Introduction to Statistical Methods in Political Science

Lecture 1: Introduction to Data and Statistics

Ignacio Urbina

Definitions

The Study of Statistics: Some Definitions

What is statistics?

- Statistics is the study of collecting, analyzing, interpreting, and presenting data.
- Involves mathematical techniques to make inferences about a population from a sample.

Why is it important?

- Uncertainty is inherent in real-world phenomena. Statistics provides tools to manage and quantify uncertainty.
- Critical for data-driven decision-making across disciplines.

Population and Sample

Population:

- A population is the complete set of all possible observations or measurements of interest.
  - Example: All residents in a country when studying public health policies.

Sample:

- A sample is a subset of the population selected for analysis.
  - Example: A group of 1,000 residents surveyed to infer public opinion on policy.

Estimator

- An estimator is a statistical method used to estimate a population parameter based on sample data.
- A more abstract definition: *An estimator is any mathematical formula (function) computed using measures collected from a sample*.
    - Example: The sample mean used as an estimator of the population mean.

Statistical Inference

- Statistical inference is the process of drawing conclusions about a population based on sample data.
  - Example: Inferring the likely values of the average income of a population based on a sample.

Dataset

- A dataset is a structured collection of data, typically organized in a tabular format.
- Rows represent individual observations, and columns represent variables.
- Example: A dataset of economic indicators across countries.

Observations

- An observation is a single data point or record in a dataset.
- Each row in a dataset typically corresponds to one observation.
- Observations are instances of measurements or responses.

Variables

- A variable is a characteristic or attribute that can take on different values.
- Each column in a dataset typically represents a variable.
- Variables are classified based on the type of data they represent.

Example Dataset

- **Sample Dataset:**

Respondent	**Age**	**Income**	X_1	X_2	**State**	**Education**
1	34	55000	1	7	NY	High school
2	29	60000	0	1	NJ	Master's
3	45	70000	1	3	MA	High school
4	40	65000	0	4	MA	High school
5	38	62000	1	2	NY	Bachelor's
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Types of Variables

- Variables can be classified as **categorical** or **continuous**.
- Understanding the type of variable is essential for choosing the correct statistical method.
- The type of variable determines the appropriate summary and analysis techniques.

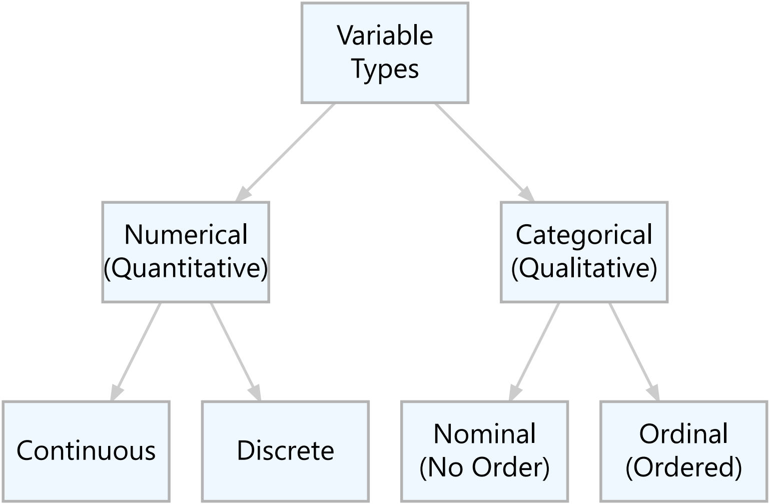
Categorical (Qualitative) Variables

- Categorical variables represent distinct categories or groups.
- Can be nominal (no order) or ordinal (ordered categories).
- Require specific transformations before statistical analysis (i.e., we need to ‘code’ them into numbers before analysis).
- Examples:
  - **Nominal**: No natural ordering among the categories. Example: Blood type (A, B, AB, O).
  - **Ordinal**: Categories have a natural order. Example: Educational level (High school, Bachelor’s, Master’s, PhD).

