Introduction to Data Science

Data: What, Why & How

Ernst C. Wit

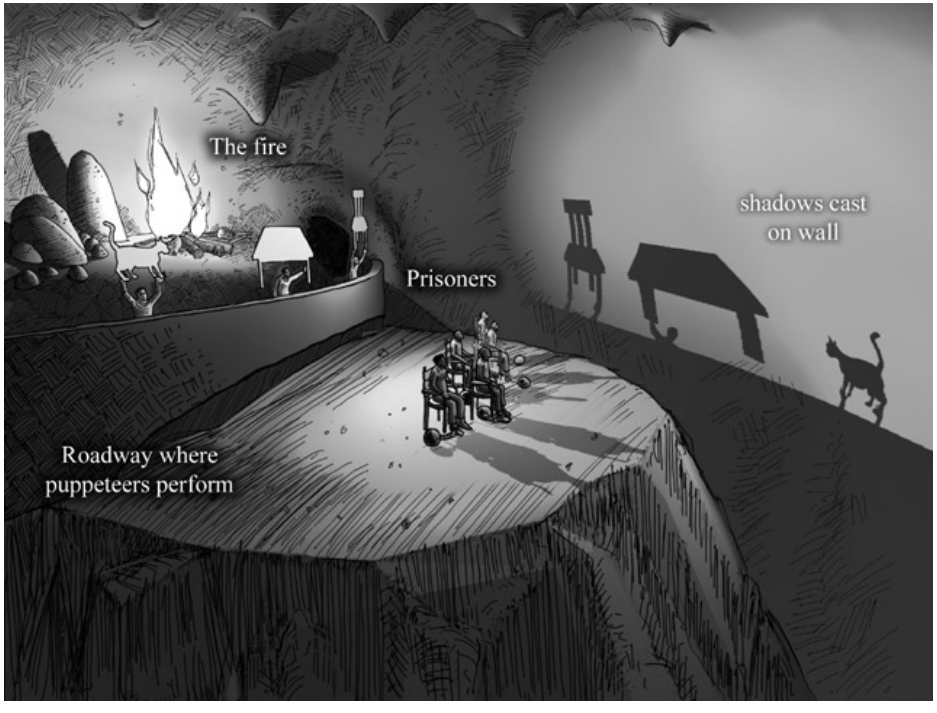
September 29, 2024

Lecture 3:

How to get Data?

Inductive Reasoning

- **Inductive reasoning**: Starting from specific observations to make general statements.
- Used in many scientific studies to form general conclusions.



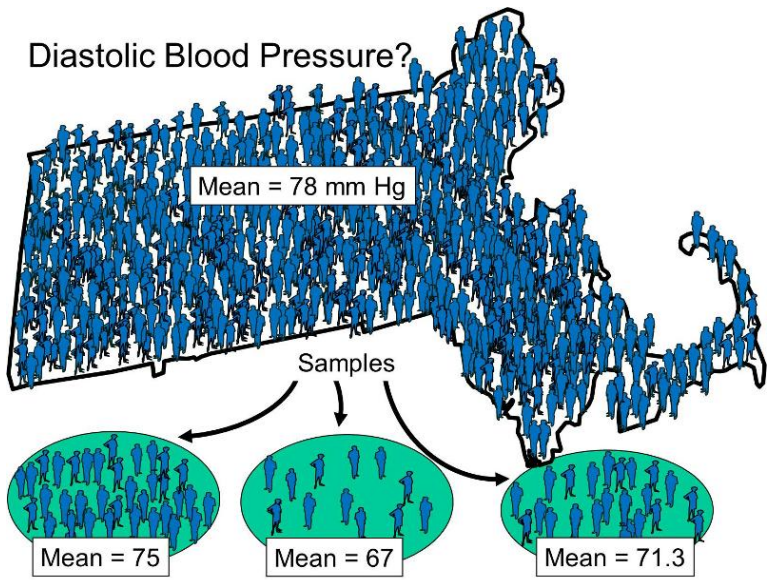
Goal: To make a statement about a **population**.

