LECTURE 2

Data Sampling and Probability

How to sample effectively, and how to quantify the samples we collect.

Data 100/Data 200, Spring 2022 @ UC Berkeley

Josh Hug and Lisa Yan

Today's Roadmap

Lecture 02, Data 100 Spring 2022

- **Review of last lecture**
- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- Probability Samples
- Application: The Gallup Poll today
- Multinomial and Binomial probabilities
- Extra: Permutations and Combinations

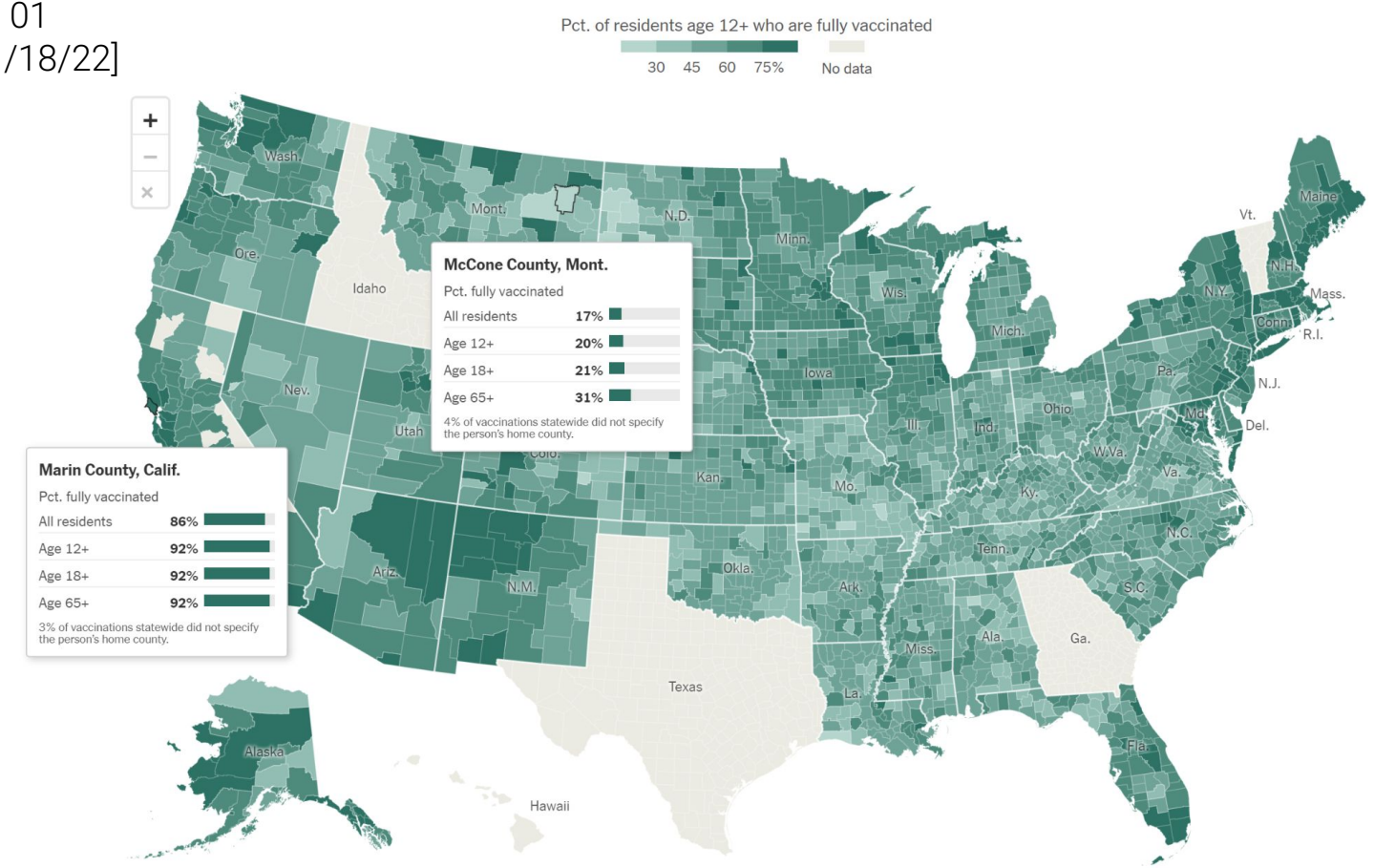
Course website walkthrough

- https://ds100.org/sp22/
- Lab 1 (out Friday): Detailed JupyterLab walkthrough
- Homework 1 (out Friday)
- Discussion 1 (online Friday, Discussion Zoom Links on Ed)
    - Discussion attendance will be taken across all sections for first 3 weeks, but try to attend the one that you have signed up for (locked in starting Week 4)
    - Early next week we will reach out re: online/conflicting discussion times.

What message is being conveyed?

From Lecture 01
[URL visited 1/18/22]