Numerical (Quantitative) Variables

- Numeric variables have values that describe a measurable quantity as a number, like ‘how many’ or ‘how much.’
- Numeric variables can be continuous or discrete.
- Examples:
  - **Continuous**: Can take any value within a range, including fractions and decimals. Example: Height, weight.
  - **Discrete**: Can only take non-negative whole numbers. Example: Number of children, number of cars owned.

Summary: Types of Variables



Example: Types of Variables

Can you correctly identify the types of variables included in this dataset?

- **Sample Dataset:**

Respondent	**Age**	**Income**	X_1	X_2	**State**	**Education**
1	34	55000	1	7	NY	High school
2	29	60000	0	1	NJ	Master's
3	45	70000	1	3	MA	High school
4	40	65000	0	4	MA	High school
5	38	62000	1	2	NY	Bachelor's
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Statistical Principles Involved in Research Studies

Defining an Empiricist Research Question

Research Question:

- A question that seeks to explore observable and measurable phenomena.
- Must be testable through empirical data and direct observation.
- Aims to contribute to existing knowledge, test theories, or solve practical problems.

Example:

- **Poorly Defined:** "Does exercise affect health?"
- **Well-Defined:** "What is the impact of a 30-minute daily aerobic exercise regimen on the cardiovascular health markers (e.g., blood pressure, cholesterol levels) of adults aged 30-50 over six months?"

Components of a Well Defined Research Question

- **Specificity:** Clearly defines variables and their relationships or differences, ensuring focus and clarity.
- **Operationalization:** Concepts are defined in measurable terms, facilitating precise data collection.
- **Feasibility:** The question is practical to investigate considering available resources, time, and ethical standards.
- **Relevance:** Addresses significant issues or gaps within the field.
- **Novelty:** Contributes new insights or perspectives to existing research.

Example: Population and Research Question

Research Question: What is the impact of a universal basic income (UBI) on economic stability and poverty reduction, specifically among low-income households?

Population of Interest: Low-income households across the entire country, with consideration of regional, demographic, and economic diversity.

Associated and Independent Variables

- **Associated Variables**: Variables that show some relationship with each other. The association can be **positive** (both increase and decrease together), **negative** (one increases while the other decreases), or **non-linear**.
- **Independent Variables**: Variables that are not associated and do not influence each other.

Example

What is the association between the total number of community activities in a neighborhood and the level of social trust among its residents?

Explanatory and Response Variables

- **Explanatory Variable**: The variable suspected of affecting the other. Often considered the cause.
- **Response (Outcome) Variable**: The outcome or effect being measured.
- **Note**: *Association between variables does not imply causality*, even if an explanatory-response relationship is identified.

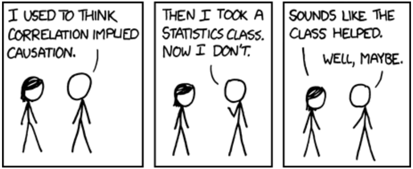

I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

Figure: Correlation (Source: XKCD)

Observational vs. Experimental Studies

Aspect	Observational Study	Experimental Study
Definition	Data collected without manipulating variables	Involves manipulating one or more variables
Purpose	Identify associations between variables	Establish cause and effect relationships
Causation	Cannot determine causality	Can determine causality with proper design

Confounding Variables and Statistical Bias

Confounding Variables

Definition: Unaccounted variables that influence both the explanatory and response (outcome) variables.

Impact: Without controlling for confounders, researchers may **misattribute** the association between variables, incorrectly inferring a direct relationship.

Statistical Bias

Definition: Systematic errors in data collection, analysis, interpretation, or review processes that can **distort** findings. **Example:**

Confounding Bias is a distortion that alters the true association between an explanatory and an outcome variable due to a third factor that is independently associated with the exposure and the outcome (Source).

Example: Assessing the Impact of Math Tutoring

Research Context:
A high school is implementing a new structured math tutoring program that provides students with an additional 2 hours of tutoring per week. Participation is voluntary. The goal is to determine whether this program positively affects students' performance in standardized math tests over the course of one academic year.

Task:
Identify and describe **one variable** that could cause **confounding bias** when assessing the association between **participation in the math tutoring program (X)** and **standardized math test scores (Y)**.