What is a Population?

Population

The entire set of individuals or items of interest.

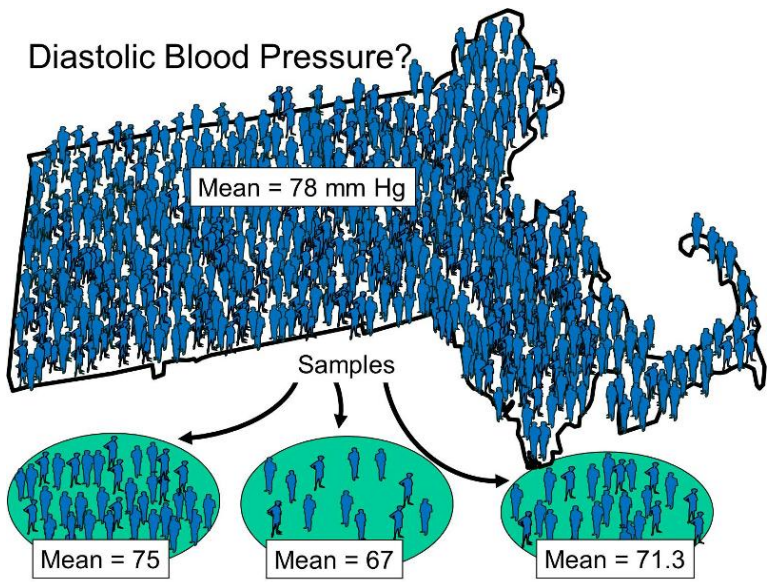
- Can be large or small depending on the context.
- Examples:
  - All residents of Lugano;
  - All current and future sufferers of breast cancer;
  - All interactions between students in IDS class.



General statements about a population

Examples of general statements about a population:

- 71% of the people in Lugano owns a car
- Efficacy of Capecitabine to treat breast cancer is 26.8%
- 57% of all interactions between IDS students are reciprocated.