Map Link

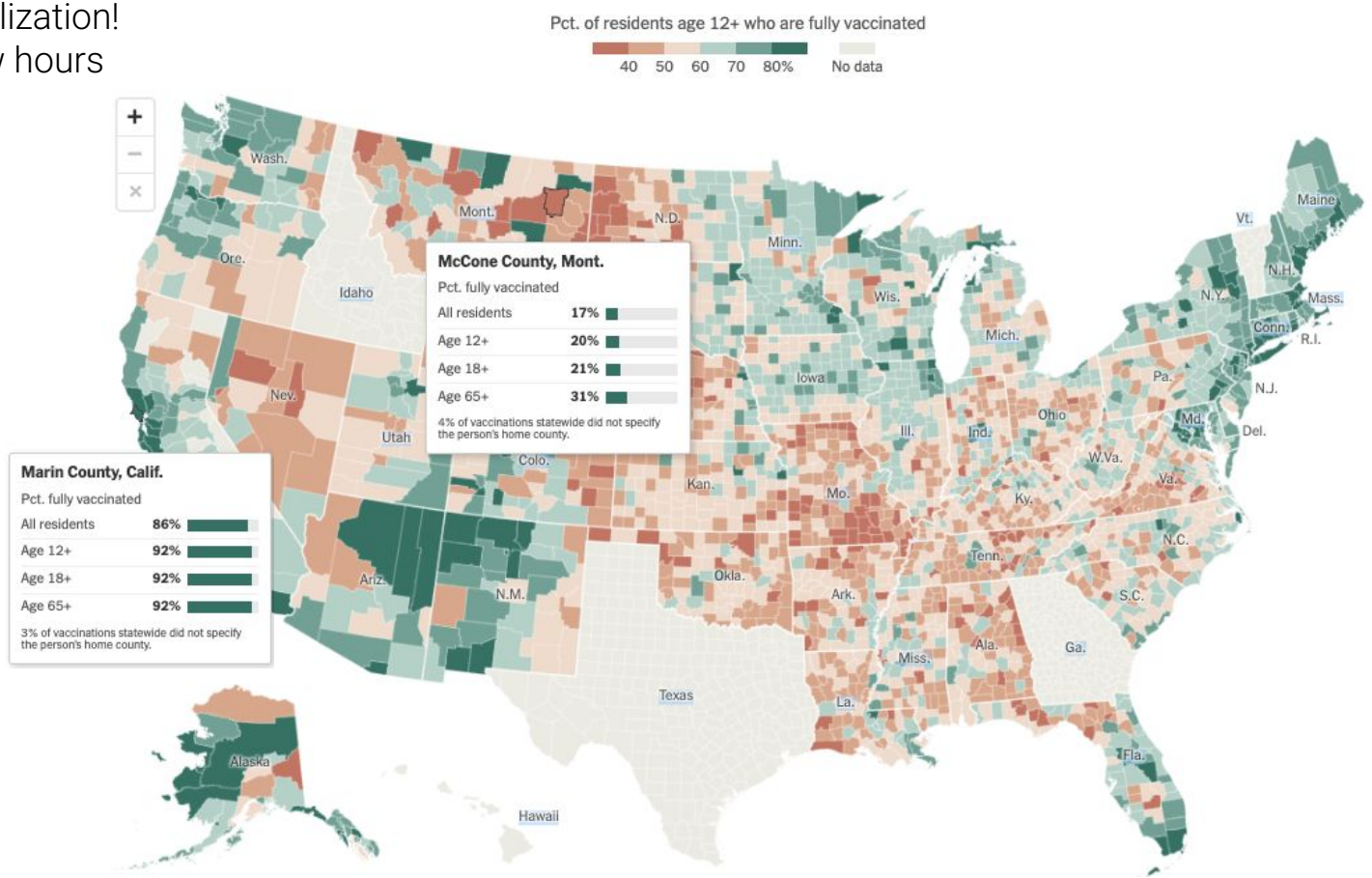


Sources: Centers for Disease Control and Prevention; Texas Department of State Health Services; Colorado

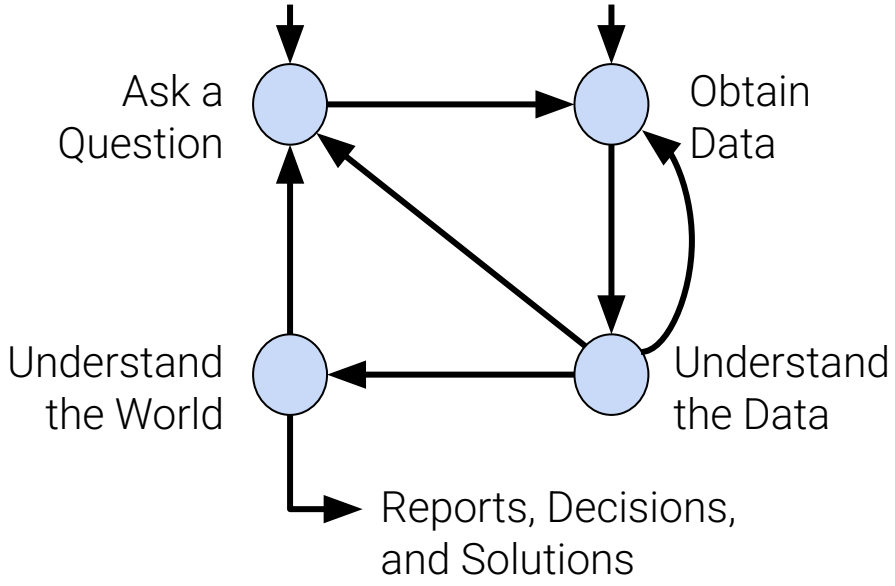
What message is being conveyed?

The new visualization! [updated a few hours after lecture]

Map Link

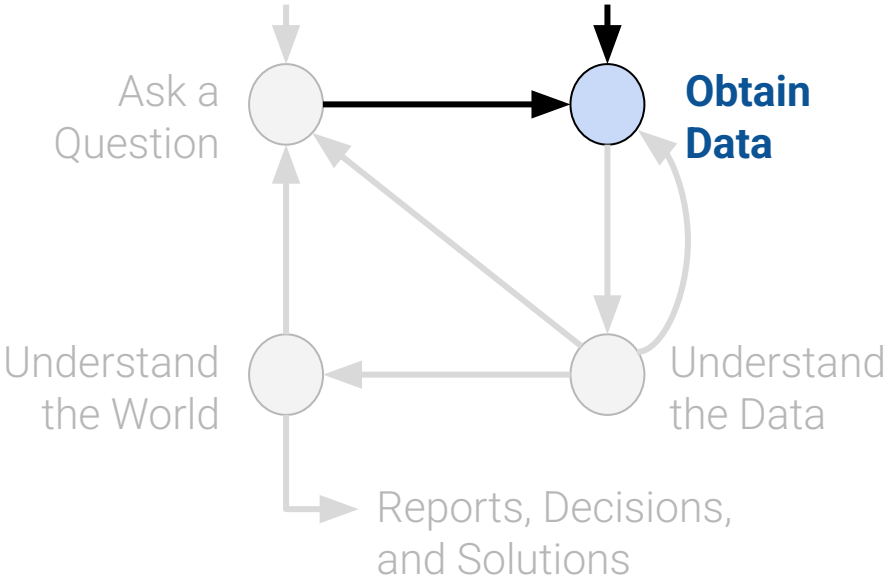


We call this the
Data Science Lifecycle.



Today

How do we collect data?



