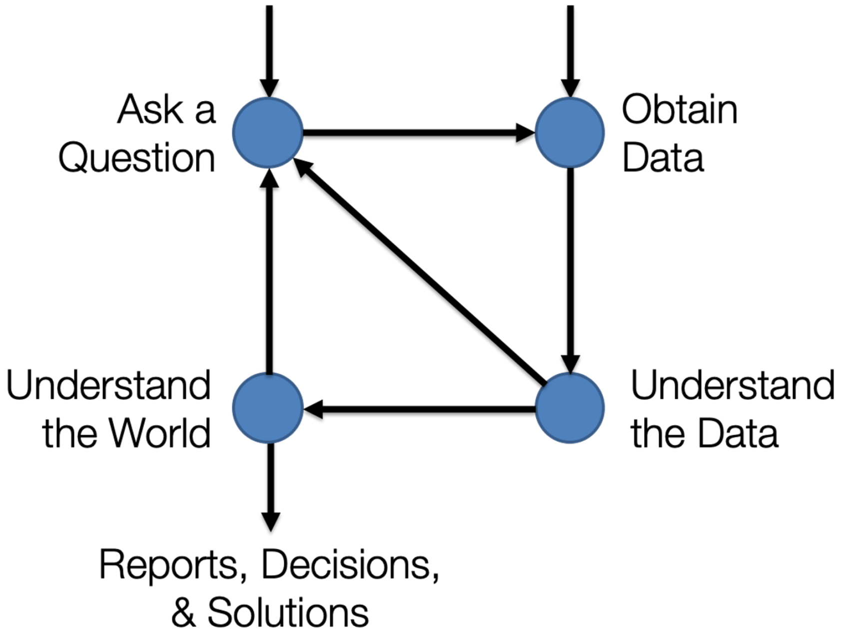
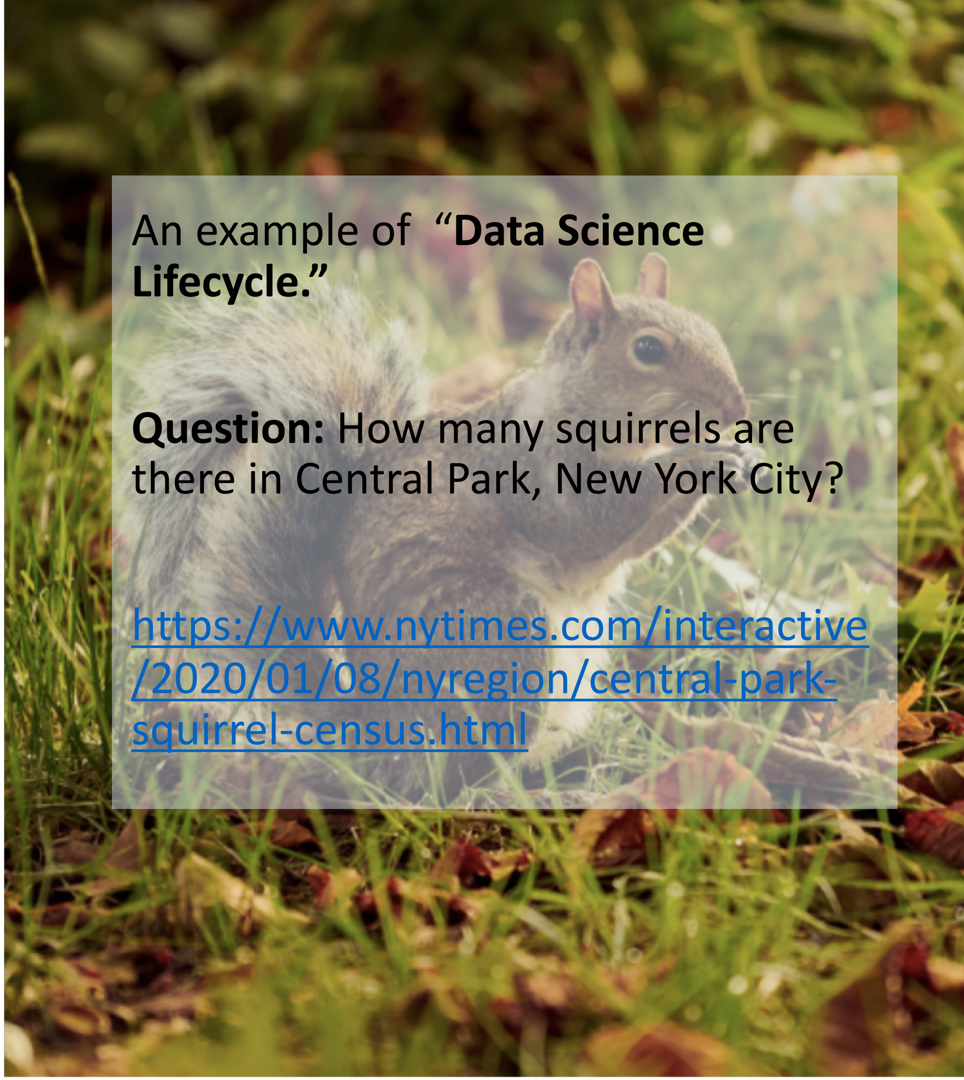
Data Sampling and Probability