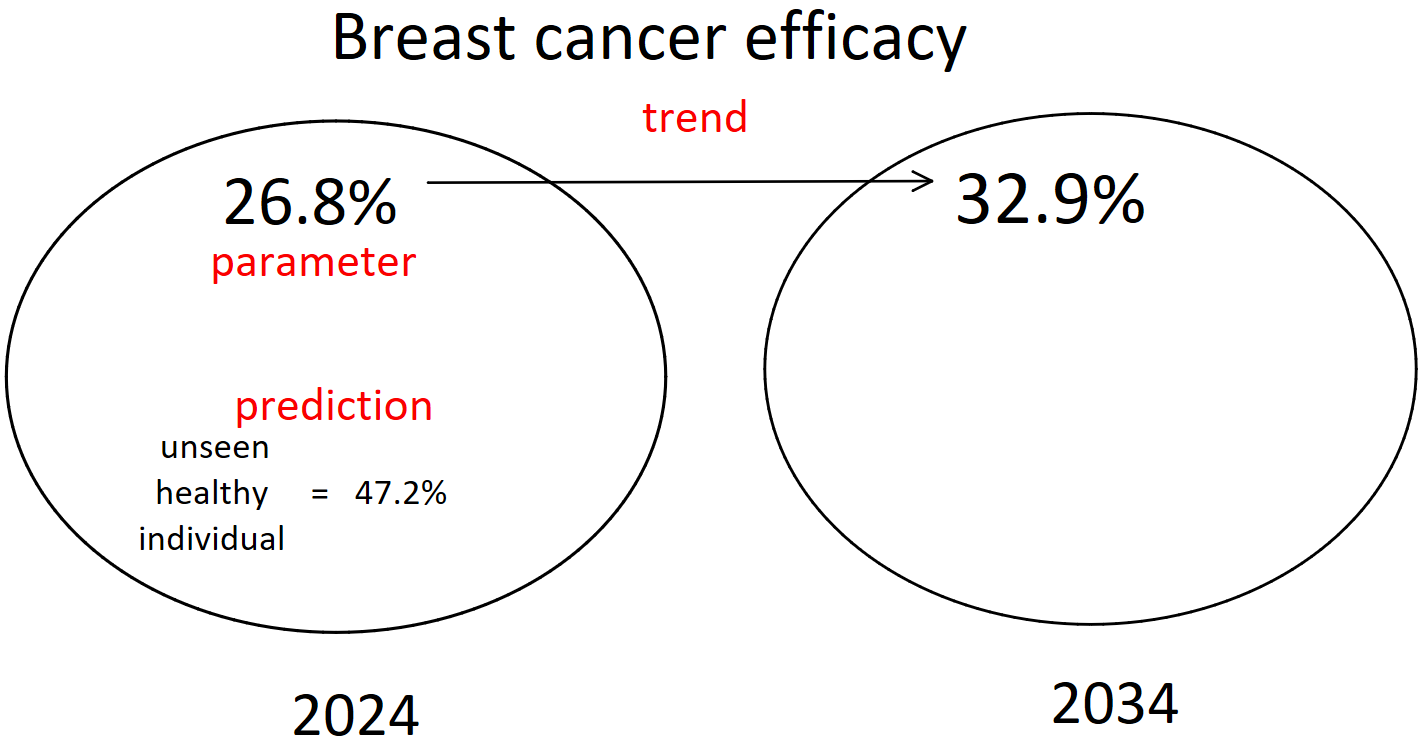


NOTE: Statements about population are true or false.

Type of general statements about a population

What type of general statements are we interested in?

- to find parameters (properties of population),
- to understand trends (about future populations),
- to make predictions (about unseen members of population).



Examples of Complete Data

Complete data refers to datasets where information is available for every individual or element in the population of interest, e.g.,

- **Census Data:**
  Population census collects data from every individual within a country.
- **Employee Records:**
  HR department has data about all employees, including their salary, job title, and employment history.
- **Amazon Product Sales:**
  Amazon tracks every sale, with complete data on product types, quantities sold, dates, and customer information.
- **School Enrollment Data:**
  Data for all students enrolled, such as attendance records, grades, and demographic information.

No Statistical Analysis with Complete Data

In principle, statistical analysis is unnecessary with complete data.

- **No Sampling Required:**
  With complete data, entire population is represented, so no need to infer population parameters from a sample.
- **Exact Population Parameters:**
  Population characteristics such as mean, variance, and proportions can be calculated directly without estimation.
- **No Uncertainty:**
  With complete data, there is no uncertainty or error from sampling, and results are deterministic.

Real-World Example:
In a national census, we know exactly how many people are unemployed, how many children are born, etc. etc.

Complete Data is Almost Always Impossible

In practice, complete data is rarely achievable:

- **Dynamic Populations:**
  Populations are often not static. They evolve over time, meaning new individuals or items are continually introduced.
- **Future Individuals or Events:**
  Induction also involves predictions about future, e.g. trends, but tuture population does not exist yet.
- **Inaccessible Data:**
  It's often impractical to gather data from every individual due to cost, time, or logistical constraints.
- **Errors and Missing Data:**
  Data collection often encounters errors or missing values.

Real-World Example:
In economic forecasting, data on future market conditions, technological advances, or demographic changes is unavailable.

