Lec 2

Data Sampling and Probability

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

Meet our teaching team!

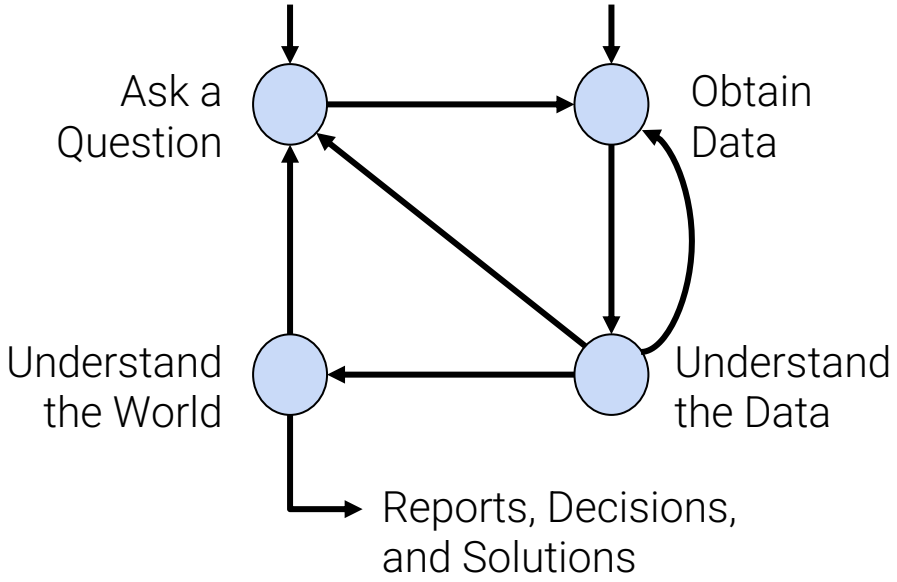
- **Name**: Boyuan Zhang
- **Major/Year**: ECE Senior
- **Email**: zhangboyuan-sgr@sjtu.edu.cn
- **Study of Interest**: Data Science, Statistics, Machine Learning
- Admitted to HKU, Master of Data Science

- **Related Course Studied**:
  - ECE4710J/STAT4710J
  - STAT1000J
  - STAT4060J
  - STAT4130J

- **Name:** Youchen Qing
- **Major/Year:** ECE Junior
- **Email:** rex29qyc@sjtu.edu.cn
- **Related Course Studied:**
  - ECE4710J/STAT4710J
  - STAT4060J
  - STAT4510J

sports enthusiast **interested in Data Science**

We call this the
Data Science Lifecycle.



Today's Roadmap

- Censuses and Surveys
- Samples
- Bias: A Case Study
- Probability Samples
- Multinomial and Binomial probabilities

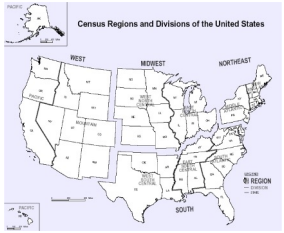
Censuses and Surveys

- **Censuses and Surveys**
- Samples
- Bias: A Case Study
- Probability Samples
- Multinomial and Binomial probabilities

China Population Census

中国人口普查
CHINA POPULATION CENSUS

The US Decennial Census



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

- Was held in Nov – Dec 2020. Released in May 2021.
- 31 provinces, autonomous regions and municipalities.
- Door-to-door collection
- Important uses:
  - Economy and social planning
  - Government policies.
  - investment in infrastructure and social welfare

- 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.

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.

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

Samples

- Censuses and Surveys
- **Samples**
- Bias: A Case Study
- Probability Samples
- Multinomial and Binomial probabilities

Sampling from a finite population

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

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.

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

Convenience samples

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

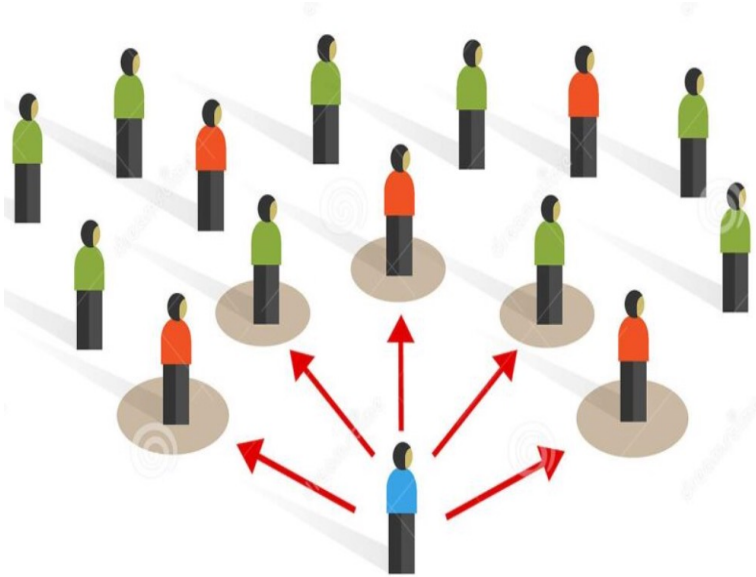
- Not a good idea for inference!
- Haphazard ≠ random.
- Sources of bias can introduce themselves in ways you may not think of!