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

HW 1 will be posted today on Canvas. (Due 2/22)

ECE 4710J will be offered in 22 Summer for the ”JiaoTong Global Classroom” Program



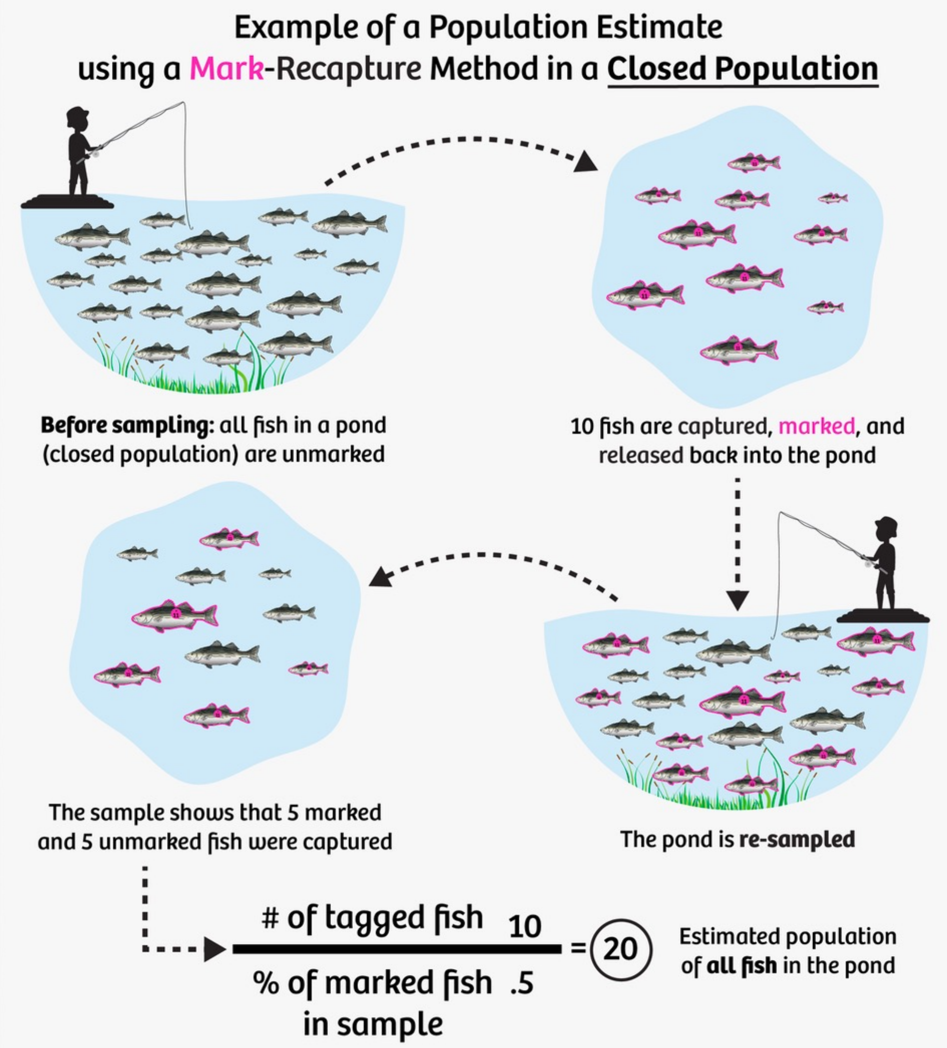


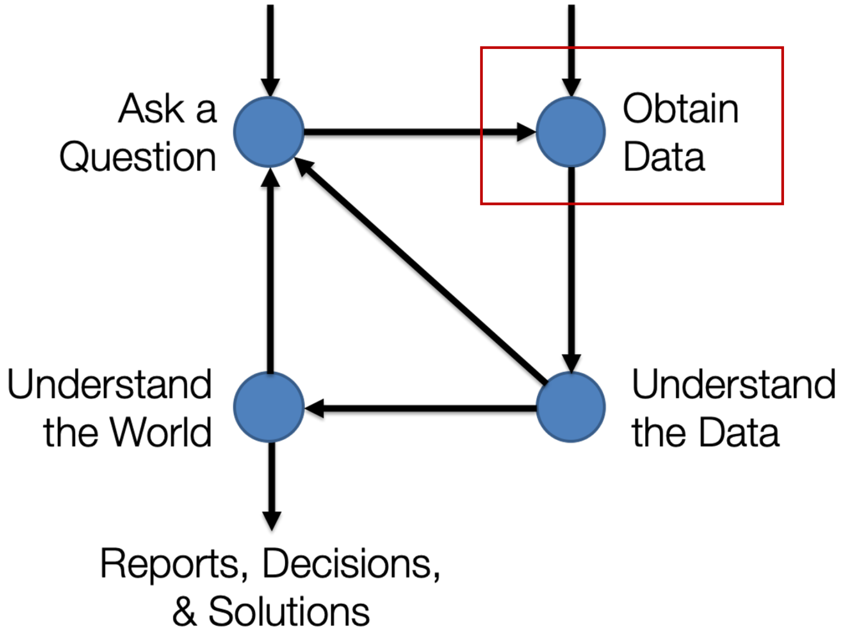
An example of “**Data Science Lifecycle.**”

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

https://www.nytimes.com/interactive/2020/01/08/nyregion/central-park-squirrel-census.html

Mark-Recapture Method





Censuses and Surveys

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.

