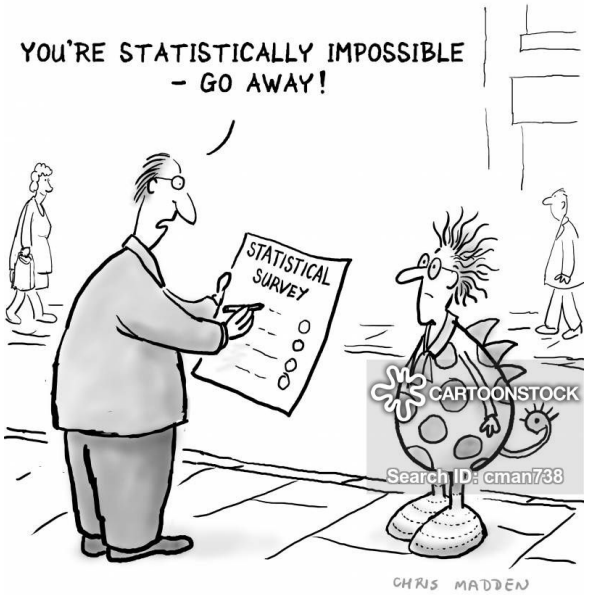
COSE471 Data Science

Probability and Data Design

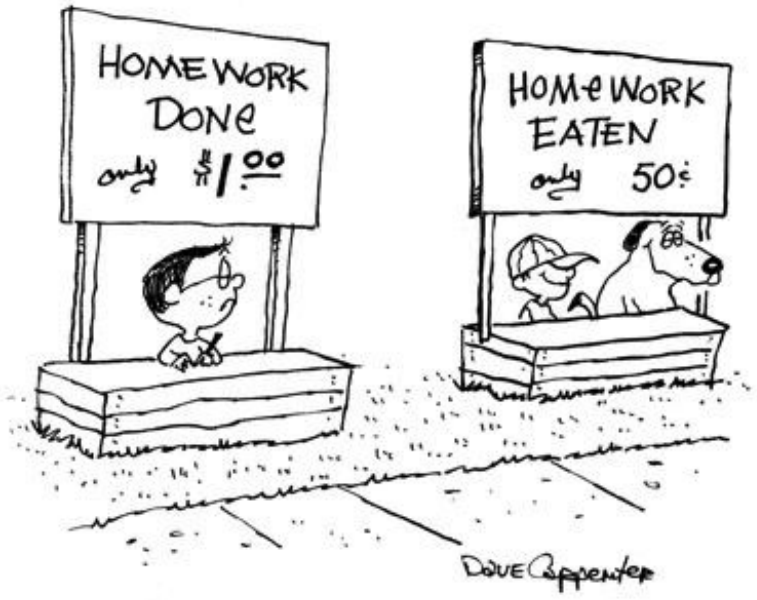


Dept. of Computer Science and Engineering

Korea University

Announcements

- Homework #1 is out today.
  - Due on March 19 (Fri) 11:59pm



Announcements

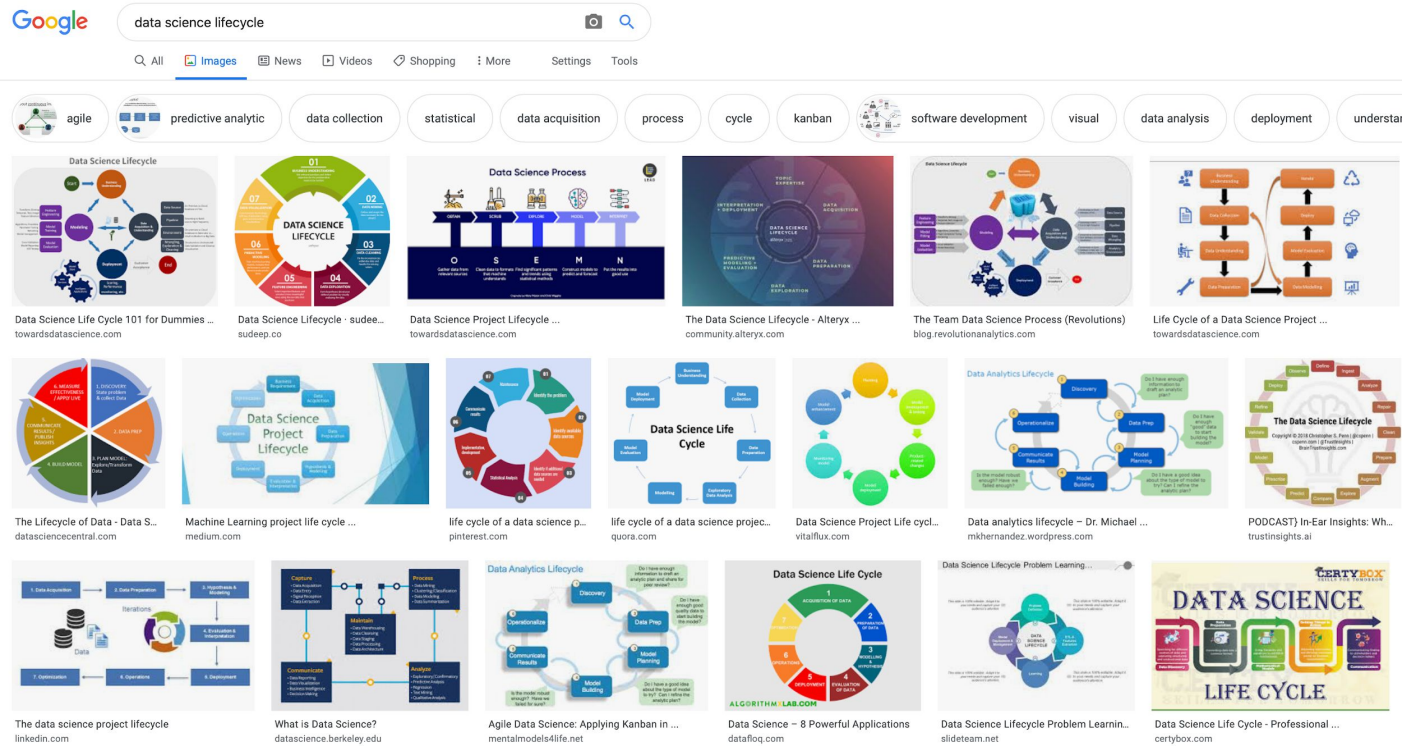
- Homework #1 is out today.
  - Due on March 19 (Fri) 11:59pm
- An invitation from Gradescope is sent out yesterday.
  - If not, let me know!
- Office hour sign-up sheet is posted in Blackboard
- Dory:
  - https://www.dory.app/c/korea.ac.kr/8898bb8b_cose471