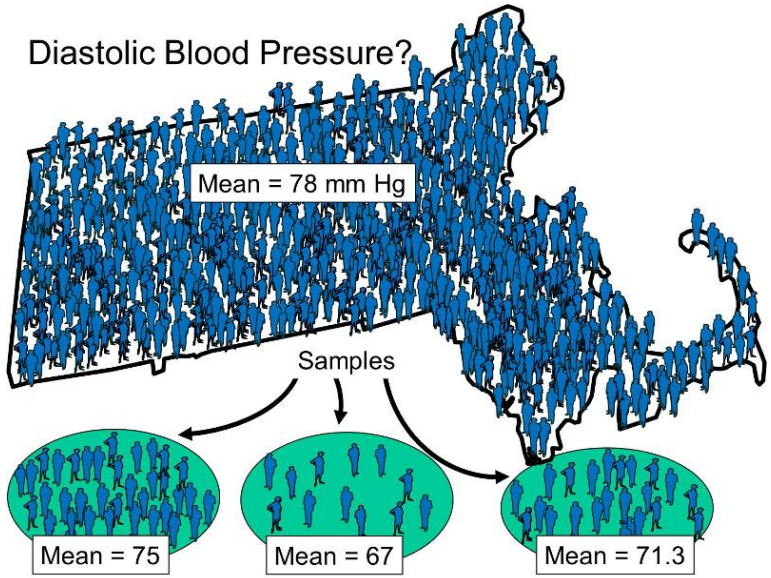
Sample: incomplete data

Sample

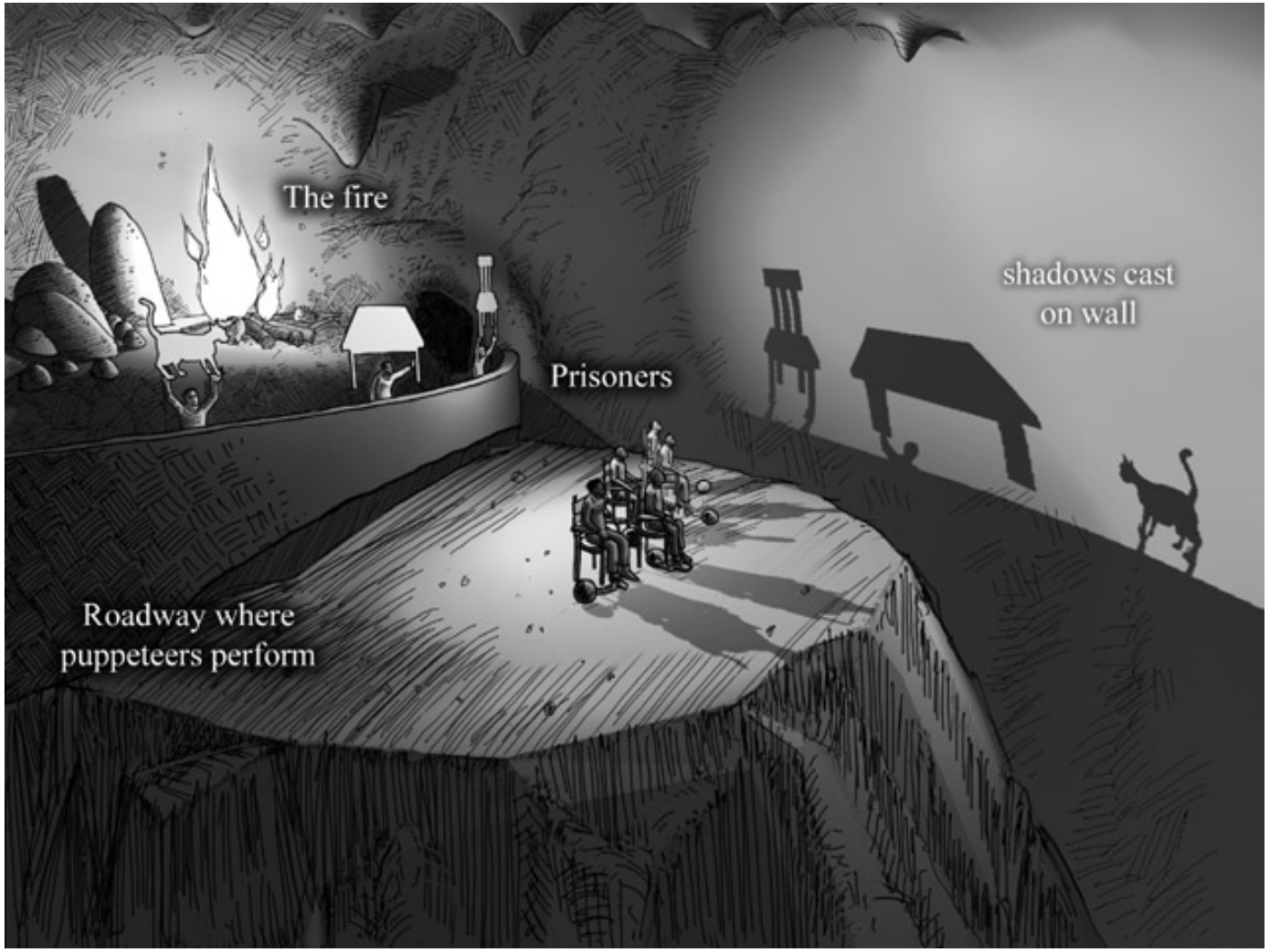
Subset of the population of interest.