Convenience samples are not random.

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.

Do you expect the weights of our sampled mice to be representative of all mice in our cage?

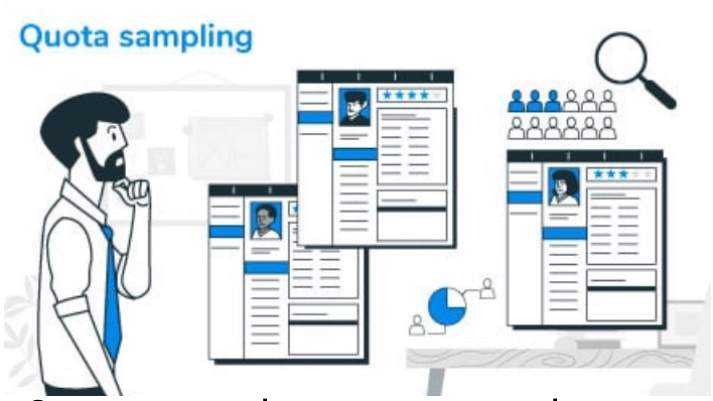
Convenience Samples



- Helps with Pilot testing:
  - handy when the researcher wants to get quick information.
  - allows quick data collection and analysis.
  - helps the researcher develop more questions for the actual study.
  - cost effective.

Quota samples

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



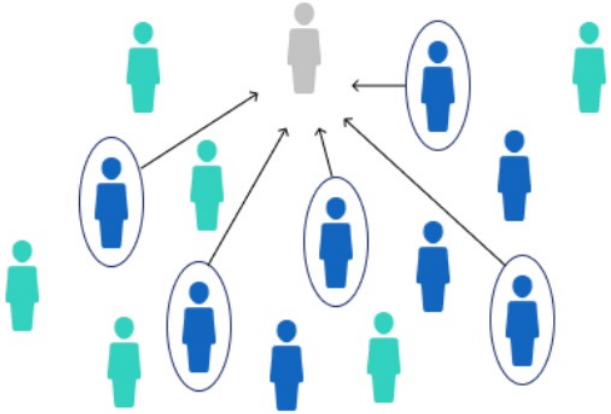
Quota samples are not random.

Issues with quota samples:

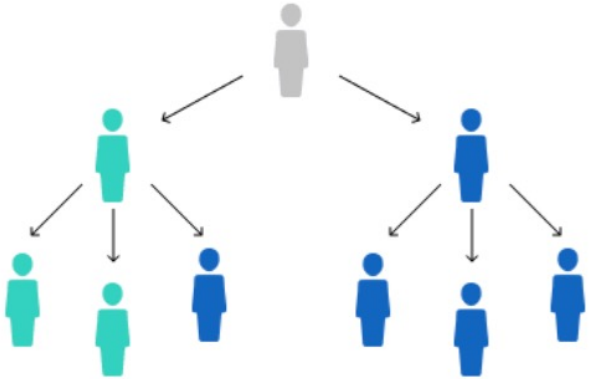
- Selection of quota is **not random**.
- By setting quotas, you require that your sample look like your population with regards to just a few aspects – but not all!
  - For example, if you set quotas for age, your sample might be representative of your population with regards to age.
  - What about gender? Ethnicity? Income?