Agenda

Welcome to the first content of COSE471! Here we plan on:

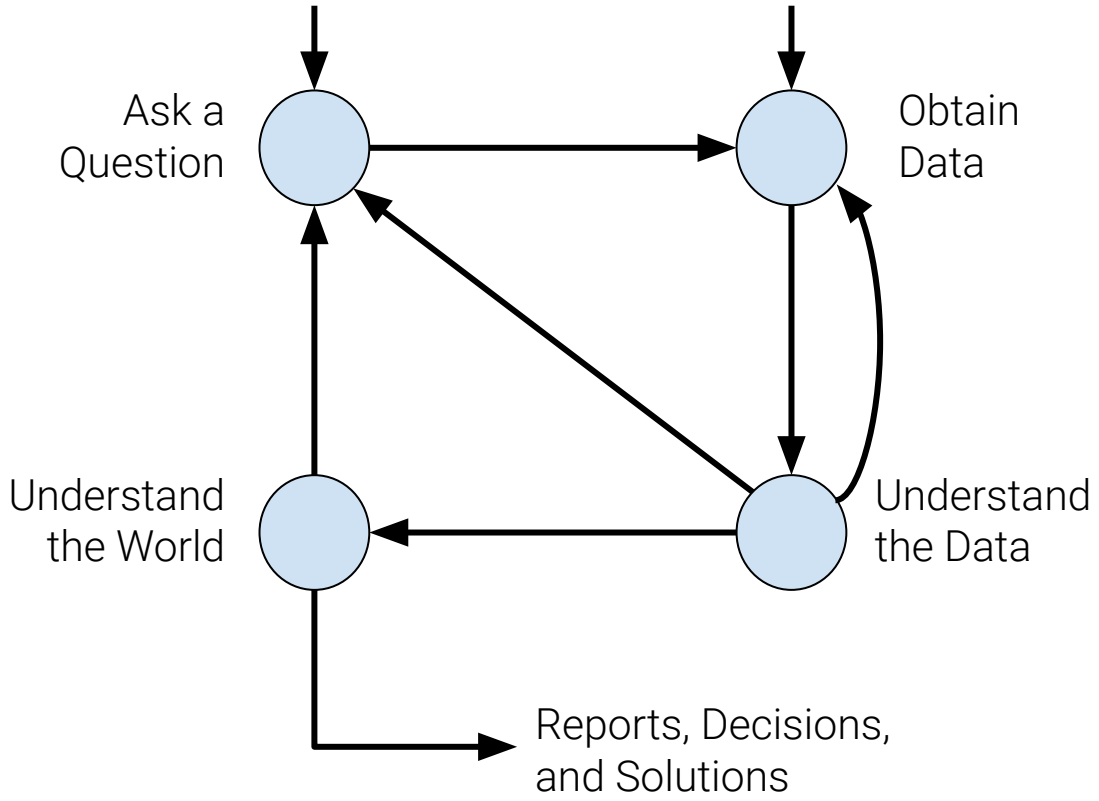
- Data Science Lifecycle
- Census, Survey, and Sampling
    - Why we need to sample in the first place.
    - What it means for our sample to be biased.
    - How to prevent these biases in our samples.
    - What exactly a sampling frame is, and why choosing a good one is important.
- Designed Experiments

Data Science Lifecycle



The “data science lifecycle” you will see out in the wild may be slightly different than the one we teach you, but the core ideas are all the same.

Data science lifecycle

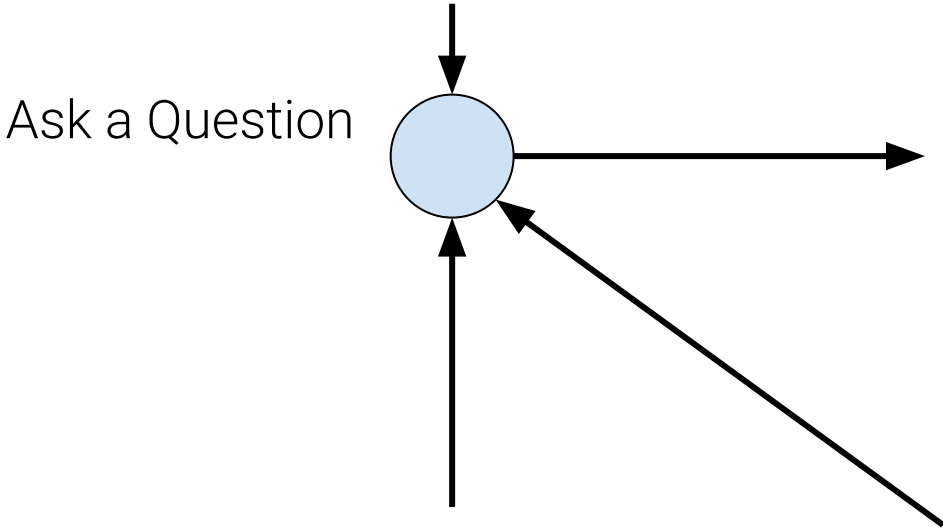


The data science lifecycle is a **high-level description** of the data science workflow.

Note the two distinct entry points!

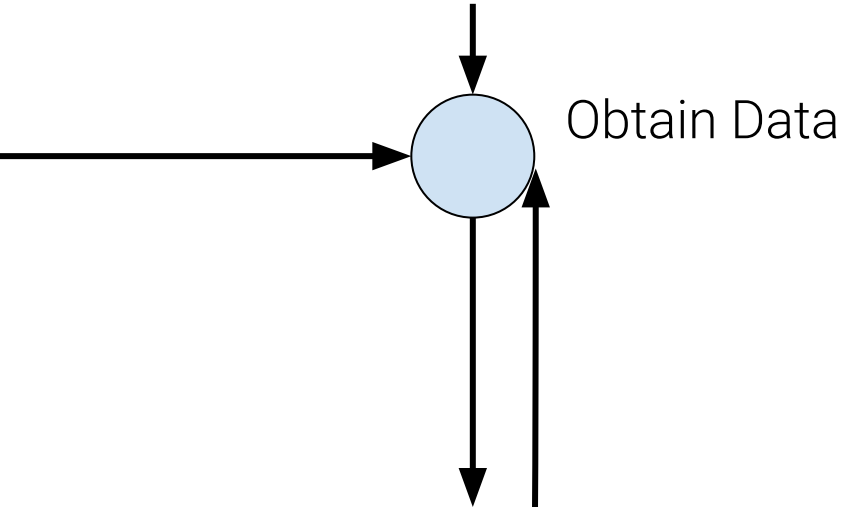
1. Question/Problem Formulation

- What do we want to know?
- What problems are we trying to solve?
- What are the hypotheses we want to test?
- What are our metrics for success?

