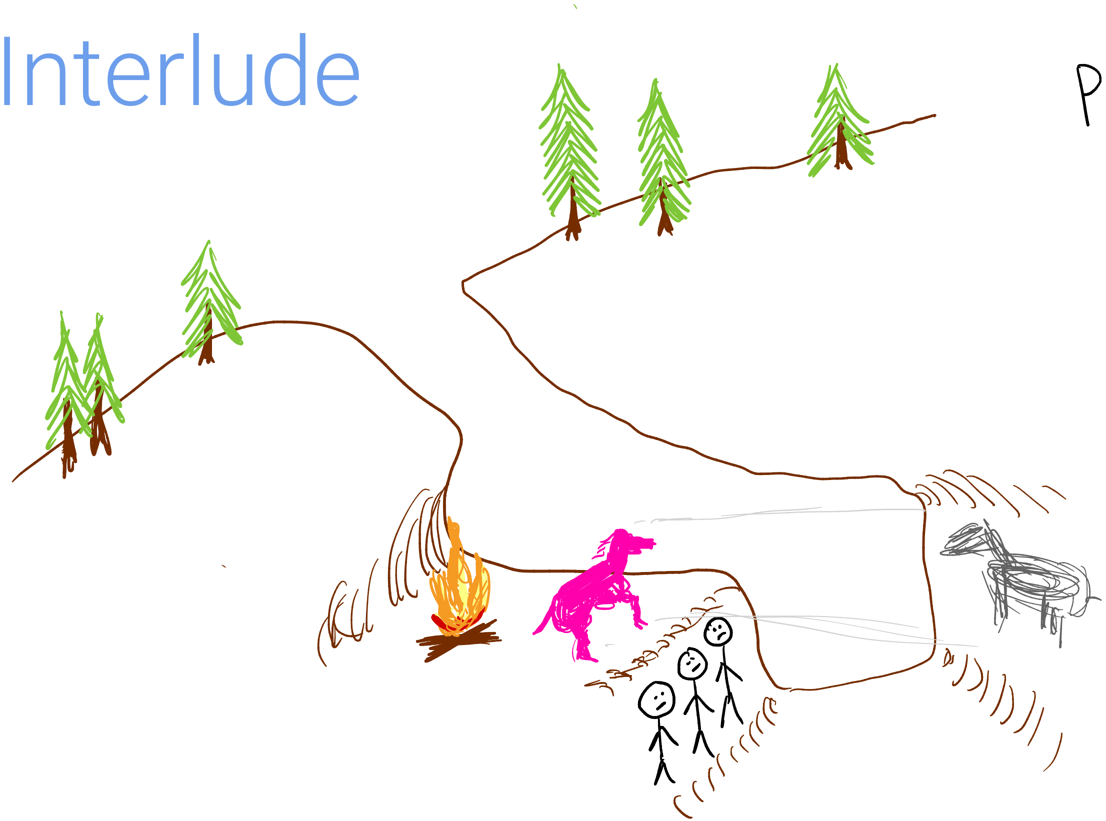
- In order to understand the world, you need to know how your data was generated
- Random Variables and their distribution formalize that process
- Many RVs reoccur and have been given names
- One of the most prominent features of an RV is its expected value.

Where we’re headed today

What is **statistical bias**?

The difference between your estimate and the truth.