Using a sample from population, we try...

- to draw conclusions about population (see below),
- to understand trends (about future populations),
- to make predictions (about unseen members of population).



Plato's cave

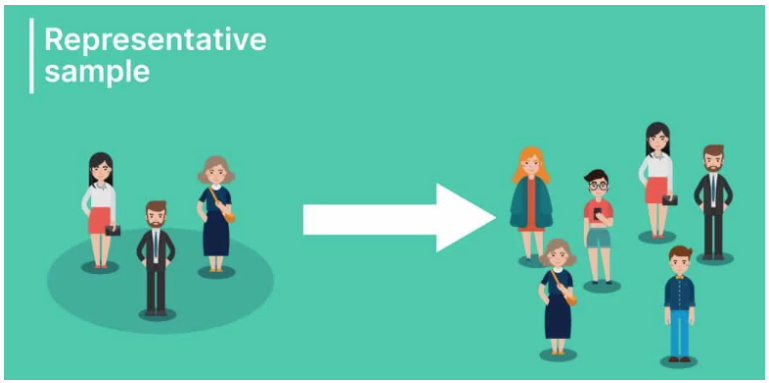


parameter: true cat, **sample:** shadow figures of a cat

Characteristics of a Good Sample

A good sample must be

- **Representative:**
  Reflect characteristics of population, including its diversity and distribution.
- **Unbiased:**
  Free from systematic errors that would favor certain samples over others.
- **Sufficiently Large:**
  Large enough so that errors cancel.



Example: Determination of the foot

Sampling strategy: *Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished. (Koebel 1500)*

Good aspects:

Example: Determination of the foot

Sampling strategy: *Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished. (Koebel 1500)*

Good aspects:

- Everyone goes to church, so "representative"
- "As they happen to pass": Arbitrary, so no bias.
- 16 is not small.

Bad aspects:

Example: Determination of the foot

Sampling strategy: *Stand at the door of a church on a Sunday and bid 16 men to stop, tall ones and small ones, as they happen to pass out when the service is finished. (Koebel 1500)*

Good aspects:

- Everyone goes to church, so "representative"
- "As they happen to pass": Arbitrary, so no bias.
- 16 is not small.