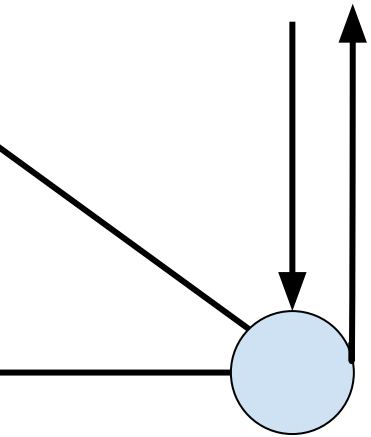


2. Data Acquisition and Cleaning

- What data do we have and what data do we need?
- How will we sample more data?
- Is our data representative of the population we want to study?



3. Exploratory Data Analysis & Visualization

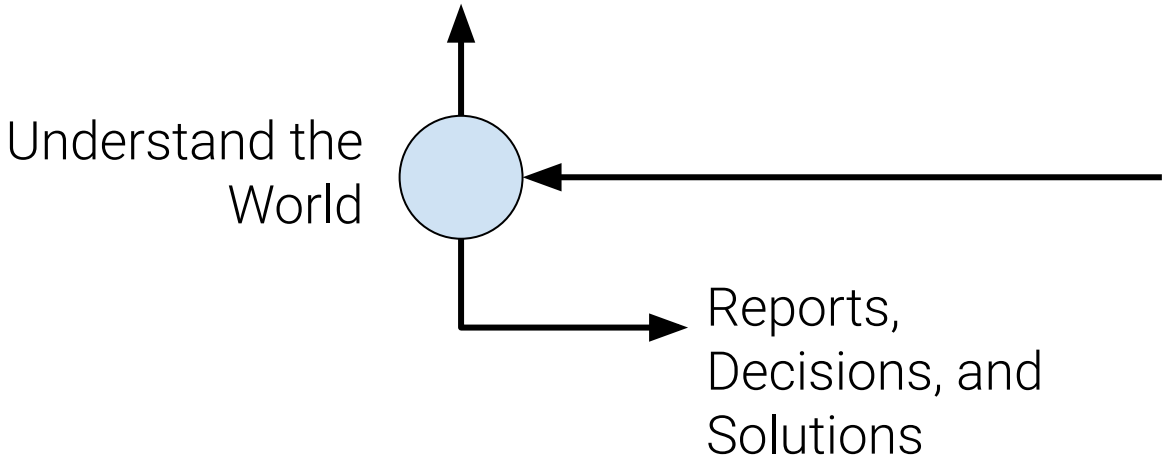


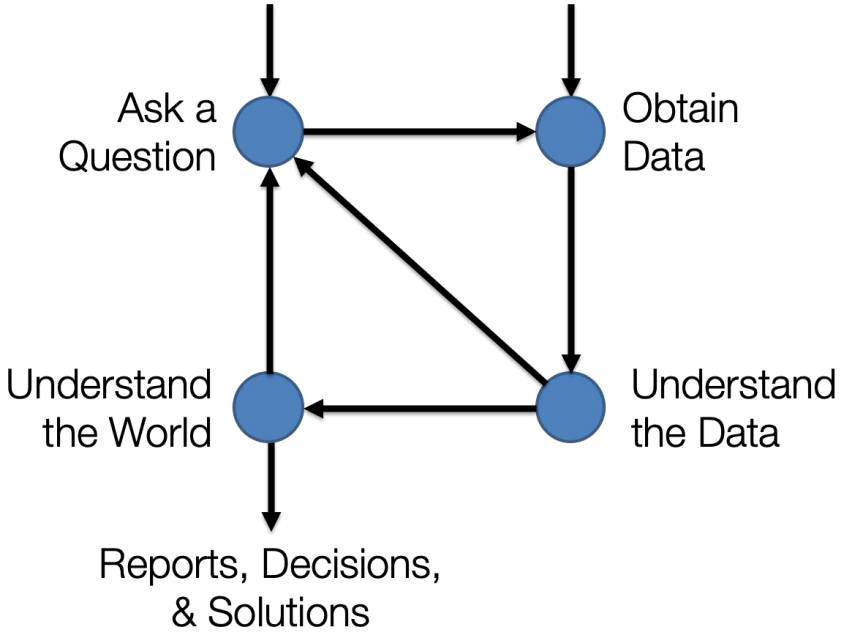
Understand the Data

- How is our data organized and what does it contain?
- Do we already have relevant data?
- What are the biases, anomalies, or other issues with the data?
- How do we transform the data to enable effective analysis?

4. Prediction and Inference

- What does the data say about the world?
- Does it answer our questions or accurately solve the problem?
- How robust are our conclusions and can we trust the predictions?





We call this the

Data Science Lifecycle.

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

https://youtu.be/CEnrGuRE0jY

The New York Times

Why Count All the Squirrels in Central Park? Why the Heck Not

The team behind the park's census of Eastern grays say an accurate tally is possible despite the critters', well, squirrelish ways.

By Andy Newman

Oct. 6, 2018

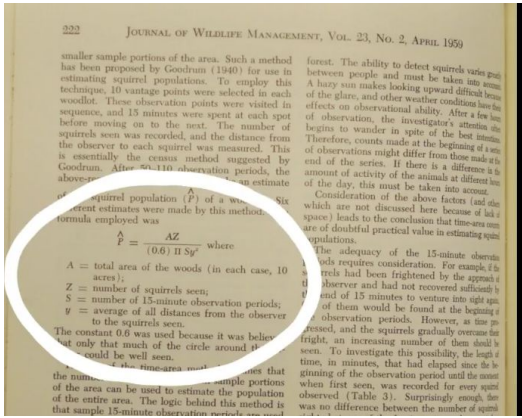
The process as reported by Denise Li (Jan 8, 2020).