Censuses and Surveys

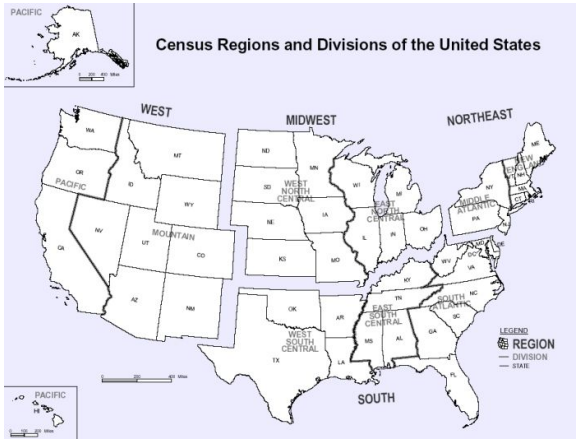
Lecture 02, Data 100 Spring 2022

- Review of last lecture
- **Censuses and Surveys**
- Sampling: Definitions
- Bias: A Case Study
- Probability Samples
- Application: The Gallup Poll today
- Multinomial and Binomial probabilities
- Extra: Permutations and Combinations

The US Decennial Census

- Was held in April 2020.
- Counts **every person** living in all 50 states, DC, and US territories. (Not just citizens.)
- Mandated by the Constitution. Participation is required by law.
- Important uses:
  - Allocation of Federal funds.
  - Congressional representation.
  - Drawing congressional and state legislative districts.

In general: a **census** is “an official count or survey of a **population**, typically recording various details of individuals.”



data.census.gov

Survey

> In general: a **census** is "an official count or **survey** of a population, typically recording various details of individuals."

A **survey** is a set of questions.

- For instance: workers survey individuals and households.

What is asked, and how it is asked, can affect:

- How the respondent answers.
- **Whether** the respondent answers.

There are entire courses on surveying!
See Stat 152 at Berkeley (Sampling Surveys).

FiveThirtyEight

Politics Sports Science & Health Economics Culture

JUN. 27, 2019, AT 12:42 PM

The Supreme Court Stopped The Census Citizenship Question — For Now

By Amelia Thomson-DeVeaux

NATIONAL

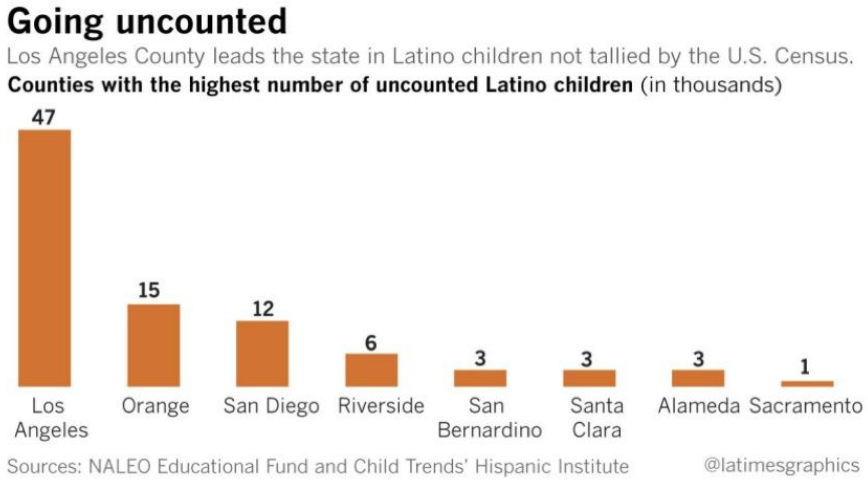
Citizenship Question To Be Removed From 2020 Census In U.S. Territories

August 9, 2019 · 3:23 PM ET

FiveThirtyEight, NPR

Undercounting in the US Decennial Census

LA Times 2010 Census



How do we know these numbers?
From other surveys.

WaPo 2000 Census

High Court Rejects Sampling In Census
Ruling Has Political, Economic Impacts

Sampling methods would estimate Americans who missed the survey.
— Most often minoritized/poor who vote Dem.
— "The better way is to improve the methods for contacting and questioning every household"

NY Times 2020 Census

In 2020 Census, Big Efforts in Some States. In Others, Not So Much.

California is spending $187 million to try to ensure an accurate count of its population. The Texas Legislature decided not to devote any money to the job. Why?

Sampling: Definitions

Lecture 02, Data 100 Spring 2022

- Review of last lecture
- Censuses and Surveys
- **Sampling: Definitions**
- Bias: A Case Study
- Probability Samples
- Application: The Gallup Poll today
- Multinomial and Binomial probabilities
- Extra: Permutations and Combinations

Sampling from a finite population

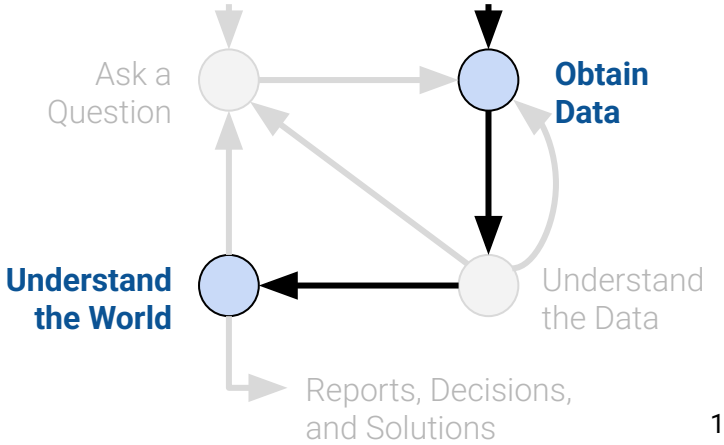
A census is great, but expensive and difficult to execute.

- Would **all** voters be willing to participate in a voting census prior to an actual election?

A **sample** is a subset of the population.

- Samples are often used to make **inferences about the population**.
- How you draw the sample will affect your accuracy.
- Two common sources of error:
  - **chance error**: random samples can vary from what is expected, in any direction.
  - **bias**: a systematic error in one direction.