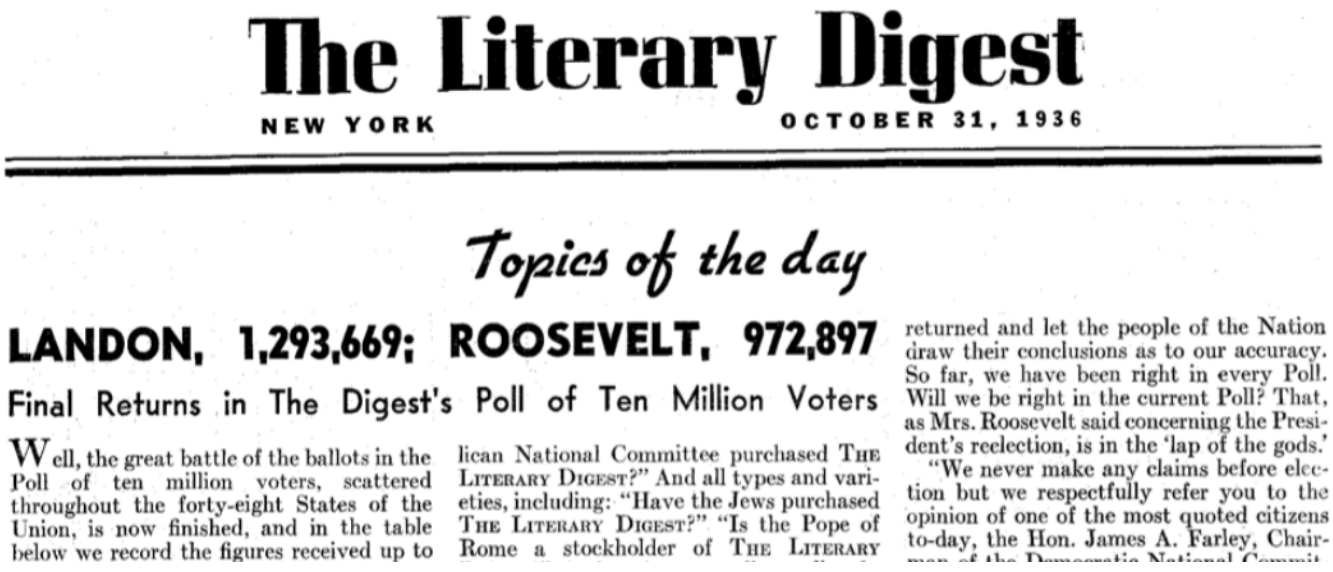
Bad aspects:

- What if a tall family comes out together?
- Is 16 men enough?

George Gallup's Contribution to Modern Polling

Gallup introduced polling techniques during 1936 US elections.

- Established method: Literary Digest polled millions of people but failed to represent general population accurately.
- Gallup used a smaller, more carefully selected **sample** that was representative of the U.S. population.
- Gallup predicted **correctly** Franklin D. Roosevelt win Literary Digest predicted falsely a landslide for Alf Landon.



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

lican 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

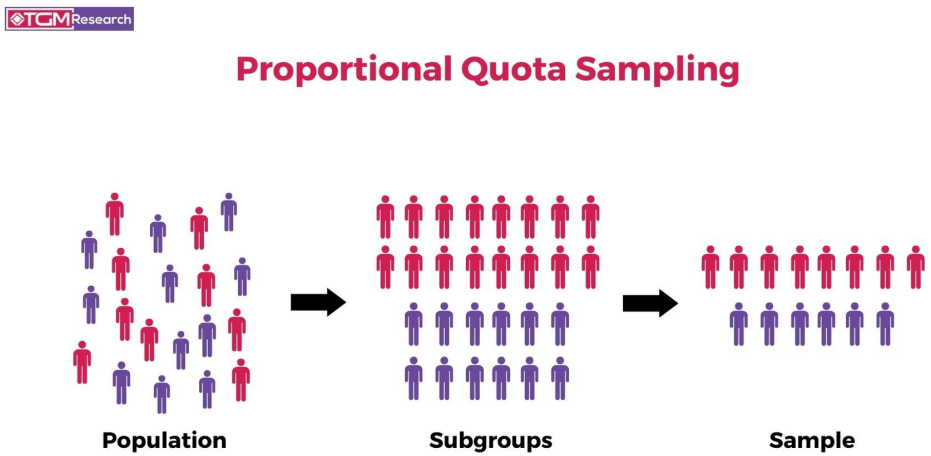
returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"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, Chair-

Importance of Gallup's Method

Gallup's success showed that

- **sample size** was less important
- than sample **representativeness**.
- He introduced **quota sampling**: key demographics were proportionally represented (e.g., age, income, geography).