The question was posed by The Squirrel Census, an independent effort that started when Jamie Allen, a writer, discovered there was no official count.

“I thought that was funny because we do keep track of other animals,” Allen said. “Somehow, the Eastern gray squirrel, being so common, had slipped under the radar.”

To interpret raw data, we model

Out in the field, I was assigned an area to walk about and search for 20 minutes. When I saw a squirrel, I plotted its location on a detailed map.



222 JOURNAL OF WILDLIFE MANAGEMENT, VOL. 23, No. 2, APRIL 1959

smaller sample portions of the area. Such a method has been proposed by Goodrum (1940) for use in estimating squirrel populations. To employ this technique, 10 vantage points were selected in each woodlot. These observation points were visited in sequence, and 15 minutes were spent at each spot before moving on to the next. The number of squirrels seen was recorded, and the distance from the observer to each squirrel was measured. This is essentially the census method suggested by 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.

To account for repeat sightings, the census team used a formula from a 1959 squirrel study to adjust the data and come up with a final count.

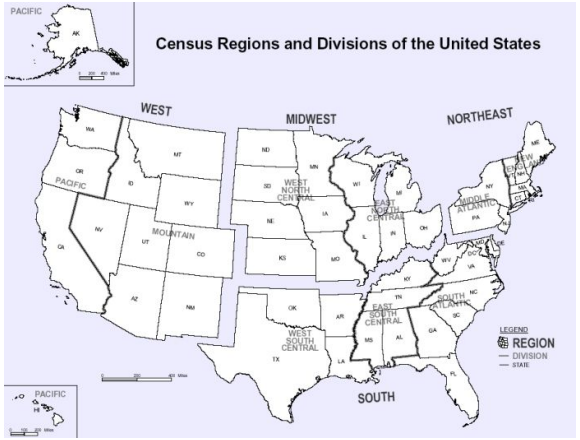
I looked up the formula later at the library.

JOURNAL OF WILDLIFE MANAGEMENT, VOL. 23, No. 2, APRIL 1959
https://www-jstor-org.libproxy.berkeley.edu/stable/pdf/3797645.pdf

Censuses and Surveys

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.



data.census.gov

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

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.

There are entire courses on surveying!

- See Stat 311 at KU.

FiveThirtyEight

Politics Sports Science & Health Economics Culture

JUN. 27, 2019, AT 12:42 PM

The Supreme Court Stopped The Census Citizenship Question — For Now

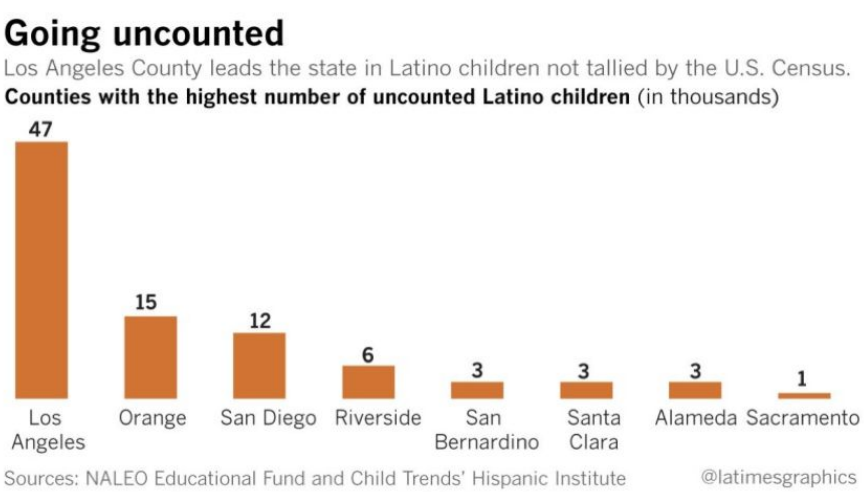
By Amelia Thomson-DeVeaux

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

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.

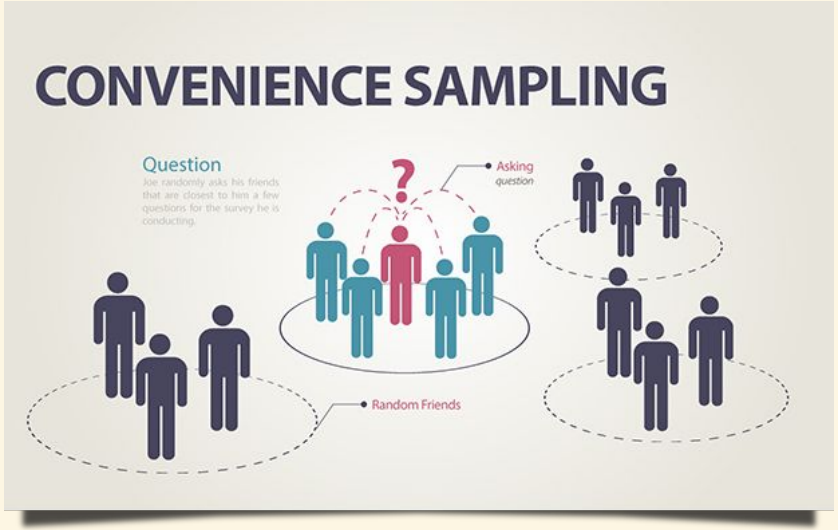
We will now look at some types of **non-random** samples, before formalizing what it means for a sample to be random.

Convenience samples

A **convenience sample** is whomever/whatever is convenient for investigator

- Sources of bias can introduce themselves in ways you may not think of!

Convenience samples are not random.



Other non-random samples

A **Self-selected sample** is whoever choose to answer.

A **Judgment sample** is whomever/whatever investigator deliberately selects

Convenience samples

A **convenience sample** is whomever/whatever is convenient for investigator

- Sources of bias can introduce themselves in ways you may not think of!

Convenience samples are not random.