Some other non-probability sampling methods

Voluntary response sample



Snowball sample



Population, sample, and sampling frame

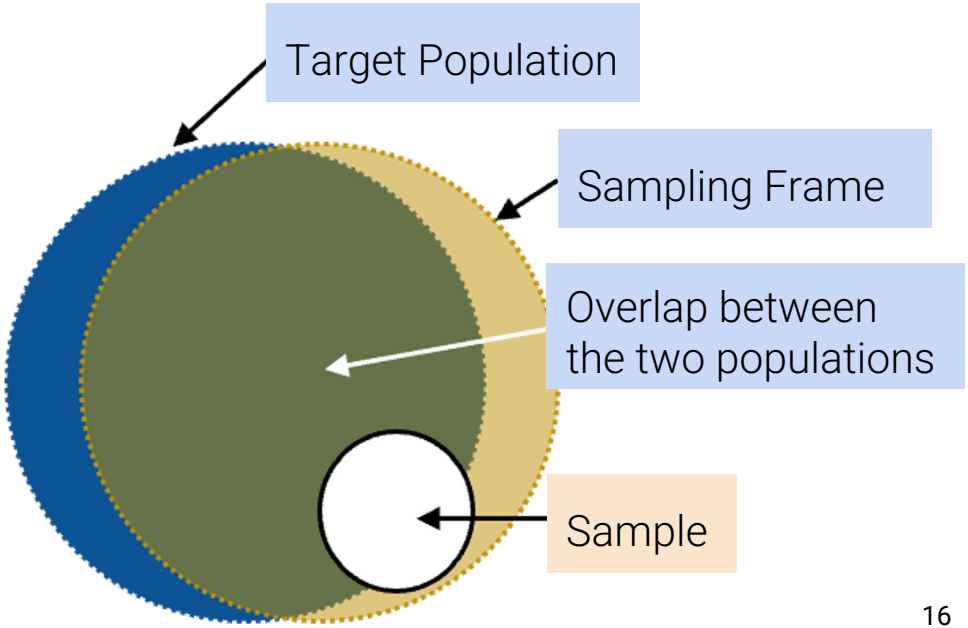
Population: The group that you want to learn something about.

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!

Quality, not quantity!

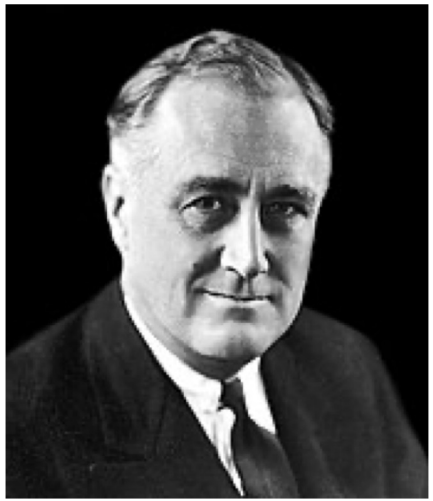
Try to ensure that the sample is representative of the population.

- 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!**

Bias: A Case Study

- Censuses and Surveys
- Samples
- **Bias: A Case Study**
- Probability Samples
- Multinomial and Binomial probabilities

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?

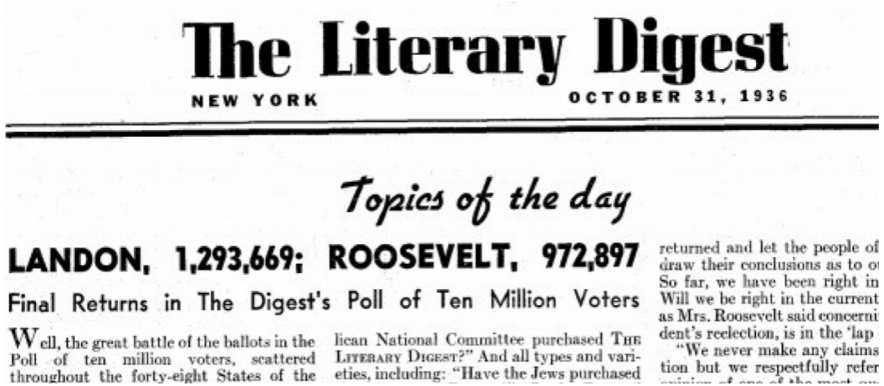
(1) The Literary Digest sample was **not representative** of the population.

- 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?

Bias

https://catalogofbias.org/biases/

Summary of results

	% Roosevelt	# surveyed
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
Actual election	61%	All voters

Big samples aren't always good!

- What you need is a representative sample.
- If your sampling method is biased, those biases will be magnified with a larger sample size.

Probability Samples

- Censuses and Surveys
- Samples
- Bias: A Case Study
- **Probability Samples**
- Multinomial and Binomial probabilities

Probability sampling

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

- Random samples **can** produce biased estimates of population characteristics.
- But with random samples we are able to estimate the bias and chance error.
    - We can quantify the uncertainty.

For our purposes, **probability samples** and **random samples** will mean the same thing.

A probability sample is a **type of sampling technique**.