Inference: quantifying degree of certainty in our models of the world.
Data 8 book



Population, sample, and sampling frame

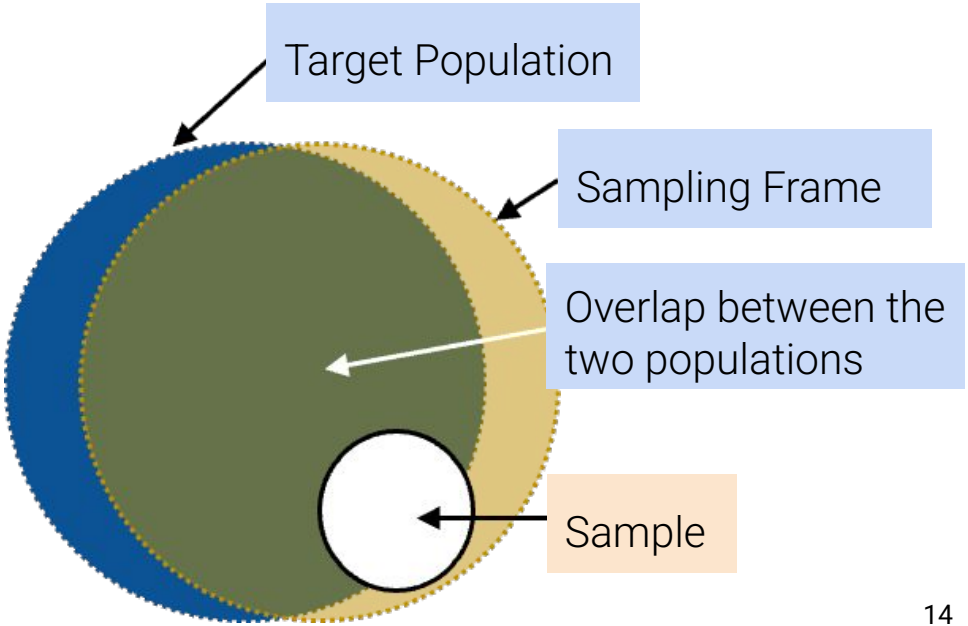
Population: The group that you want to learn something about.

Sampling Frame: The list from which the sample is drawn.

- If you're sampling people, the sampling frame is the set of all people that could possibly end up in your sample.

Sample: Who you actually end up sampling.

- A subset of your sampling frame.



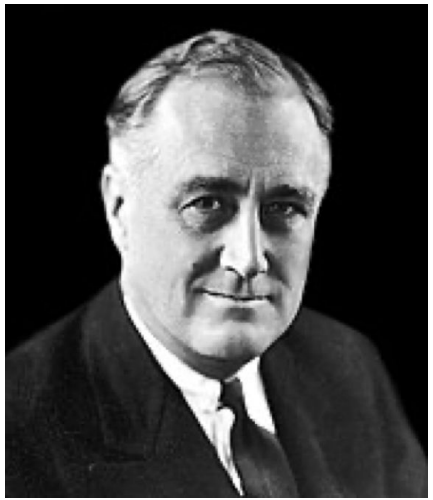
There may be individuals in your **sampling frame** (and hence, your sample) that are **not** in your population!

Bias: A Case Study

Lecture 02, Data 100 Spring 2022

- Review of last lecture
- Censuses and Surveys
- Sampling: Definitions
- **Bias: A Case Study**
- Probability Samples
- Application: The Gallup Poll today
- Multinomial and Binomial probabilities
- Extra: Permutations and Combinations

Case study: 1936 Presidential Election



Roosevelt (D)

Landon (R)

In 1936, President Franklin D. Roosevelt (left) went up for re-election against Alf Landon (right). As is usual, **polls** were conducted in the months leading up to the election to try and predict the outcome.

(Election result spoiler: Landon was not a U.S. President)

The Literary Digest: Election Prediction

The Literary Digest was a magazine. They had successfully predicted the outcome of 5 general elections coming into 1936.

They sent out their survey to **10,000,000** individuals, who they found from:

- Phone books.
- Lists of magazine subscribers.
- Lists of country club members.

	% Roosevelt	**# surveyed**
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000

How could this have happened?
They surveyed 10 million people!

The Literary Digest: What happened?

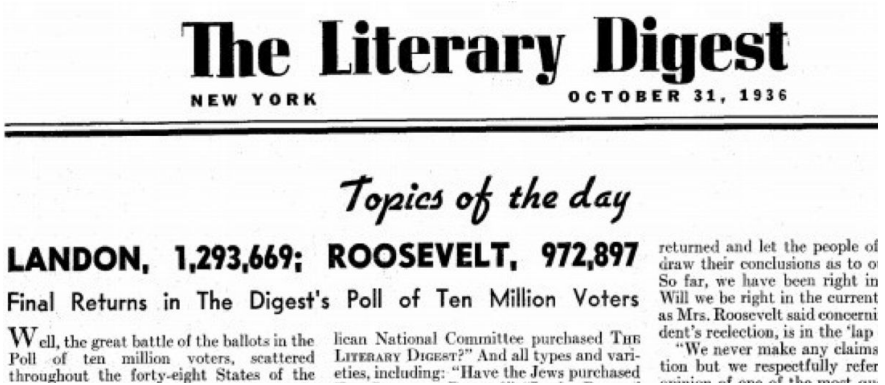
(1) The Literary Digest sample was **not representative** of the population.

- The Digest's **sampling frame**: people in the phonebook, subscribed to magazines, and went to country clubs.
- These people were more affluent and tended to vote Republican (Landon).

	% Roosevelt	# surveyed
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000

(2) Only 2.4 million people **actually filled out the survey!**

- 24% response rate (low).
- Who knows how the 76% **non-respondents** would have polled?



The Literary Digest

NEW YORK — OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the

lican National Committee purchased The Literary Digest?" And all types and varieties, including: "Have the Jews purchased

returned and let the people of draw their conclusions as to o So far, we have been right in Will we be right in the current as Mrs. Roosevelt said concerni dent's reelection, is in the 'lap

"We never make any claims tion but we respectfully refer

Gallup's Poll: Election Prediction

George Gallup, a rising statistician, also made predictions about the impending 1936 elections.

Not only was his estimate much closer than The Literary Digest's estimate, but he did it with a **sample size of only 50,000**!

George Gallup also predicted what The Literary Digest was going to predict, within 1%, with a **sample size of only 3000 people**.

- He predicted the Literary Digest's **sampling frame** (phonebook, magazine subscribers, country clubs).
- So he sampled those same individuals!

	% Roosevelt	**# surveyed**
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000
George Gallup's poll	56%	50,000
George Gallup's prediction of Digest's prediction	44%	3,000

Samples, while convenient, are subject to chance error and **bias**.