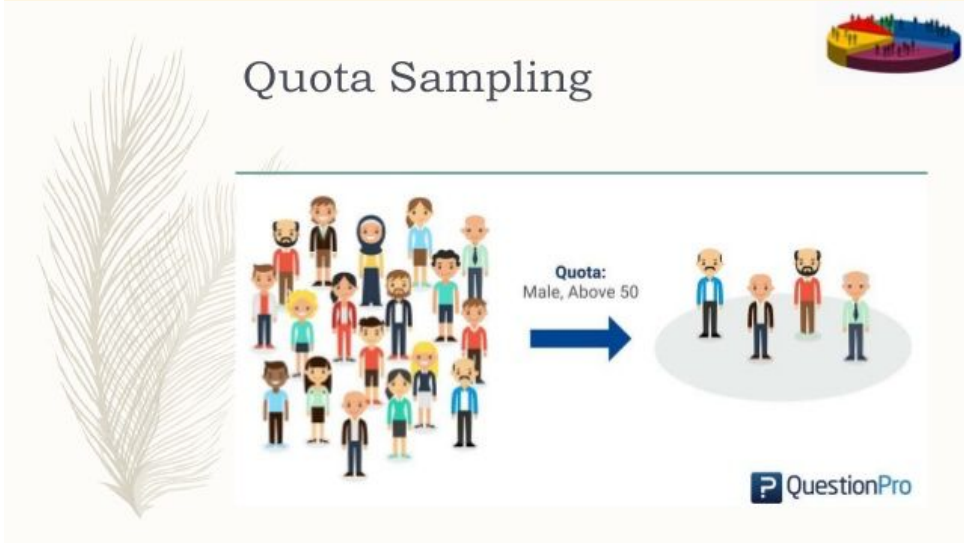
Example: Suppose we have a cage of mice, and each week, we want to measure the weights of these mice. To do so, we take a convenience sample of these mice, and weigh them.

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

Quota samples

Quota sampling is a procedure

- that restricts the selection of the sample by controlling the number of respondents by one or more criterion
- For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.



Quota samples

Quota sampling is a procedure

- that restricts the selection of the sample by controlling the number of respondents by one or more criterion
- For example, an interviewer may be told to sample 200 females and 300 males between the age of 45 and 60.

Quota samples are non-random

- For example, interviewers might be tempted to interview those who look most helpful.
- May be biased because not everyone gets a chance of selection.

Quality, not quantity!

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

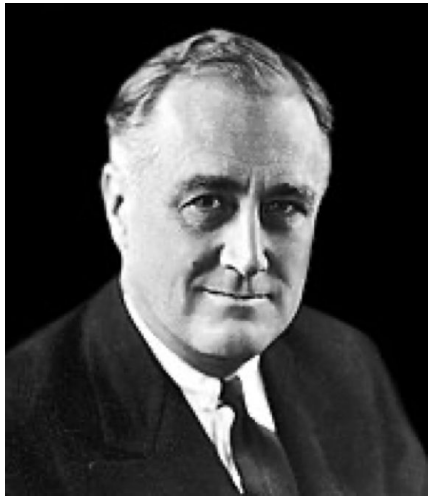
- Don't just try to get a big sample.
- If your method of sampling is bad, and your sample is big, you will have a **big, bad sample!**

This is a phenomenon you will explore in-depth in Homework 1, where you will perform an analysis of the 2016 US Presidential Elections.

Big Bad Wolf

Sampling Bias

Case study – 1936 US Presidential Election



Roosevelt (D)

Landon (R)

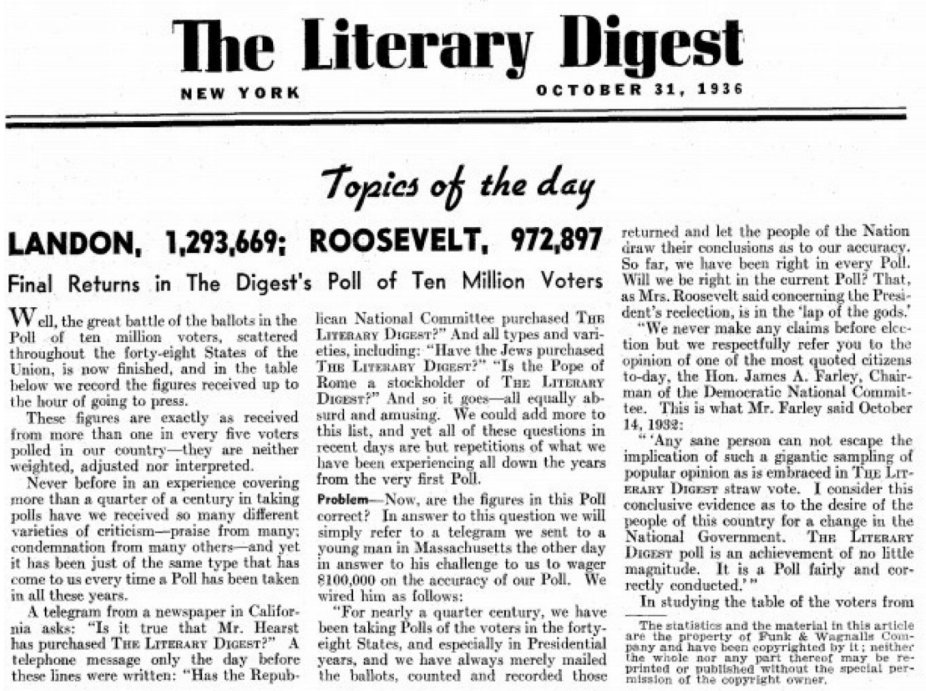
In 1936, President Franklin D. Roosevelt (left) went up for re-election against Alf Landon (right). As is usual, polls were conducted in the months leading up to the election to try and predict the outcome.

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!

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?