Experiments in Social Science

Context:
Social science experiments help us understand cause-and-effect relationships by manipulating a specific variable (the *treatment*) under controlled conditions.

Experimental Treatment (Definition):
The specific intervention or condition applied exclusively to the **treatment group**. This allows researchers to measure its impact on the response (outcome) variable compared to the **control group** (untreated).

Principles of Experimental Design

- **Control**: Accounting for confounding factors to ensure observed effects are due to the treatment.
- **Randomization**: Randomly assigning subjects to treatment and control groups to eliminate selection bias.
- **Replication**: Repeating the experiment or using a sufficiently large sample to ensure the results are reliable.
- **Blocking**: Grouping subjects with similar characteristics to reduce variability and better isolate the treatment effect.

Blinding in Studies

- **Single Blinding**: Participants are unaware of whether they are in the treatment or control group.
- **Double Blinding**: Both participants and experimenters are unaware of the group assignments.
- Blinding helps reduce bias, particularly the Performance and Experimenter bias.
  - E.g., in a clinical study, if participants in the control group systematically seek other treatments, there could be performance bias.
  - Or, if researchers/clinicians treat participants differently depending on which group they are in, this could imply experimenter bias (see: Experimenter Effect).
  - **Placebo Effect**: Change in participants' outcome variable due to their belief in the treatment rather than the treatment itself.

Sampling: Basic Principles

Sampling Methods and Sampling Bias

Sampling Method (Definition):
A sampling method refers to the process used to select individuals or observations from a population to be included in a sample.

Sampling Bias (General):

- Sampling bias occurs when the process of selecting a sample **skews the results**, making some members of the population systematically more (or less) likely to be included than others.

Sampling Methods and Sampling Bias

Types of Sampling Bias:

- **Non-response Bias:** Arises when individuals selected for the sample do not respond or are unwilling to participate, and those non-responders **differ in important ways** from those who do respond.
- **Self-Selection Bias:** Occurs when participation in a survey, study, or experiment is **voluntary**, allowing individuals to decide on their own whether to be included. As a result, those who opt-in may **systematically differ** from those who do not, leading to a non-representative sample.
- **Coverage Bias:** Happens when some members of the population are inadequately represented in the sample due to limitations in the **sampling frame**. This occurs when certain groups have **no chance** or a **lower chance** of being included, leading to a distorted view of the population.
  - **Sampling Frame:** The list or method used to identify and select individuals from the population for inclusion in the sample.

Probabilistic Sampling

- **Probabilistic Sampling (def.):** A method where each member of the population has a known, non-zero chance of being selected in the sample. This approach ensures that the sample is more representative of the population.
  - **Examples:** Simple random sampling, stratified sampling, cluster sampling.
- **Advantages:** Reduces sampling bias, allows for generalization to the population, and facilitates the use of statistical inference.

Sampling Techniques

- **Simple Random Sampling**: Every subject has an equal probability of being selected.
- **Stratified Sampling**: Population is divided into strata; a random sample is drawn from each stratum.
- **Cluster Sampling**: Population is divided into clusters (i.e., naturally forming groups that are diverse within and similar between); a random sample of clusters is selected, and all subjects within those clusters are studied.
- **Multistage Sampling**: Corresponds to taking samples in stages using smaller and smaller sampling units at each stage.
  - For example, clusters are chosen randomly, and then a random sample from within each cluster is selected.

Stratified vs. Cluster Sampling

Stratified Sampling

- **Objective:** Ensure representation across distinct subgroups
- **When to Use:**
  - Population is heterogeneous and can be divided into meaningful strata
  - Detailed analysis is required within each subgroup
  - High precision in estimates across strata
- **Example:** Survey different income brackets to study economic disparities

Cluster Sampling

- **Objective:** Reduce cost and logistical complexity for large, dispersed populations
- **When to Use:**
  - Population naturally forms clusters
  - Easier to collect data from entire clusters than from scattered individuals
  - Ideally, each cluster is diverse internally (a "mini-population")
- **Example:** Conduct a health survey by sampling entire schools as clusters

Convenience (Non-probabilistic) Sampling

- **Convenience (Non-probabilistic) Sampling (def.):** Individuals who are easiest to reach or most accessible are chosen for the sample, rather than randomly or systematically selected.
- **Potential Issues:**
  - **Bias:** May result in a sample that is not representative of the population. Findings may not be applicable to the broader population due to the non-random nature of sample selection.
  - **Examples:** Surveying people in a mall or using participants from a single location or organization.