Surveys

- 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

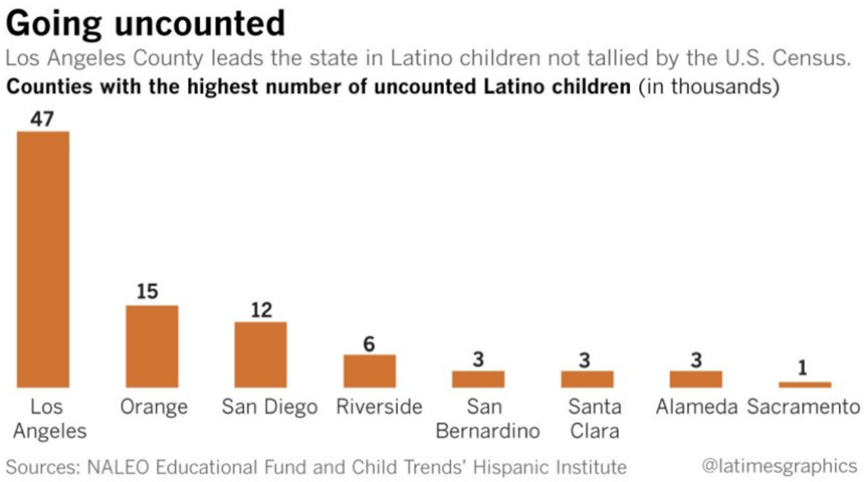
By Amelia Thomson-DeVeaux

NATIONAL

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

August 9, 2019 · 3:23 PM ET

Issues with the US Decennial Census



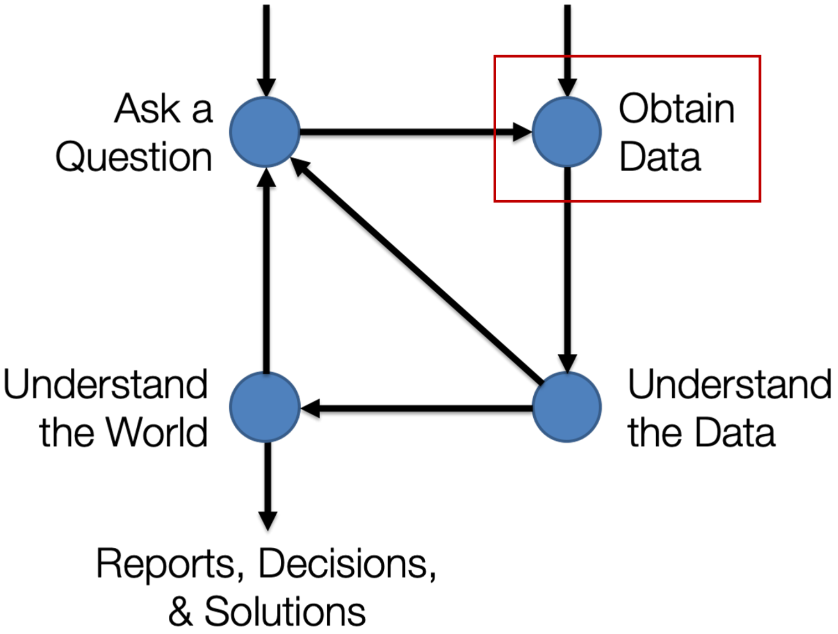
How do we know these numbers?
From other surveys.

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?

High Court Rejects Sampling In Census

Ruling Has Political, Economic Impacts