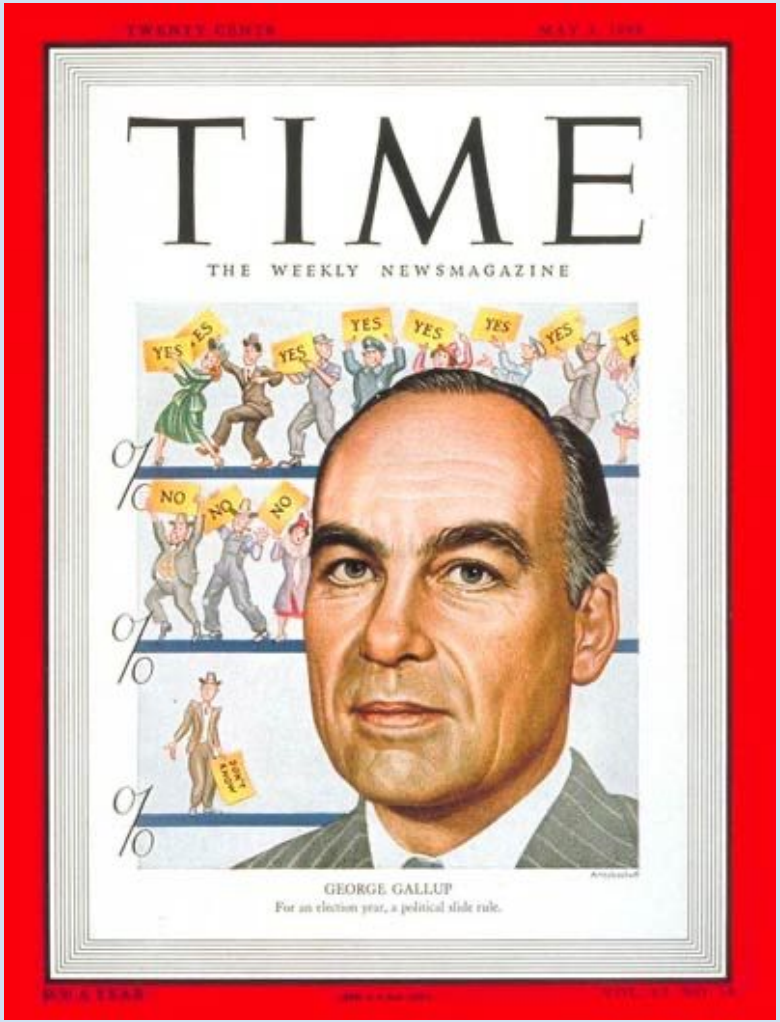
They surveyed 10 million people!

- 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**!



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

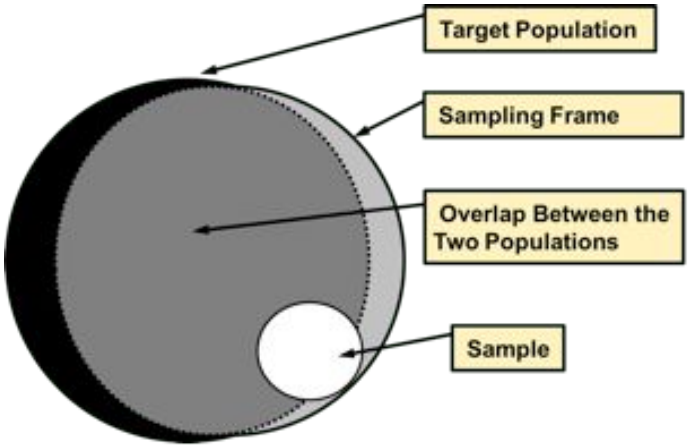
	% Roosevelt	# surveyed
The Literary Digest poll	43%	10,000,000
George Gallup's poll	56%	50,000
George Gallup's prediction of Digest's prediction	44%	3,000
Actual election	61%	All voters

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.

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!

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!

Probability Samples

Probability sampling

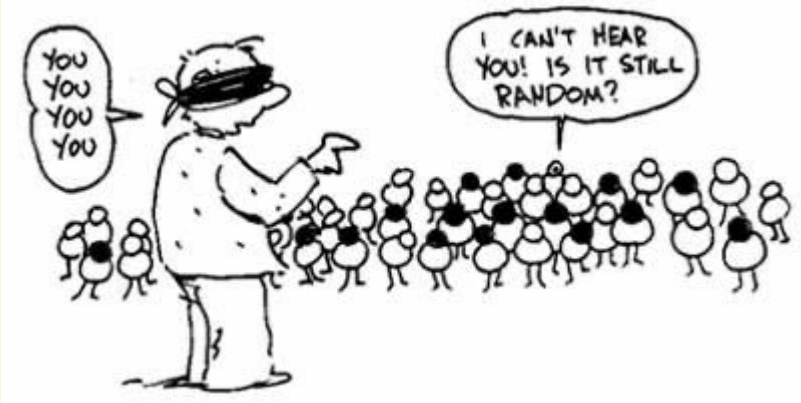
Sample is selected based on probabilistic procedure

- Allows assigning a precise probability to the event that each particular sample is drawn from the population
- Allows to quantify uncertainty/confidence about an estimator, prediction, or hypothesis test

Probability sampling

Be suspicious

- Whenever standard errors, p-values, or confidence levels are reported **without a proper explanation of the sampling procedure**.
- They could be meaningless or seriously wrong



Probability sampling

In order for a sample to be a probability sample:

- You **must** be able to provide the chance 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.

Simple Random Sample

A useful representation for sampling is a box model

- where the population of interest is represented by a box of N tickets, each with values written on them (data!)

less biased

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

- For a small sample compared to the population, SRS is very close to sampling at random with replacement.

Simple Random Sample

- How many ways are there so select an SRS of size n from a population of size N?

- What is the chance that a particular element of the population is selected by SRS?

Example scenario

Consider the following sampling scheme:

- Suppose a class roster has 1200 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).

Pause here and answer these questions!

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 1200 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 + 1190], 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 1200 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 + 1190], 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 1200 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 + 1190], 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.

A very common approximation

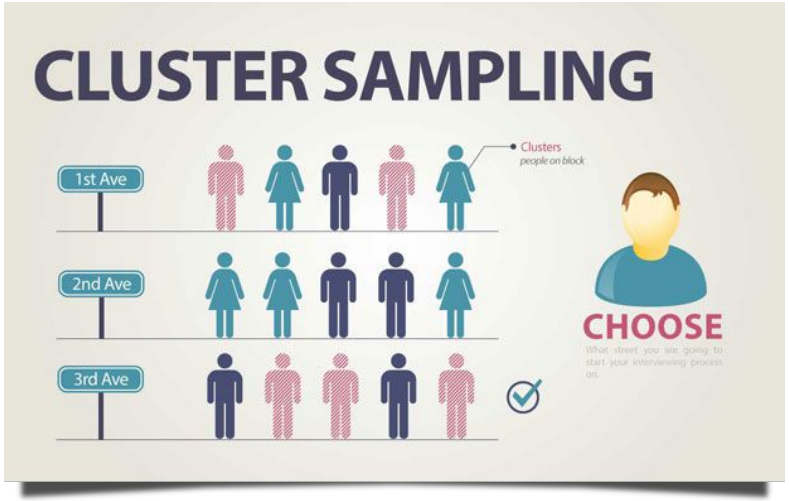
- A common situation in data science:
  - We have an enormous population.
  - We can only afford to sample a relatively small number of individuals.