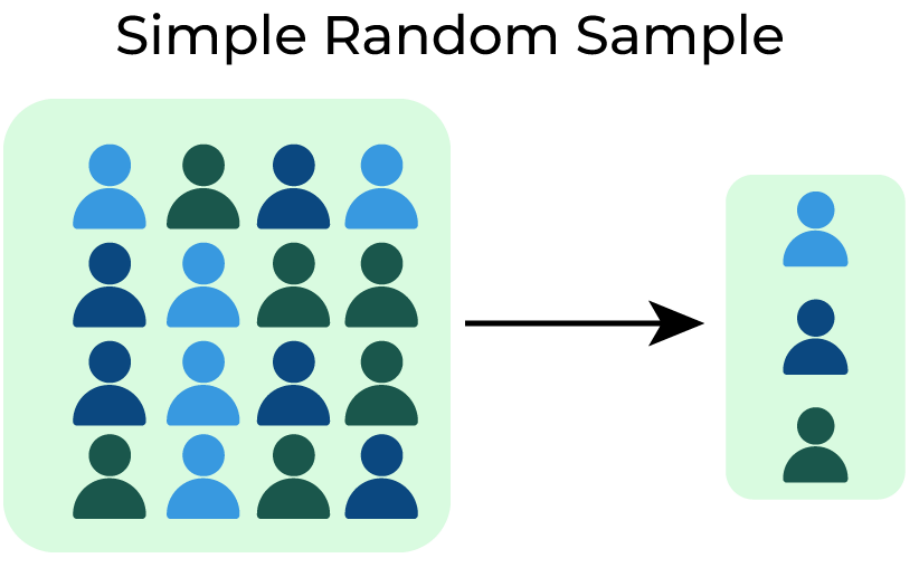


Literary Digest's sample was biased toward wealthier individuals (those with phones).

Simplest Strategy to get Good Sample

How to get a good sample?

- **Random:**
  Each member of population should have equal chance to be in sample:
  - reduces bias and
  - ensures sample is not systematically different from population.
- **Independent Observations:**
  Selection of one individual should not influence that of others.



Example: Sampling in Surveys

Polling during an election campaign.

- Population: All eligible voters in a country.

Example: Sampling in Surveys

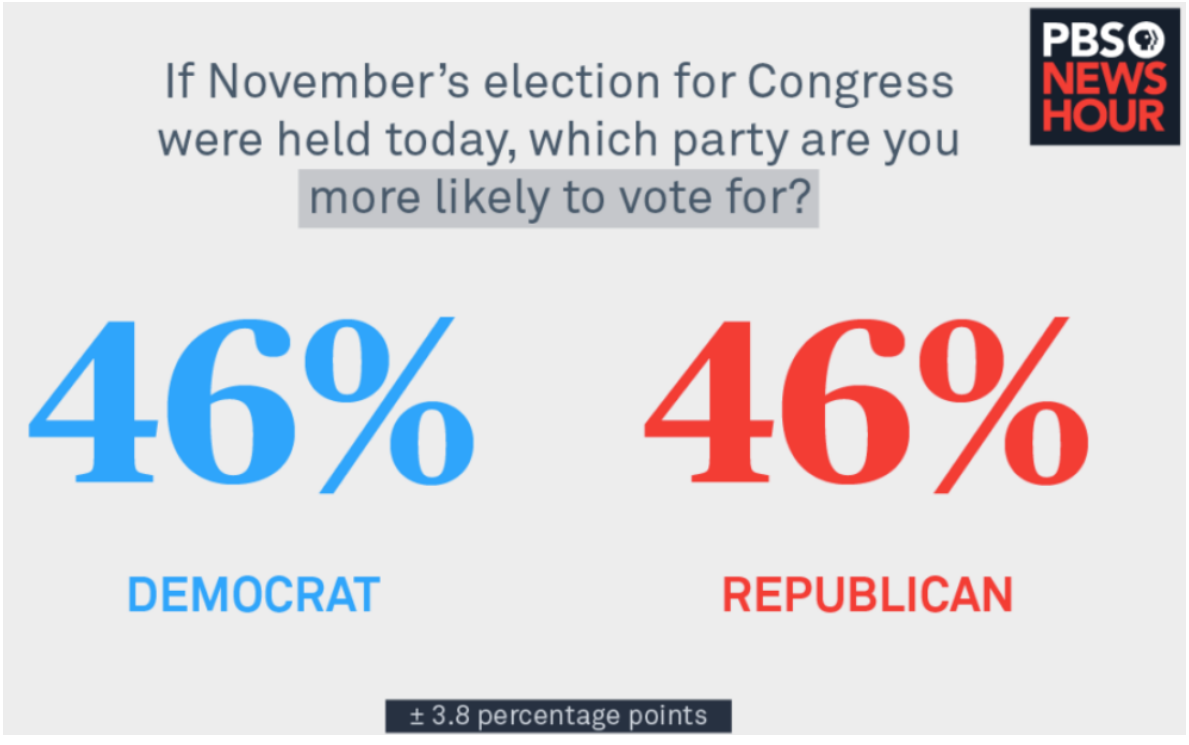
Polling during an election campaign.