Interlude



Plato's Allegory of the Cave

World of Forms

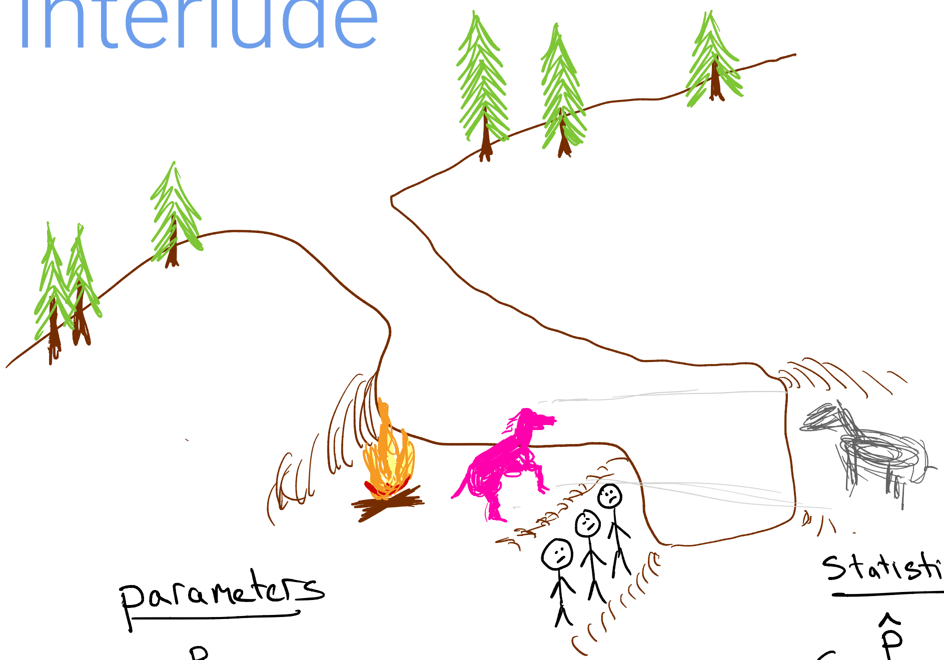
- Non-physical essence of all things.

World of Representation

- The material world that we observe.

Philosopher: Person who seeks knowledge of forms.

Interlude



parameters

$$\theta \begin{cases} p \\ \mu \\ \sigma \\ \hat{\beta} \end{cases}$$

Statistics

$$\hat{\theta} \begin{cases} \hat{p} \\ \hat{\mu},\ \bar{x} \\ \hat{\sigma},\ s \\ \hat{\beta}_1 \end{cases}$$

Metaphor of the Cave

World of Parameters

- Constants that define the structure of the world

World of Data/Statistics

- Observable information generated by RVs and their parameters.
- Statistic: numerical summary of data.

Statistician: person who uses statistics to learn about parameters.

What is a statistic?

- A single piece of data.
- A $\underbrace{\text{numerical summary}}_{\text{function}}$ of a $\underbrace{\text{dataset}}_{\text{realizations of R.V.s}}$.

$$\hat{\theta} = f(x_1, x_2, \dots x_n)$$

What is an estimator?

- A statistic designed to estimate a parameter

Choosing a statistic/estimator

Example 1: Squirrels

Question: How many squirrels are there in Central Park, New York City?

Goodrum. After 50–110 observation periods, the above-recorded data were used to make an estimate of the squirrel population ($\hat{P}$) of a woodlot. Six different estimates were made by this method. The formula employed was

$$\hat{P} = \frac{AZ}{(0.6)\,\Pi\, Sy^2} \quad \text{where}$$

A = total area of the woods (in each case, 10 acres);
Z = number of squirrels seen;
S = number of 15-minute observation periods;
y = average of all distances from the observer to the squirrels seen.

The constant 0.6 was used because it was believed that only that much of the circle around the observer could be well seen.

unfortunate notation!

Parameter: total # squirrels in Central Park $\quad P$

The Data: z, y

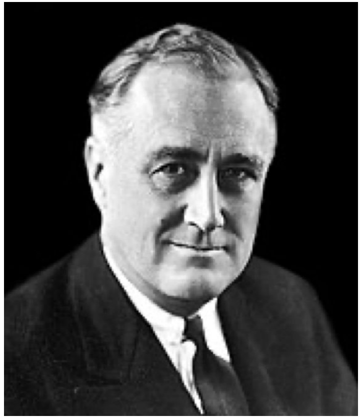
The Estimator

$$\hat{P} = f(z, y; A, S, .6, \pi)$$

Choosing a statistic/estimator

Example 2: FDR vs. Langdon

Question: what proportion of Americans will vote for FDR?



Roosevelt (D)

Landon (R)

The Parameter: the total proportion of votes for FDR, p

The Data: $X_1, X_2, \ldots X_{10M}$

The Estimator:

$$\hat{p} = f(x_1, x_2, \ldots x_{10M}; n)$$

$$= \frac{x_1 + x_2 + \ldots x_{10M}}{n}$$

What is **statistical bias**?