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

Samples

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.

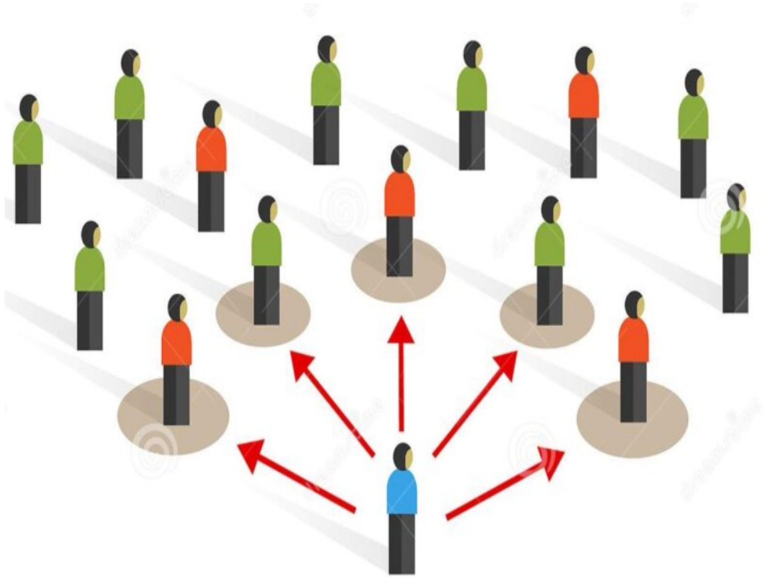
- Not a good idea for inference!
- Haphazard ≠ random.
- Sources of bias can introduce themselves in ways you may not think of!

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?

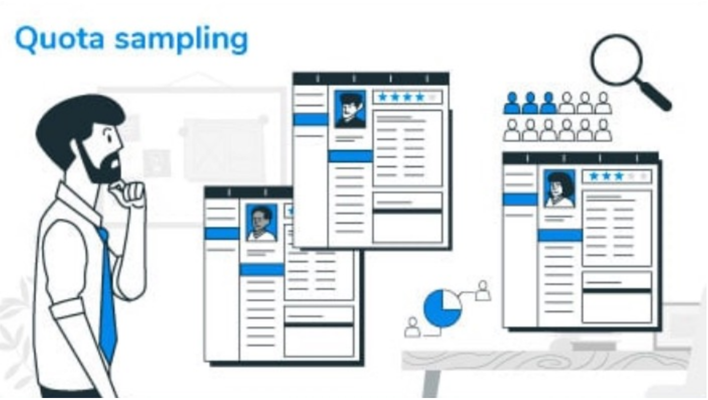
Convenience Samples



- 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.