Adapted from *Statistics* by Freedman, Pisani, and Purves, W.W. Norton NY

Common Biases

Selection Bias

- Systematically excluding (or favoring) particular groups.
- How to avoid: Examine the sampling frame and the method of sampling.

Response Bias

- People don't always respond truthfully.
- How to avoid: Examine the nature of questions and the method of surveying.

Non-response Bias

- People don't always respond.
- How to avoid: Keep your surveys short, and be persistent.
- People who don't respond aren't like the people who do!

Which types of bias do you think the Literary Digest sample had?

Probability Samples

Lecture 02, Data 100 Spring 2022

- Review of last lecture
- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- **Probability Samples**
- Application: The Gallup Poll today
- Multinomial and Binomial probabilities
- Extra: Permutations and Combinations

Quality, not quantity!

Try to ensure that the sample is representative of the population.

- Don't just try to get a big sample.
- If your method of sampling is bad, and your sample is big, you will have a **Big Bad Sample!**

We will now look at some common **non-random** samples, before formalizing what it means for a sample to be random.

*

*Big Bad Wolf

Common Non-Random Samples

A **convenience sample** is whoever you can get ahold of.

Example: Suppose we have a cage of mice, and each week, we want to measure the weights of these mice. To do so, we take a convenience sample of these mice and weigh them.

- Not a good idea for inference!
- Haphazard ≠ random.
- Sources of bias can introduce themselves in ways you may not think of!

A **quota sample** is where you first specify your desired breakdown of various subgroups, and then reach those targets however you can.

Example: You want to sample individuals in your town, and you want the age distribution of your sample to match that of your town's census results.

- Reaching quotas "however you can" is not random.
- Your sample will look like your population with respect to a few aspects—but not all.
  - Quotas for age will represent age.
  - What about gender? Ethnicity? income?

Probability Sample (aka Random Sample)

Why sample at random? One reason is to reduce bias, but that's not the main reason!

- Random samples **can** produce biased estimates of population characteristics.
  - For example, if we're estimating the maximum of a population.
- But with random samples we are able to **estimate the bias and chance error**.
  - We can **quantify the uncertainty**.

A **probability sample** drawn from a random sampling scheme has the following properties:

- You **must** be able to provide the chance that any specified **set** of individuals will be in the sample.
- All individuals in the population **need not** have the same chance of being selected.
- You will still be able to measure the errors, because you know all the probabilities.

Example Scheme 1: Probability Sample

Suppose I have 3 students (**A**ndrew, **B**ella, **D**ominic):
I decide to sample 2 of them as follows:

- I choose **A** with probability 1.0
- I choose either **B** or **D**, each with probability 0.5.

All subsets of 2:	{**A**, **B**}	{**A**, **D**}	{**B**, **D**}
Probabilities:	0.5	0.5	0

This is a **probability sample** (though not a great one).

- Of the 3 people in the population, I know the chance (aka **probability**) of getting each subset.
- Suppose I'm measuring the average distance students live from campus.
  - This scheme does not see the entire population!
  - My estimate using the single sample I take has some **chance error** depending on if I see AB or AD.
  - This scheme **biases** towards A's response

We'll learn more about quantifying error/bias later in a few weeks.

Common random sampling schemes

A **random sample with replacement** is a sample drawn **uniformly** at random **with** replacement.

- Random doesn't always mean "uniformly at random," but in this specific context, it does.

A **simple random sample (SRS)** is a sample drawn **uniformly** at random **without** replacement.

- **Every individual (and subset of individuals) has the same chance of being selected.**
- Every pair has the same chance as every other pair.
- Every triple has the same chance as every other triple.
- And so on.

A raffle could use either sampling scheme, depending on if winners are eligible for multiple prizes.

Example Scheme 2: Simple Random Sample?

Consider the following sampling scheme:

- A class roster has 1100 students listed alphabetically.
- Pick one of the first 10 students on the list at random.
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).

1. Is this a probability sample?

2. Does each student have the same probability of being selected?

3. Is this a simple random sample?

(2 min pause to think)

Example Scheme 2: Simple Random Sample?

Consider the following sampling scheme:

- A class roster has 1100 students listed alphabetically.
- Pick one of the first 10 students on the list at random.
- To create your sample, take that student and every 10th student listed after that (e.g. Students 8, 18, 28, 38, etc).

1. Is this a probability sample?

Yes.

For a sample [n, n + 10, n + 20, …, n + 1090], where 1 <= n <= 10, the probability of that sample is 1/10.

Otherwise, the probability is 0.

Only 10 possible samples!

2. Does each student have the same probability of being selected?

Yes.

Each student is chosen with probability 1/10.

3. Is this a simple random sample?

No.

The chance of selecting (8, 18) is 1/10; the chance of selecting (8, 9) is 0.

A very common approximation for sampling

A common situation in data science:

- We have an enormous population.
- We can only afford to sample a relatively small number of individuals.

If the **population is huge** compared to the sample, then
random sampling with and without replacement are pretty much the same.

Example: Suppose there are 10,000 people in a population.
Exactly 7,500 of them like Snack 1; the other 2,500 like Snack 2.

What is the probability that in a random sample of 20, **all people like Snack 1**?

SRS (Random Sample Without Replacement)

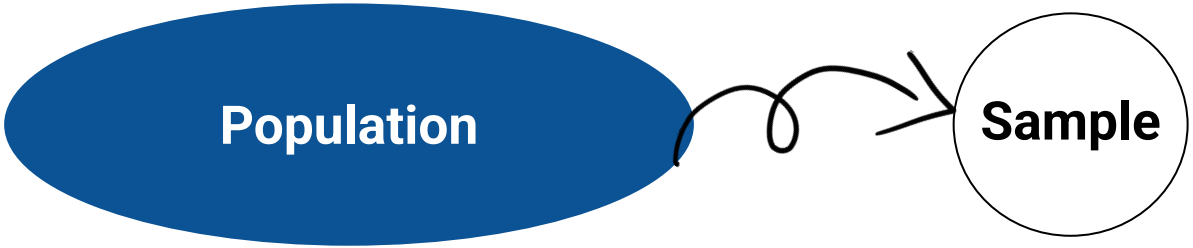
$$\left(\frac{7500}{10000}\right)\left(\frac{7499}{9999}\right)\cdots\left(\frac{7482}{9982}\right)\left(\frac{7481}{9981}\right) \approx .003151$$

Random Sample With Replacement

$$(0.75)^{20} \approx 0.003171$$

Probabilities of sampling with replacement are much easier to compute!