Representativeness of a Sample

Definition: A sample is **representative** if its key characteristics (e.g., demographics, behaviors) closely match those of the overall population, allowing for valid generalizations.

Discussion:

- *Survey Vendors*: Many commercial survey vendors sell **representative samples** based on a few demographic features but not true probabilistic samples, meaning results may still be affected by sampling bias.
    - The most common form of representative sample is quota sampling.
- Even a probabilistic sample can be **non-representative** if it is too small or improperly stratified.

Terms and Concepts Covered

Terms and Concepts Covered in this Lecture

- Statistics
- Population
- Sample
- Estimator
- Statistical Inference
- Dataset, Observations, & Variables
- Categorical/Continuous Variables
- Response/Explanatory Variable
- Observational/Experimental Study
- Confounding Variables
- Bias
- Sampling Techniques
- Voluntary Response
- Simple Random Sampling / Stratified / Cluster / Multistage Sampling
- Research Question
- Population of Interest
- Associated/Independent Variables
- Blinding
- Control
- Randomization
- Replication
- Blocking
- Sampling Bias
- Convenience Sampling Bias

Introduction to Statistical Methods in Political Science

Lecture 2: Summarizing Data I - Descriptive Statistics

Ignacio Urbina

Descriptive Statistics

Keeping the Goal in Sight

- Recap: Inferential statistics → Learning about the properties and characteristics of a population using samples.
- Recall that we call a ‘statistic’ any quantitative value that is a function of our sample data and, through a specific function, maps them into a single measurement meant to represent a feature of the data.

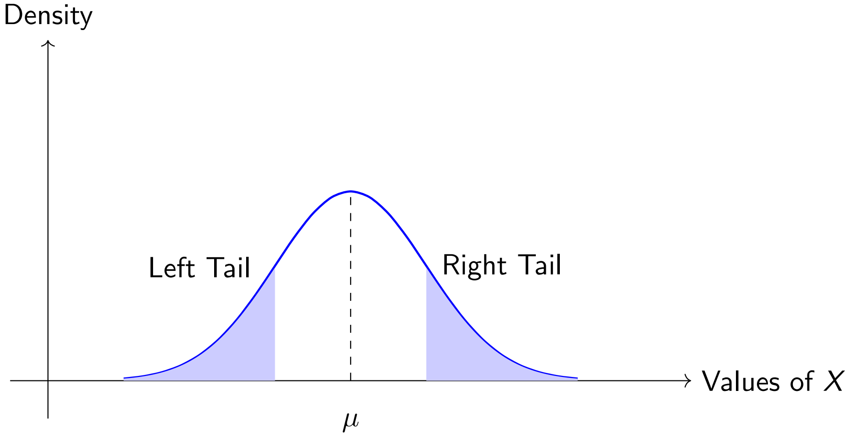
Descriptive Statistics

- Before making inferences about the population, it is essential to learn how to effectively describe the properties and characteristics of the data we collected in our sample.
- We call this process “descriptive statistics” and use different measures (statistics) to describe our data.
- We have three types of descriptive statistics: univariate, bivariate, and multivariate.

Distribution of a Variable

- At a fundamental level, when doing descriptive statistics, our goal is to provide a summarized description of the distribution of our data.
- *Def.* **Distribution of a variable**. The distribution of a variable is a function (often represented in a graph) that shows the possible values of a variable and how often they occur.