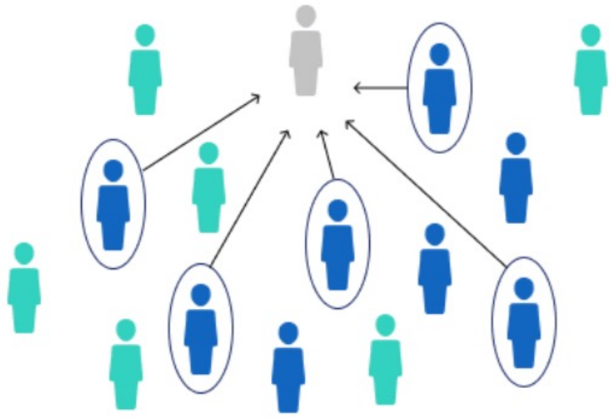
Quota samples are not random.

Issues with quota samples:

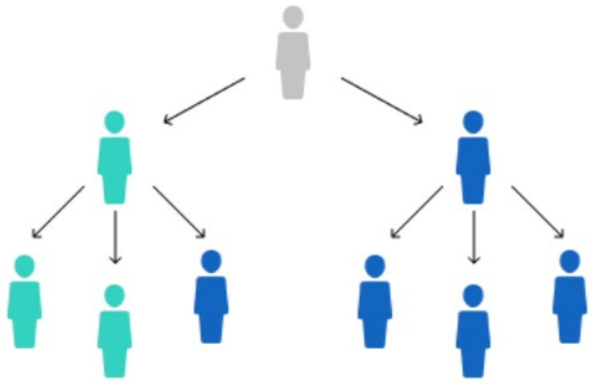
- Selection of quota is **not random**.
- By setting quotas, you require that your sample look like your population with regards to just a few aspects – but not all!
    - For example, if you set quotas for age, your sample might be representative of your population with regards to age.
    - What about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

Some other non-probability sampling methods

Voluntary response sample



Snowball sample



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

https://catalogofbias.org/biases/

Common Biases

Selection Bias

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

Response Bias

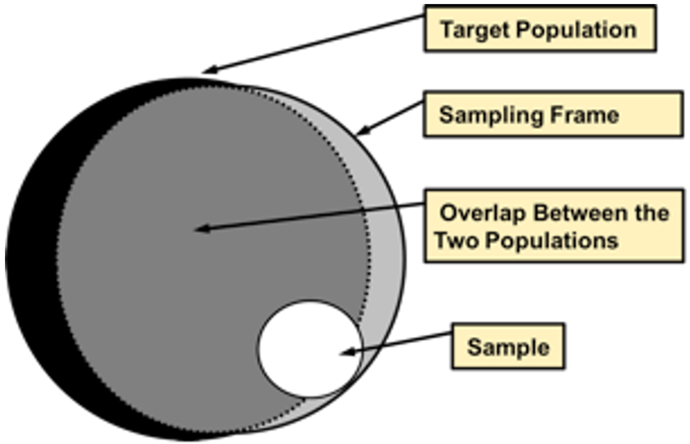
- People don’t always respond truthfully. (Reviews of products)
- How to avoid: Examine the nature of questions and the method of surveying.

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!

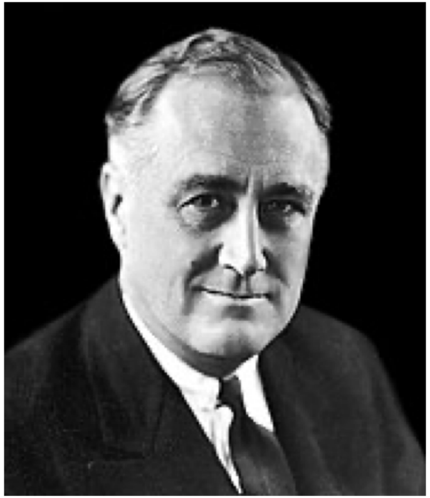
Population, samples, and sampling frame

- **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.