So far



If a sample was **randomly sampled with replacement** from the population:

- It is a probability sample.
- We can quantify error and bias (to be covered later).
- **Given the population distribution**, we can compute the probability of us getting a **particular sample**.

We almost **never** know the population distribution, unless we take a census!!
But this framing helps **quantify our certainty** in any analysis/inference using our sample.

Application: The Gallup Poll today

Lecture 02, Data 100 Spring 2022

- Review of last lecture
- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- Probability Samples
- **Application: The Gallup Poll today**
- Multinomial and Binomial probabilities
- Extra: Permutations and Combinations

The Gallup Poll today

https://media.gallup.com/PDF/FAQ/HowArePolls.pdf 1997

> The actual number of people that need to be interviewed for a given sample is to some degree less important than the soundness of the **fundamental equal probability of selection principle**…

Gallup U.S. Election polls:

- **Sampling Frame**: "civilian, non-institutionalized population" of adults in telephone households in continental US
- **Random Digit Dialing** to include both listed/unlisted phone numbers (avoid **selection bias**)
- **Within household selection process** to randomly select if ≥1 adult in household
  - If no answer, recall multiple times (avoids **non-response bias**)

→ **Simple Random Sample**! (of the defined sampling frame)

According to Gallup's report:

> …**question wording** is probably the **greatest source of bias and error** in the data, followed by question order.

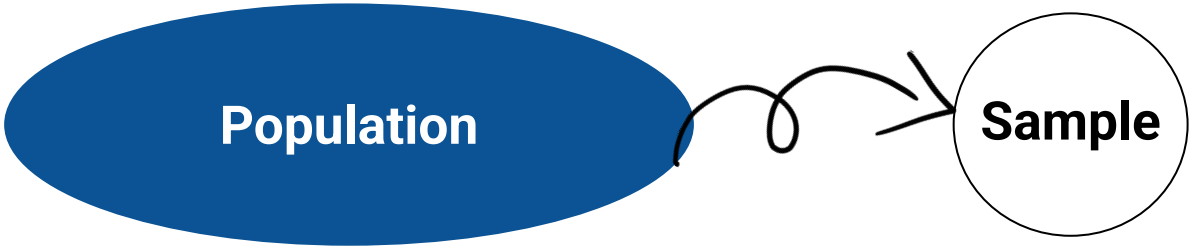
Response bias is their (claimed) biggest issue!

Multinomial and Binomial Probabilities

Lecture 02, Data 100 Spring 2022

- Review of last lecture
- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- Probability Samples
- Application: The Gallup Poll today
- **Multinomial and Binomial probabilities**
- Extra: Permutations and Combinations

So far



If a sample was **randomly sampled with replacement** from the population:

- It is a probability sample.
- We can quantify error and bias (to be covered later).
- **Given the population distribution**, we can compute the probability of us getting a **particular sample**.

We almost **never** know the population distribution, unless we take a census!!
But this framing helps **quantify our certainty** in any analysis/inference using our sample.

Special case: Random sampling with replacement of a **Categorical population distribution** produces **Multinomial/Binomial Probabilities**.

The scenario

Binomial and multinomial probabilities arise when we:

- Sample at random, **with replacement**.
- Sample a fixed number (n) times.
- Sample from a **categorical distribution**.
  - If 2 categories, **Binomial**:
    Bag of marbles: 60% blue 40% not blue
  - If >2 categories, **Multinomial**:
    Bag of marbles: 60% blue 30% green 10% red

Goal: **Count the number of each category** that end up in our sample.

- `np.random.multinomial` returns these counts (Homework 1).
- We'll derive the multinomial probabilities in this section as a review of probability.

Note: In this section, we will make use of the binomial coefficient and factorials.
A refresher is at the end of this lecture covering the derivation.

Binomial probability: Two categories

Suppose we sample at random with replacement 7 times from a bag of marbles:

60% **blue** marbles 40% **not** blue marbles.

Q1. What is P(**bnbbbnn)?**

Q2. Fill in the blank with **<**, **=**, or **>**: P(4 **blue**, 3 **not** blue) ____ P(**bnbbbnn**)

(1 min pause to think)

Binomial probability: Two categories

Suppose we sample at random with replacement 7 times from a bag of marbles:

60% **blue** marbles 40% **not** blue marbles.

Q1. What is P(**bnbbbnn**)**?**

By the product rule from Data 8, since the sample is drawn with replacement:

P(**bnbbbnn**) $= 0.6 \times 0.4 \times 0.6 \times 0.6 \times 0.6 \times 0.4 \times 0.4 = (0.6)^4(0.4)^3$

Q2. Fill in the blank with **<**, **=**, or **>**: P(4 **blue**, 3 **not** blue) **>** P(**bnbbbnn**)

Why? **bnbbbnn** is a specific **order**. It is far more restrictive and specific than the **count** 4 **blue**, 3 **not** blue.

Binomial probability (cont.)

Q2. "4 **blue**, 3 **not** blue" can occur in **several equally likely** ways.
For instance, P(**b**n**bbb**n**n**) = P(**bbbb**n**nn**) = P(**nnn**b**bbb**) = … = $(0.6)^4(0.4)^3$.

Typo

P(4 **blue**, 3 **not** blue) is the **total** chance of all of those ways.

and thus,

$$\text{P(4 blue, 3 not blue)} = \frac{7!}{4!\,3!}(0.6)^4(0.4)^3$$

$$= \underbrace{\binom{7}{4}}\,\underbrace{(0.6)^4(0.4)^3}$$

binomial probability

This expression arises from the **sum rule** and **product rule** of probability.

# of ways to choose 4 of 7 places to write **b**
(other 3 get filled with **n**)

$$\binom{7}{4} = \frac{7!}{4!3!}$$

For a particular outcome, probability of this **ordered series** of **b**'s and **n**'s (Q1)

Multinomial probability: Multiple categories

Now suppose we sample at random with replacement 7 times from a bag of marbles:

60% **blue** marbles 30% are **green** 10% are **red**.

Q1. What is P(**bgbbbgr**)?