Distribution of a Continuous Variable



Distribution of Categorical Ordinal Data

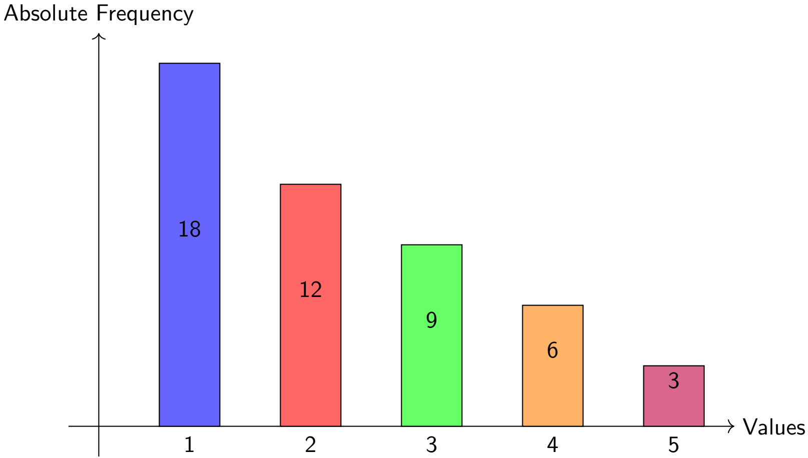


Figure: Categorical Ordinal Distribution

Summarizing Data

- Datasets often include hundreds, thousands, and, on some occasions, millions of observations. While looking at the dataset is always instructive, it is impossible to make sense of it just by doing so.
- Hence, we need to summarize our data. By this, we mean **aggregating** —or putting all the data together—into a few pieces of information that are far easier to read and interpret.

How to Describe Data

- When we summarize data we, again, seek to provide a summarized description of the distribution of a variable.
- In **univariate descriptive statistics**, we will describe the distribution of a variable using three approaches:
  - Measures of Central Tendency
  - Measures of Dispersion Around the Center
  - Shape of the Distribution

Measures of Central Tendency

- We are often interested in describing a distribution by providing one value representing its center.
- By “center,” we mean a numeric value that balances the distance between all the other points in the distribution in some specific way.
- Depending on a variable’s specific type of distribution and the type of variable (categorical or numerical), we will use the **average**, **median**, or **mode** to best represent the center of the distribution.

Central Tendency: the Average (Arithmetic Mean)

The following is the function (formula) for the **average** or **sample mean**:

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{x_1 + x_2 + \cdots + x_{N-1} + x_N}{N}$$

Where:

- *i* represents an arbitrary index we use to label our observations in our sample
- The total number of observations is denoted by the letter N, and we call it “sample size.” (also N rows in our dataset)

The Average: Some Notes

- Why is the average useful? The average is useful because it provides a single measure summarising the entire dataset with just one value.
  - This simplifies our understanding of and ability to work with data. Despite being affected by outliers, the average is still important in various statistical analyses.
- *Caveat*: the average does not necessarily equal the value most likely to be randomly drawn from the data.

On Outliers or Extreme Values

- Sometimes the data will have extreme values, which are values that according to the pre-established criteria (we will review this in time) can be deemed as extremely far from the center of the distribution, such that these values are very unlikely.
- In small samples, these outliers can disturb the usefulness of the mean as a description of central tendency, such that the information they provide is of lesser qualitative significance.

Example of Influence of Outliers

- Imagine we have a sample size of $N_{\text{Sample 1}} = 10$. We have measured a random sample of people's yearly income. Assume we find $\bar{X}_{\text{Sample 1}} = 40,000$.
- Now consider another sample, this time of size $N_{\text{Sample 2}} = 1,000$. Assume we find that $\bar{X}_{\text{Sample 2}} = 42,500$.
- Assume in each sample, we forgot to include an additional data point: $X_{N+1} = 150,000$. How does the mean change in each case?

Central Tendency: the Median

- When numerical data have a natural ordering, we can also use an alternative measure of central tendency: the **median**.
- The median is the *value of the distribution that lies in the middle* such that 50% of the values are to the left and 50% to the right.
- In other words, with the median, our notion of distance from the "center" is more concerned with rank order than numeric absolute distance.