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

Case study – 1936 Presidential Election



Roosevelt (D)

Landon (R)

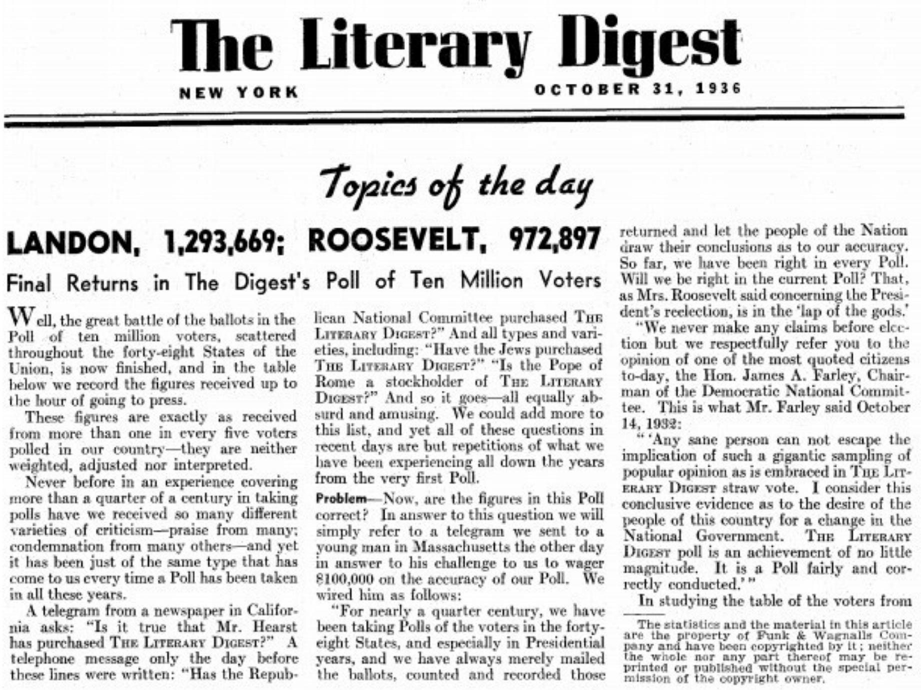
In 1936, President Franklin D. Roosevelt (left) went up for re-election against Alf Landon (right). As is usual, polls were conducted in the months leading up to the election to try and predict the outcome.

The Literary Digest

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.



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 Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased THE LITERARY DIGEST?" A telephone message only the day before these lines were written: "Has the Republican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased THE LITERARY DIGEST?" "Is the Pope of Rome a stockholder of THE LITERARY DIGEST?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager $100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"'Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in THE LITERARY DIGEST straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. THE LITERARY DIGEST poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted.'"

In studying the table of the voters from

The Literary Digest

The Literary Digest's **prediction**:

43% Roosevelt, 57% Landon

The **actual** outcome of the election:

61% Roosevelt, 37% Landon

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

- Their sample was **not representative** of the population.
    - They sampled people who owned phones, subscribed to magazines, and went to country clubs, who at the time were more affluent.
    - These people tended to vote Republican (Alf Landon).
- Only 2.4 million people **actually filled out the survey!**
    - 24% response rate (low).
    - Who knows how the other 76% would have polled?

Gallup’s Poll

George Gallup, a rising statistician, also made predictions about the impending 1936 elections. He predicted that Roosevelt would win with **56% of the vote**.

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%. **How was he able to predict what they were going to predict, with such accuracy?**

- He predicted that they would survey people in the phone book, people who subscribed to magazines, and who were part of country clubs.
- So he sampled those same individuals!
- He was able to predict their prediction by sampling only 3000 people.