- 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!**

Cluster Sample

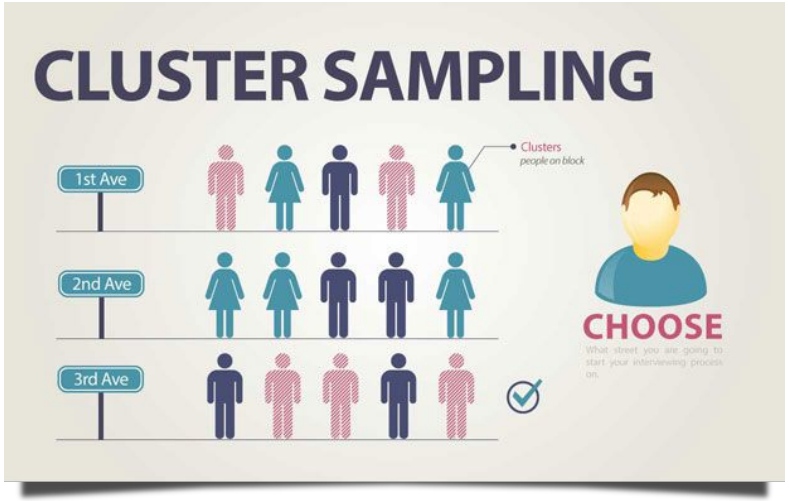
1. The population is divided into clusters of individuals.
2. One then uses SRS to select entire clusters instead of individuals.



Cluster Sample

Makes data collection easier!

- E.g. easier to poll entire towns (Anam) of a few hundred people each than to poll thousands of people distributed across the entire area (Seoul).
- But, tends to produce greater variation in estimation.
    - We need to take larger samples than with SRS.



Stratified Sample

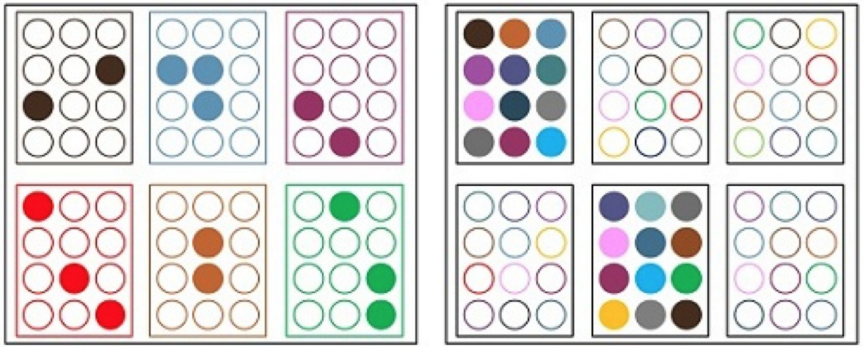
1. The population is divided into strata of individuals, e.g., based on demographics.
2. Select SRS of individuals in each stratum.

Stratified vs. Cluster Sample

Cluster vs. Stratified Sample

- In cluster sampling, we use a single SRS to select groups
- In Stratified sampling, we use one SRS per group to select individuals

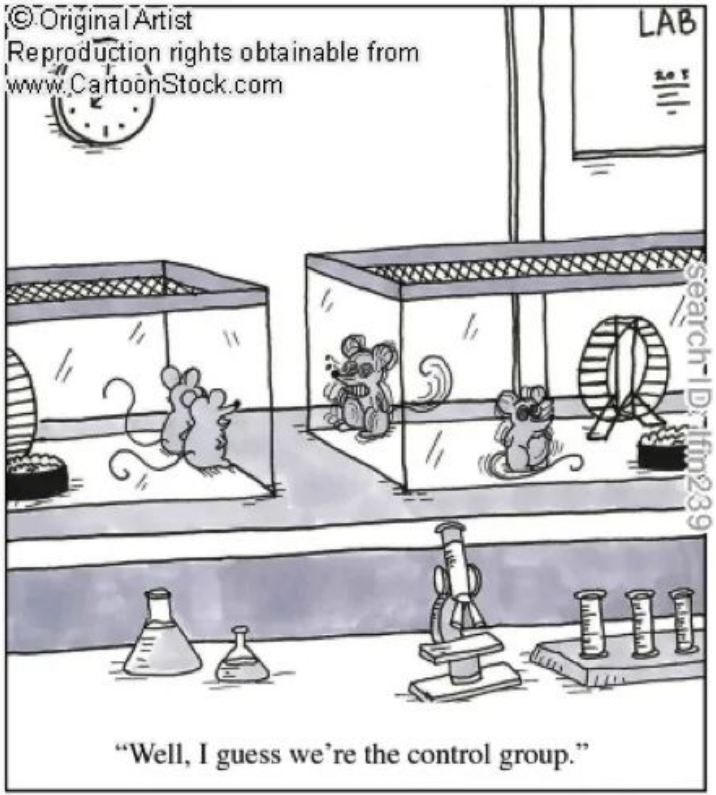
Stratified sampling results in increased precision and representation.



Designed Experiments

Examine the association/effect of a treatment on an outcome when the variable of interest is under the control of the investigator.

- E.g. clinical trial to test effect of new drug on patients with Alzheimer's disease.
- E.g. A/B test for two versions of a website



Randomized controlled trial (RCT)

A type of designed experiment in which participants in the trial are randomly allocated to either