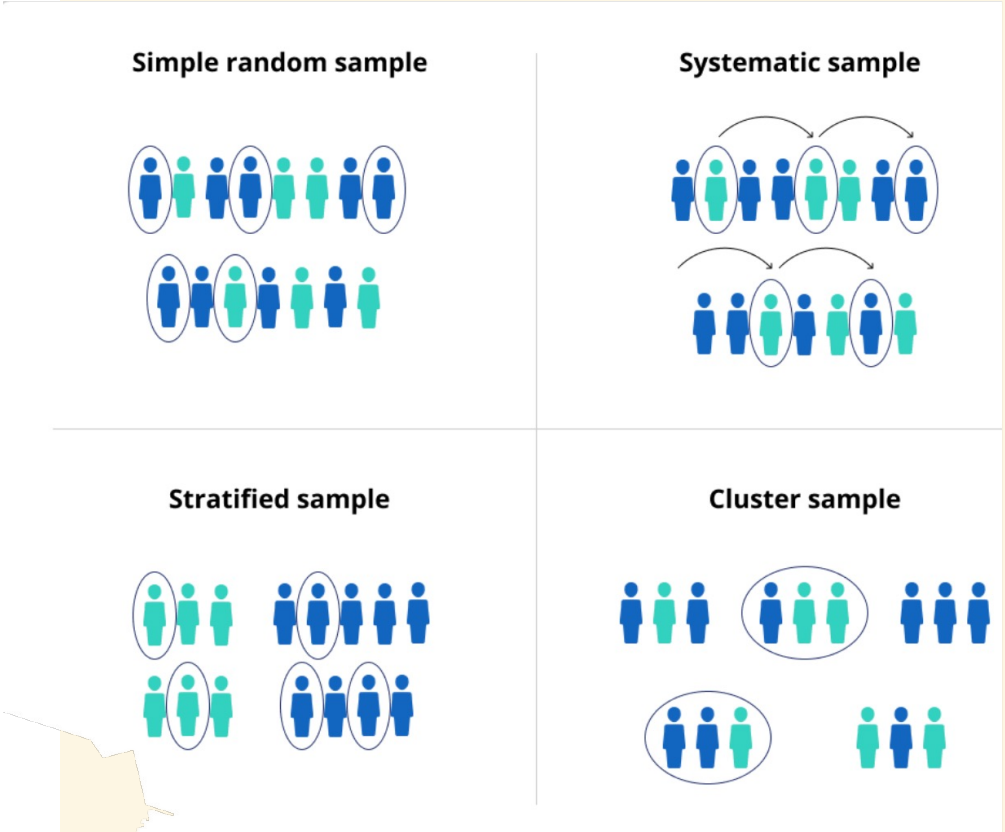
Chance Error

- Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error.**
- Drawing another sample will result in a **different** chance error.
- Estimate = Parameter + Bias + Chance error

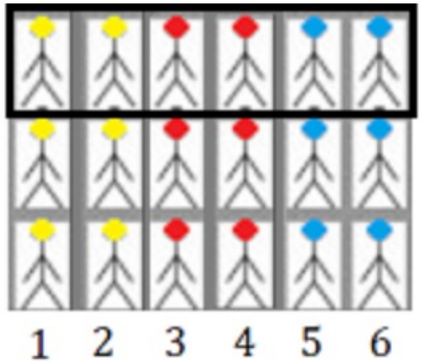
- The chance error will get smaller as the sample size get larger, but it is unavoidable.
- This is not the case for bias: Increasing the sample size just repeats the bias on a large scale.

Probability Sampling examples

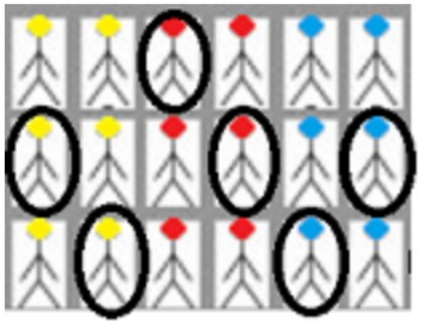
- Example
  - You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.
  - The company has offices in 10 cities across the country (all with roughly the same number of employees in similar roles). You don't have the capacity to travel to every office to collect your data, so you use random sampling to select 3 offices – these are your clusters



Quota sampling Vs Stratified sampling



Quota Sampling



Stratified Sampling

- With quota sampling, **random sampling methods are not used (called "non probability" sampling).**
- With stratified sampling, you **use a random sampling method**

Probability sampling

In order for a sample to be a probability sample:

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

Not all probability samples are necessarily good.

For instance, suppose I have three students: Allen, Ken, John, and I want to sample two of them.