Like before, use product rule to determine probability for a particular **order**:

$$P(\text{bgbbbgr}) = 0.6 \times 0.3 \times 0.6 \times 0.6 \times 0.6 \times 0.3 \times 0.1 = (0.6)^4(0.3)^2(0.1)^1$$

Q2. What is P(4 **blue**, 2 **green**, 1 **red**) ?

$$\frac{7!}{4!\,2!\,1!}(0.6)^4(0.3)^2(0.1)^1$$

multinomial probability

Like before, use **addition rule** and **multiplication rule**:

# of ways to choose 4 of 7 places to write **b**, then choose 2 places to write **g** , (other 1 get filled with **r**)

For a particular outcome (say, Q1), probability of this **ordered series** of **b**'s, **g** 's, and **r**'s

Generalization of multinomial probabilities

If we are drawing at random with replacement **n** times, from a population broken into three separate categories (where $p_1 + p_2 + p_3 = 1$):

- Category 1, with proportion $\mathbf{p_1}$ of the individuals.
- Category 2, with proportion $\mathbf{p_2}$ of the individuals.
- Category 3, with proportion $\mathbf{p_3}$ of the individuals.

Then, the **multinomial probability** of drawing $\mathbf{k_1}$ individuals from Category 1, $\mathbf{k_2}$ individuals from Category 2, and $\mathbf{k_3}$ individuals from Category 3 (where $k_1 + k_2 + k_3 = n$) is

$$\frac{n!}{k_1!k_2!k_3!}p_1^{k_1}p_2^{k_2}p_3^{k_3}$$

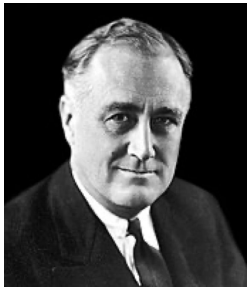
At no point in this class will you be forced to memorize this! This is just for your own understanding. In practice (as you will see in Homework 1), we use `np.random.multinomial` to compute these quantities.

Homework 1

Prerequisites (Calculus, Probability, Linear Algebra, Programming)

1936 U.S. Election:

- The *Literary Digest*'s sampling scheme was biased and did not represent the population. Their prediction was way off.
- But can we **quantify** this takeaway? What is the likelihood that the *Digest*'s differences arose simply due to **chance error** in their sample?



Roosevelt (D)

Landon (R)

We know the actual population distribution (i.e., election results).

- Assume the *Digest* did random sampling with replacement from the population.
- Simulate many different samples and generate many different predictions
- Draw a conclusion.

You have seen this process before in Data 8!! [Hypothesis Testing](#).

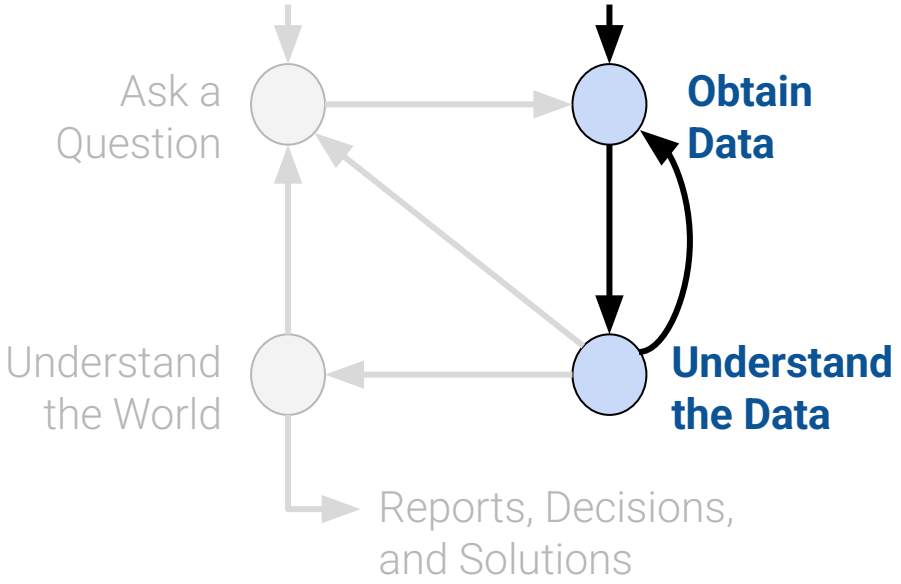
	% Roosevelt	**# surveyed**
Actual election	**61%**	**All voters** (~45,000,000)
The Literary Digest poll	43%	10,000,000

Wrapping up

We will return to the question of sampling in a few weeks.

For now, we will switch gears by moving to the **next phase** of the Data Science Life Cycle:

- Once we have obtained data, how do we analyze it, clean it, and visualize it in Python?



Extra: Permutations and Combinations

Lecture 02, Data 100 Spring 2022

- Review of last lecture
- Censuses and Surveys
- Sampling: Definitions
- Bias: A Case Study
- Probability Samples
- Application: The Gallup Poll today
- Multinomial and Binomial probabilities
- **Extra: Permutations and Combinations**

Extra: permutations and combinations

YouTube recordings (Summer 2021):
Part 1, Part 2

Disclaimer

- **This is not a class on probability or combinatorics.**
  - In other words, this is not Stat 140 or CS 70.
- This content on its own is not in scope.
  - In other words, you will never be asked questions like “how many permutations of MISSISSIPPI are there”.
  - Instead, we present it in order to (re)explain where the binomial coefficient comes from.
- This is purely meant to serve as a refresher, and is for your understanding only.
- Also, the video will walk through this material in perhaps a more natural fashion.

Permutations

Suppose I have five people, boringly named A, B, C, D, and E. **In how many ways can I arrange them in a line?**

- There are 5 options for who can end up first in line (anyone could).
- Given that, there are 4 options for who can end up second in line.
  - Could be anyone, other than whoever was first (5 - 1 = 4).
- Given that, there are 3 options for who can end up third in line, and so on.

Think of each blank as a position in line, and the number in the blank as the number of people that could end up there.

__5__ __4__ __3__ __2__ __1__

The result is **5 * 4 * 3 * 2 * 1** = 120, which we denote as **5!** (read “five factorial”). In general,

n! = n * (n - 1) * (n - 2) * … * 3 * 2 * 1

Permutations

How many ways can I arrange 3 of {A, B, C, D, E} in a line?

- 5 options for who is first.
- 4 options for who is second.
- 3 options for who is third.
- Nobody after these three.