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 (Random) Samples

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

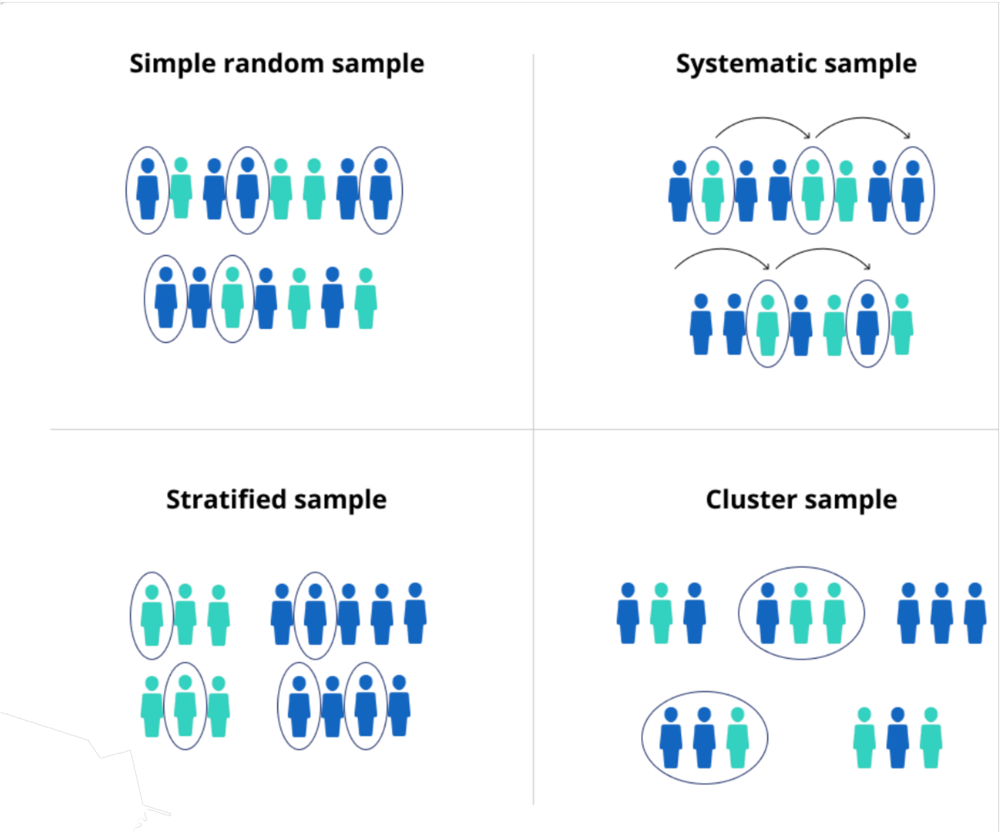
- Since the sample is drawn at random, the estimate will be different from the parameter due to **chance error.**
- Drawing another sample will result in a **different** chance error.
- Estimate = Parameter + Bias + Chance error

- The chance error will get smaller as the sample size get larger, but it is unavoidable.

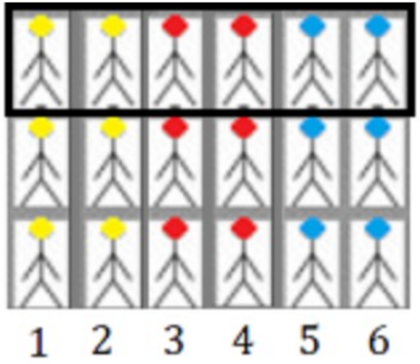
- This is not the case for bias: Increasing the sample size just repeats the bias on a large scale.

Probability Sampling examples

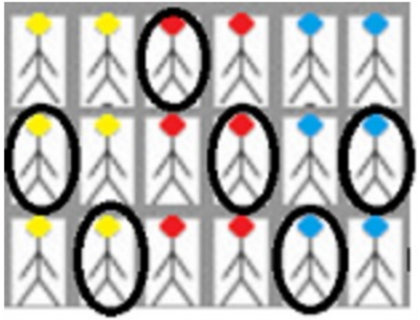
- **Example**
  - You want to select a simple random sample of 100 employees of Company X. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.
  - The company has offices in 10 cities across the country (all with roughly 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 being selected.