- Population: All eligible voters in a country.
  better: all *actual* voters in upcoming election.
- Sample:

Example: Sampling in Surveys

Polling during an election campaign.

- Population: All eligible voters in a country.
  better: all *actual* voters in upcoming election.
- Sample:A random selection of 1,000 voters.

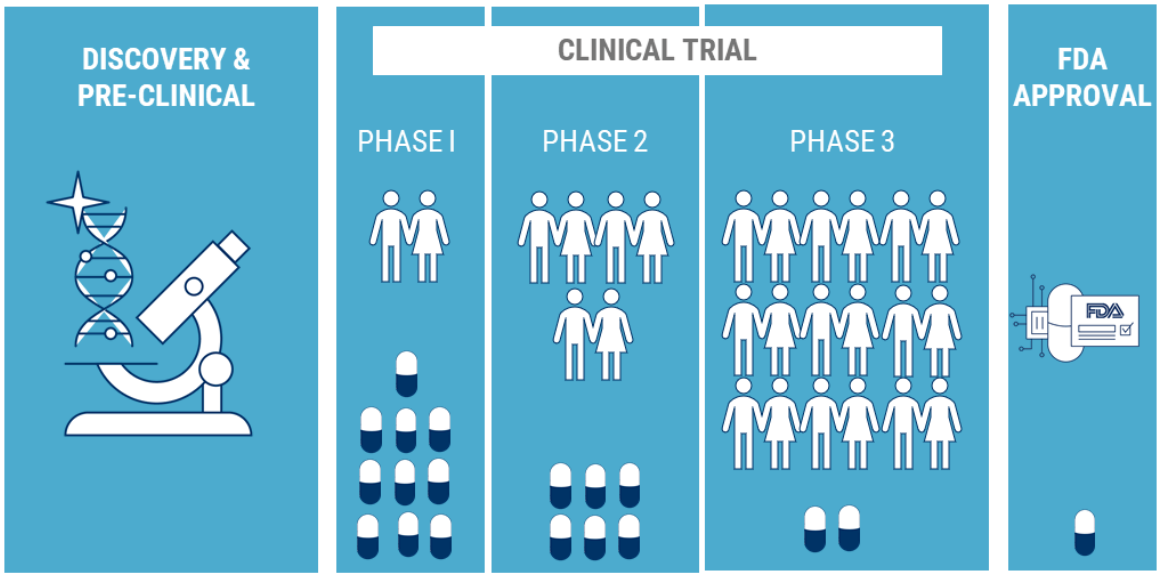


Goal: Use sample to predict outcome of election.

Example: Sampling in Medical Studies

Testing a new drug.

- Population: All people with a certain medical condition.
- Sample: 500 patients selected to participate in a clinical trial.



Source: cbinsights.com

Goal: Use sample to ascertain drug's effectiveness.

Summary of Key Points

- **Inductive Reasoning:** Using specific observations to make general statements about a population.
- **Population vs. Sample:**
  - Population: The entire set of individuals or items of interest.
  - Sample: A subset of the population used to make inferences.
- **Importance of Sampling:**
  - Complete data is rare, so sampling is crucial.
  - Samples must be representative, unbiased, and large enough.
- **Random Sampling:** everyone has an equal chance of being selected: *unbiased and representative*.