__5__ __4__ __3__

This result is **5 * 4 * 3 = 60**. Note, we can also write 5 * 4 * 3 as

5 * 4 * 3 = (5 * 4 * 3 * 2 * 1) / (2 * 1)

= 5! / 2! = **5! / (5 - 3)!**

What are the implications of this?

Permutations

In general: if I have **n objects**, and want to **select k** of them in a way that **order matters**, then the number of ways I can do this is

$$\frac{n!}{(n-k)!}$$

Again: **this is not a class that covers counting!**

- This result, by itself, will (almost certainly) never appear again in this class.
- It's more here to serve as an intermediate step in what comes next.

Combinations

Now, suppose I want to select three people from {A, B, C, D, E}, but in a way that **order does not matter**.

- If order mattered, ABE, EAB, BAE, etc. would count as different arrangements.
- But if order does not matter, then the above three arrangements are all the same – they contain the same 3 people!
- If order does not matter, there are **fewer** ways to make our selections.

How can we use our previous answer to help us here?

- How many times did we overcount?
- Each unique group of three people is counted 3! = 6 times.
  - ABE, AEB, BAE, BEA, EAB, EBA are really all the same now.
  - We need to divide our previous answer by the number of times we overcounted!

Combinations

Dividing our previous answer by **3!** yields $\frac{\frac{5!}{2!}}{3!} = \frac{5!}{2!3!}$

This quantity is the **number of ways we can select 3 objects from a set of 5, in a way that order does not matter.**

More generally, the number of ways we can select **k** objects from a set of **n**, in a way that order does not matter (and where k, n are both non-negative integers, and k <= n) is:

The symbol on the left is referred to as the **binomial coefficient**, and is read “n choose k.”

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

Combinations

Here are some examples on how we can (and will) use the binomial coefficient.

- **How many ways can we flip a coin (whose flips are independent of one another) 4 times and see 2 heads?**
  - Equivalent question: how many different ways can we order the string “HHTT”?
  - There are 4 “positions.” Choose 2 of them to be H (the remaining 4 will be T).
  - This is **4 choose 2**, or 6. (Enumerated: HHTT, HTHT, HTTH, TTHH, THTH, THHT.)
- Suppose we have a bag of marbles that contains marbles, some of which are blue. **How many ways can we draw 7 marbles, such that 4 are blue and 3 are not blue?**
  - We have 7 draws. Choose 4 of them to be blue (the remaining automatically are not).
  - This is **7 choose 3**, or 35.
- Note: The above answers (and algebra) imply that $\binom{n}{k} = \binom{n}{n-k}$
  - Choosing k successes is equivalent to choosing n - k failures.

What about order?

You may be wondering – why did we use the **binomial coefficient** to determine the number of orderings of 2 heads and 2 tails (HHTT, HTHT, HTTH, TTHH, THTH, THHT), when the whole point was that **order doesn't matter**?

The order in which we declare the positions is what doesn't matter here. The contents of each position, though, do matter.

"first flip is heads"
"second flip is tails"
"third flip is tails"
"fourth flip is heads"

"third flip is tails"
"first flip is heads"
"fourth flip is heads"
"second flip is tails"

"first flip is heads"
"third flip is tails"
"fourth flip is heads"
"second flip is tails"

"second flip is tails"
"third flip is tails"
"first flip is heads"
"fourth flip is heads"

These are all equivalent! They all equate to the ordering **HTTH**.

The order in which you tell me what flip goes where is not of importance. This is why we use choosing here.

Another interpretation

How many ways can we flip a coin (whose flips are independent of one another) 4 times and see 2 heads?

- Equivalent problem to determining the number of rearrangements of HHTT.
- Let's label them uniquely: H1, H2, T1, T2.
- There are 4 objects, so there are 4! orders.
- But, some of these orderings are really the same!
  - "H1 H2 T1 T2" is really the same as "H2 H1 T2 T1" and "H1 H2 T2 T1."
  - These should all count as one "ordering."
  - There are 2! ways to arrange the two Hs amongst themselves.
  - There are 2! ways to arrange the two Ts amongst themselves.
- Dividing out the repetition yields, as we saw before,

$$\frac{4!}{2!2!}$$

Extension to multiple categories

How many ways can I select 7 marbles from a bag such that **4 are blue**, **2 are green**, and **1 is red**? (e.g. **bgbbbgr**, **bbbrgbg**, **brbgbgb**, etc.)

Again, there are two interpretations.

$$\binom{7}{4} \cdot \binom{3}{2} \cdot \binom{1}{1}$$

$$\frac{7!}{4!2!1!}$$

Extension to multiple categories

How many ways can I select 7 marbles from a bag such that **4 are blue**, **2 are green**, and **1 is red**? (e.g. **bgbbbgr**, **bbbrgbg**, **brbgbgb**, etc.)

Again, there are two interpretations.

$$\binom{7}{4} \cdot \binom{3}{2} \cdot \binom{1}{1}$$

$$\frac{7!}{4!2!1!}$$

7 open positions. Choose 4 to fill with blue marbles.

Of the 3 remaining positions, fill 2 with green marbles.

In the final position, put 1 red marble.

Extension to multiple categories

How many ways can I select 7 marbles from a bag such that **4 are blue**, **2 are green**, and **1 is red**? (e.g. **bgbbbgr**, **bbbrgbg**, **brbgbgb**, etc.)

Again, there are two interpretations.

$$\binom{7}{4} \cdot \binom{3}{2} \cdot \binom{1}{1}$$

$$\frac{7!}{4!2!1!}$$

We have 7 marbles total, which we can arrange in 7! ways. The 4 blues are identical, so divide by 4!, and similarly divide by 2! for the greens.

Extension to multiple categories

How many ways can I select 7 marbles from a bag such that **4 are blue**, **2 are green**, and **1 is red**? (e.g. **bgbbbgr**, **bbbrgbg**, **brbgbgb**, etc.)

Again, there are two interpretations.

$$\binom{7}{4} \cdot \binom{3}{2} \cdot \binom{1}{1} = \frac{7!}{4!2!1!}$$

Unsurprisingly, they are equal!

LECTURE 2

Data Sampling and Probability

Content credit: Fernando Pérez, Suraj Rampure, Ani Adhikari, Joseph Gonzalez, and Lisa Yan