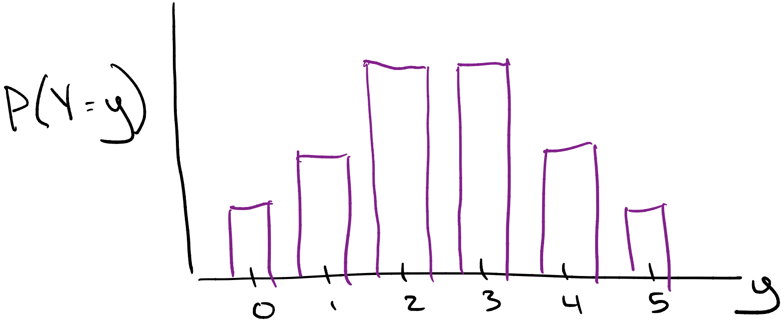
The difference between your estimate and the truth.

$$\hat{\theta} - \theta$$

$\rightarrow$ Not quite...

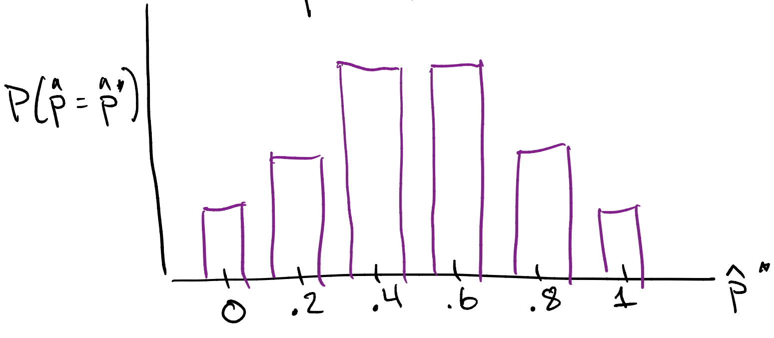
Don’t forget about sampling variability

Example: The total number of heads in a series of 5 coin flips. $\rightarrow Y = X_1 + X_2 + \dots X_5$



OR if we want to estimate p:

$\rightarrow \hat{p} = \frac{1}{n} Y$

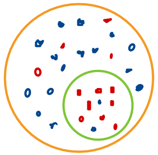


- Y is an RV, therefore $\hat{p}$ is an RV.
- Since estimators are fxns of RV's, they are RVs → subject to <u>sampling variability</u>

What is **statistical bias**?

Expected Value of Estimator

parameter



The difference between your estimate and the truth.

$$E(\hat{\theta}) - \theta$$

θ_{other}

$$\hat{\theta} = f(X_1, X_2 \dots)$$

Q: What if the data wasn't generated by θ?

A: It will not be representative of the population.

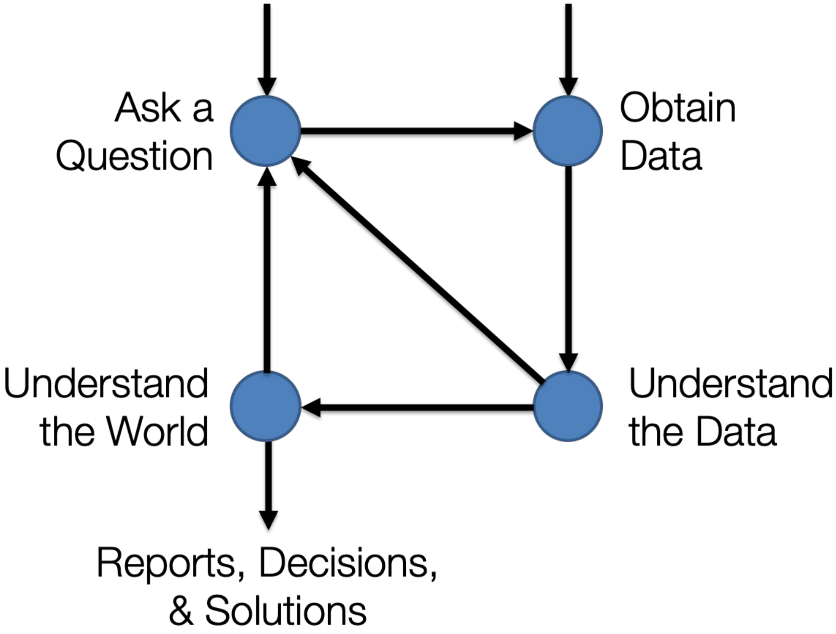
↳ selection bias

Q: What if your estimator isn't great?

A: Biased Estimator

Ex. $\hat{p}_b = \frac{1}{n-1} Y_n$

What's Next?



Ask a Question
Obtain Data
Understand the World
Understand the Data
Reports, Decisions, & Solutions