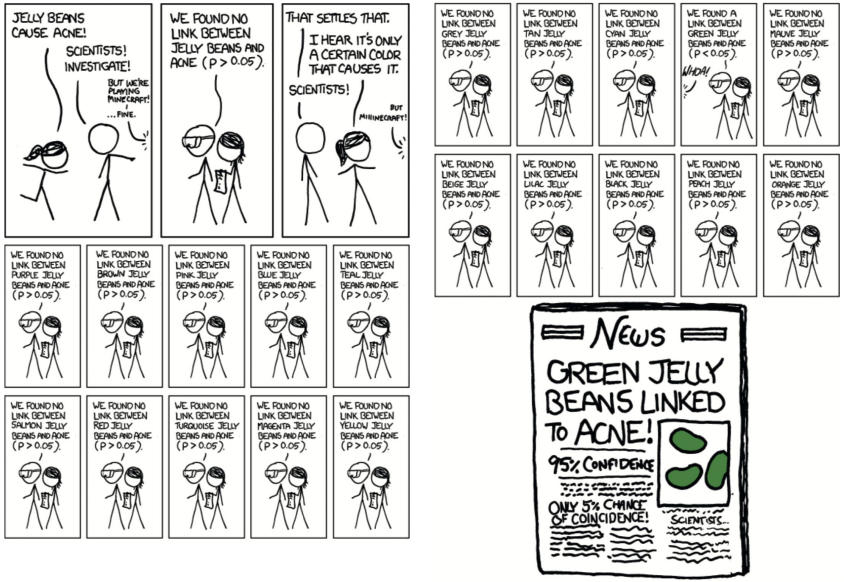
- the **group receiving the treatment** under investigation
- or a **control group** receiving standard treatment, no treatment, or a placebo.

RCT is often considered the gold standard for many types of investigations, e.g. clinical trials

Observational studies

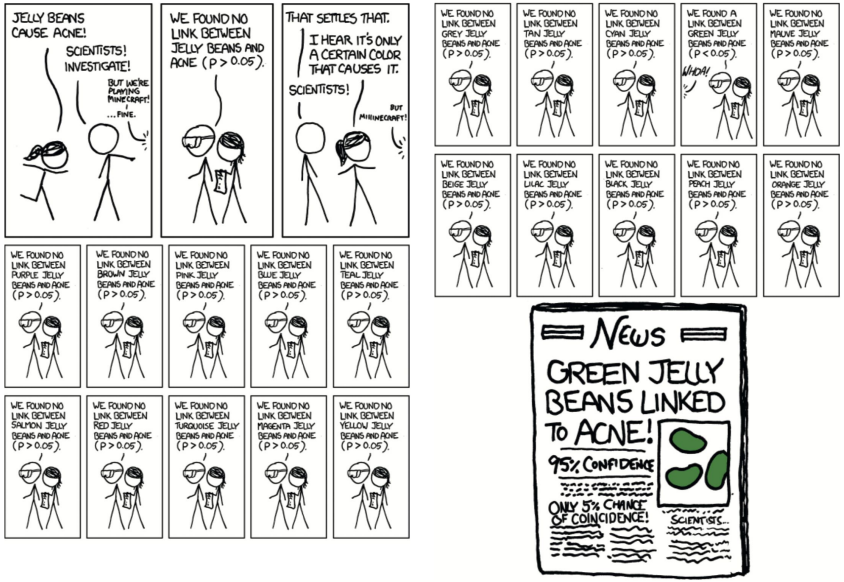
Examine the association/effect of a treatment on an outcome when the variable of interest is not under the control of the investigator

- E.g. Study effect of smoking on health



Observational studies

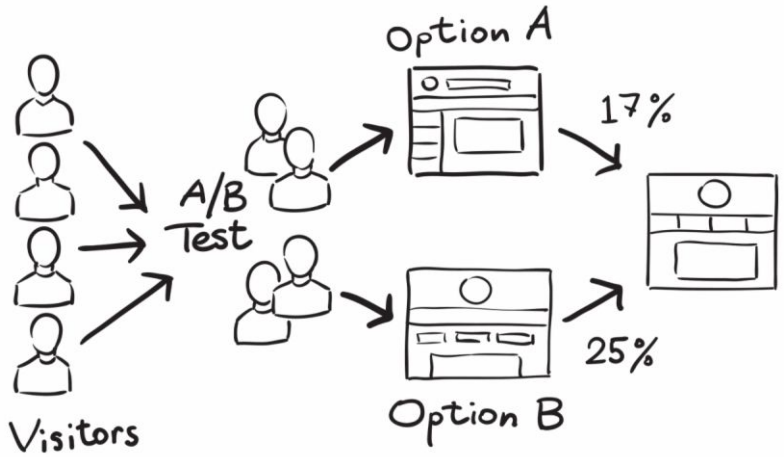
- **Case-control study**: Two existing groups differing in outcome (case vs. control) are identified/sampled and compared on the basis of variables potentially associated with the outcome.
- **Cross-sectional study**: Data are obtained at a specific point in time for each subject
- **Longitudinal study**: Data are obtained at multiple timepoints for each subject



A/B Testing

Determine whether two samples were **drawn from the same population**, i.e. have the same data generating distribution.

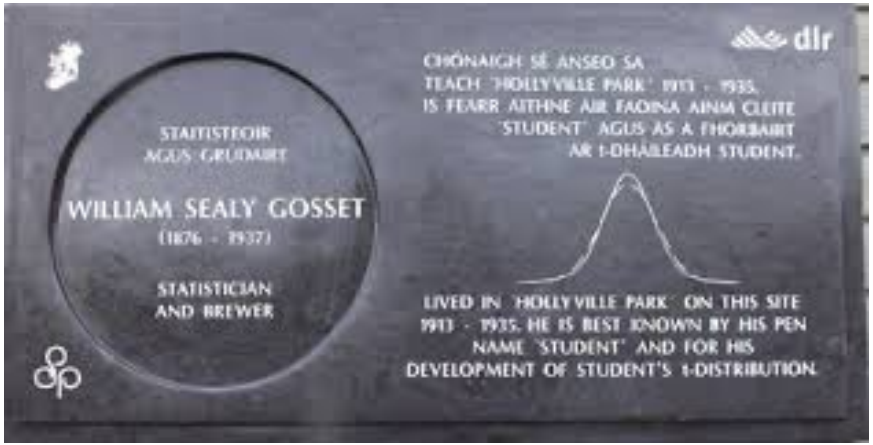
- Widely used in industry for marketing, website, and mobile app design.
    - e.g. comparing two different types of subject headers in e-mailing campaigns.



A/B Testing

Google engineers ran their first A/B test in 2000 in an attempt to determine the optimum number of results to display on the search engine's result page.

- T-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guiness brewery and whose pseudonym was "Student"



Summary

Welcome to the first content of COSE471! Here we plan on:

- Data Science Lifecycle
- Census, Survey, and Sampling
    - Why we need to sample in the first place.
    - What it means for our sample to be biased.
    - How to prevent these biases in our samples.
    - What exactly a sampling frame is, and why choosing a good one is important.
- Designed Experiments