- I choose Allen with probability 1.
- I choose either Ken or John, each with probability ½.

This is a probability sample (but it's not great).

Does it have chance error?

Does it have bias?

Some 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** is a sample drawn **uniformly** at random **without** replacement.

Every individual (and subset of individuals) has the same chance of bei

- Every individual 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 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!

Example scenario

Consider the following sampling scheme:

- Suppose a class roster has 100 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, etc).

Is this a probability sample?

Does each student have the same probability of being selected?

Is this a simple random sample?

Example scenario

Consider the following sampling scheme:

- Suppose a class roster has 100 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, etc).

Is this a probability sample?

- Yes. If my sample is [n, n + 10, n + 20, …, n + 90], where 1 <= n <= 10, the probability of that sample is 1/10.
- Otherwise, the probability is 0.
- Only 10 possible samples!

Does each student have the same probability of being selected?

Is this a simple random sample?

Example scenario

Consider the following sampling scheme:

- Suppose a class roster has 100 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, etc).

Is this a probability sample?

- Yes. If my sample is [n, n + 10, n + 20, ..., n + 90], where 1 <= n <= 10, the probability of that sample is 1/10.
- Otherwise, the probability is 0.
- Only 10 possible samples!

Does each student have the same probability of being selected?

- Yes. Each student is chosen with probability 1/10.

Is this a simple random sample?

Example scenario

Consider the following sampling scheme:

- Suppose a class roster has 100 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, etc).

Is this a probability sample?

- Yes. If my sample is [n, n + 10, n + 20, …, n + 90], where 0 <= n <= 10, the probability of that sample is 1/10.
- Otherwise, the probability is 0.
- Only 10 possible samples!

Does each student have the same probability of being selected?

- Yes. Each student is chosen with probability 1/10.

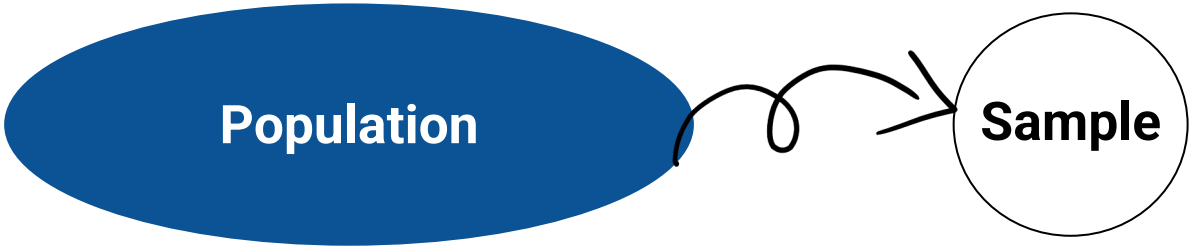
Is this a simple random sample?

- No. The chance of selecting (8, 18) is 1/10; the chance of selecting (8, 9) is 0.

Multinomial and Binomial probabilities

- Censuses and Surveys
- Samples
- Bias: A Case Study
- Probability Samples
- **Multinomial and Binomial probabilities**

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
- We'll derive the multinomial probabilities in this section as a review of probability.

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, since the sample is drawn with replacement:

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

=

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**) = … = $(0.6)^4(0.4)^3$.

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

and thus,

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

$$= \underbrace{\binom{7}{4}}_{\text{\# of ways}}\underbrace{(0.6)^4(0.4)^3}_{\text{ordered series}}$$

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.

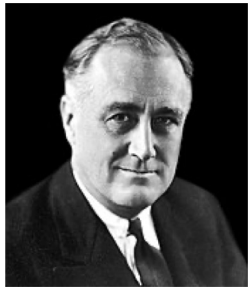
Then, the **multinomial probability** of drawing $\mathbf{k_1}$ individuals from Category 1, $\mathbf{k_2}$ individuals from Category 2, and $\mathbf{k_3}$ individuals from Category 3 (where $k_1 + k_2 + k_3 = n$) is

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

Revisit the “Literary Digest”

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 **Hypothesis Testing**.

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

Mark-Recapture Method

In the simplest case, a one-stage mark-recapture study produces the following data

- M : number of animals marked in first capture
- C : number animals in second capture
- R : number of marked animals in second capture.
- We are interested in N : number of animals in the population

$$\widehat{N} = \frac{MC}{R}$$

This population estimate would arise from a probabilistic model in which the number of recaptured animals is distributed binomially

- R ~ Binomial(C, p), where p = M/N (prerequisite: N is large, M/N > 0.1)