- 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

- 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**.
  - For instance, if our population size is in the thousands, and we’re sampling 100 people, removing those 100 doesn’t change the population very much.
- **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 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?

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?

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.

Binomial and multinomial probabilities

Given the population distribution, we can compute the probability of us getting a particular 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.
    - For example, a bag of marbles in which 60% are **blue** and 40% are **not** blue.
    - Or where 60% are **blue**, 30% are **green**, and 10% are **red**.
- Want to count the number of each category that end up in our sample.

Binomial probabilities

Suppose we sample at random with replacement 7 times from a bag of marbles, 60% of which are **blue** and 40% of which are **not** blue.
"4 blue, 3 not blue" can occur in several equally likely ways.
For instance, P(**bnbbbnn**) = P(**bnbbbnn**) = P(**bnbbbnn**) = … = $(0.6)^4(0.4)^3$.

P(4 blue, 3 not blue) is the **total** chance of all of those ways. The number of ways in which we can draw 4 blue marbles and 3 not blue marbles is

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

and thus,

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

Generalization of binomial probabilities

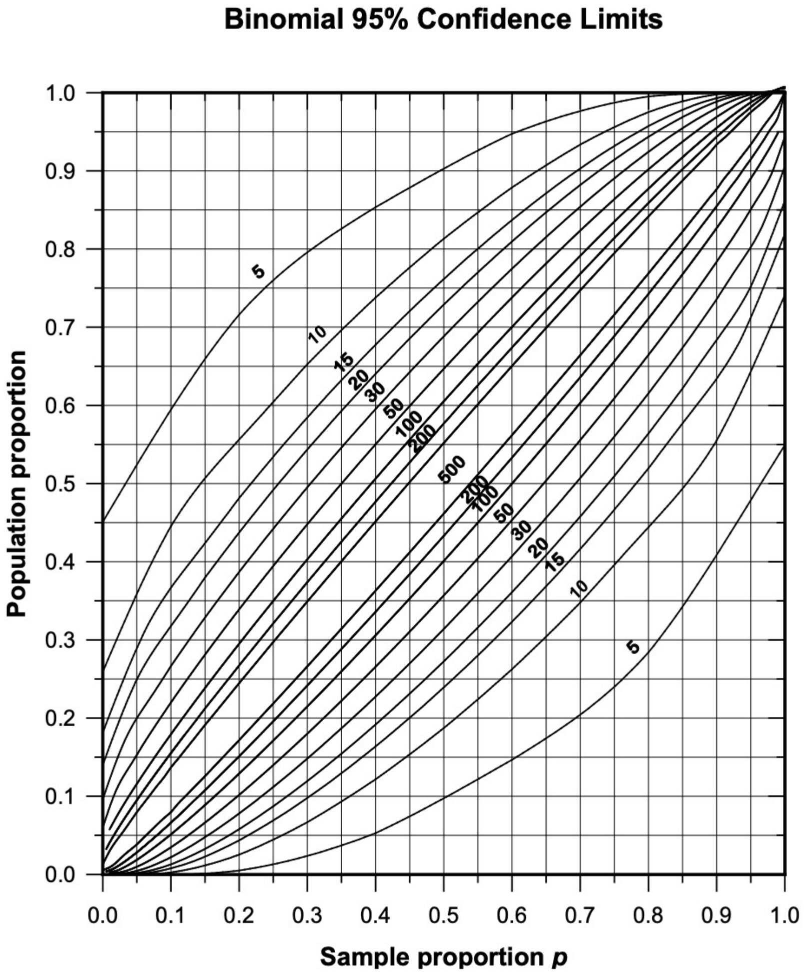
If we are drawing at random with replacement **n** times, from a population in which a proportion **p** of the individuals are called "successes" (and the remaining **1 - p** are "failures"), then the probability of **k successes** (and hence, **n - k failures**) is

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$$

- Note: since the number of successes and the number of failures we draw must sum to n, saying "k successes" is equivalent to saying "k successes and n - k failures."
  - For instance – if you flip a coin 10 times and see 6 heads, we know there were 4 tails.
- This is essentially the **binomial distribution**.

Revisit “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)



Revisit “The Literary Digest’s result”