Central Tendency: the Median

Here's the general formula for the median depending on whether N is odd or even:

$$median(X) = \begin{cases} X_{\{\frac{N+1}{2}\}} & \text{if } N \text{ is odd} \\ \frac{1}{2}(X_{\{\frac{N}{2}\}} + X_{\{\frac{N}{2}+1\}}) & \text{if } N \text{ is even} \end{cases}$$

- In the presence of outliers, the median can be a more resilient measure of central tendency than the mean, especially in small samples.
- Why? Because the median only considers the rank order of the values of the distribution, therefore, it is robust against outliers.

Central Tendency: The Mode

- Another central tendency summary statistic is the mode, which is particularly useful for categorical values.
- *Def.* **Mode**: The mode is the most frequent value of the distribution.
- Because of this definition, the mode is also considered the most likely value of the distribution.
- Yet, in practice, the mode is only useful for categorical values (It is important to think why this is the case).

Some Notes: The Mode

- Note that the mode is also a robust statistic in that it is resilient against outliers.
- When a distribution has only one mode, we call it "unimodal;" when it has two, we call it "bimodal."
- While we might be tempted to always prefer the mode as a measure of central tendency for categorical variables, in practice, we should always look at all the *appropriate* measures of central tendency jointly when asking about the central tendency of a distribution.

Distributions: Absolute, Relative, and Cumulative Frequency

More on Distributions: Absolute and Relative Frequency

- Sometimes we want to ask how likely is one specific value of a variable relative to others. This is particularly relevant for variables that take on integer values (nominal and ordinal categorical variables).
- *Def.* **Absolute Frequency** of the value of a variable. The absolute frequency of a value is the count of the number of times that value occurs in the data set.
- *Def.* **Relative frequency** of the value of a variable. The relative frequency of a value, f_i, is the proportion of the total number of data points that that value, x_i, represents.
  - It is calculated by dividing the absolute frequency of the value by the total number of data points.

Pew Research Center's American Trends Panel (ATP)

- **Design and Sampling:** Multimode, probability-based panel with roughly 10,000 U.S. adults, selected randomly to ensure national representativeness.
- **Recruitment Method:** Initially recruited via random digit dialing (2014-2017), switched to address-based sampling (ABS) from the U.S. Postal Service's CDS file (2018-present).
- **Survey Modes:** Online surveys (computer, tablet, smartphone) and phone interviews with live interviewers, starting in 2024 to include phone survey options.
- **Weighting:** Multistep process to adjust for sampling stages and nonresponse, aligning survey samples with population benchmarks.

Example of a categorical ordinal surveyed in ATP Wave 116

- In ATP Wave 116, one of the questions posed to the participants was: "*How confident are you that votes cast by absentee or mail-in ballot across the United States will be counted as voters intend in the elections this November?*"
- The responses (excluding "No answer") are categorized into several confidence levels, allowing respondents to express their perceptions.

Confidence	Absolute Frequency
Not at all confident	407
Not too confident	611
Somewhat confident	967
Very confident	534

Example of a categorical ordinal surveyed in ATP Wave 116

- Using the absolute frequency of responses compute the relative frequencies.
- $N = 2519$

Confidence	Absolute Frequency	Relative Freq. (f_i)
Not at all confident	407	
Not too confident	611	
Somewhat confident	967	
Very confident	534	

Example of a categorical ordinal surveyed in ATP Wave 116

- Using the relative frequency of responses, compute the cumulative frequencies.

Confidence	Absolute Frequency	Relative Freq. (f_i)
Not at all confident	407	0.162
Not too confident	611	0.243
Somewhat confident	967	0.384
Very confident	534	0.212

More on Distributions: Cumulative Frequency

- When variables have a specific ordering relationship (either increasing/decreasing in magnitude or qualitative intensity), we can compute the cumulative distribution of the variable.
- *Def.* **Cumulative frequency** of the value of a variable. The cumulative frequency is the running total of the frequencies.
- *Def.* **Cumulative relative frequency** of the value of a variable (F_i). The cumulative frequency is the running total of the relative frequencies.
  - In other words, the cumulative relative frequency tells the sum of each proportion or percentage including and leading up to each data value

Example of a categorical ordinal surveyed in ATP Wave 116

- Using the relative frequency of responses, compute the cumulative relative frequencies.

Confidence	Type of Frequency Absolute	 Relative (f_i)	 Cumulative (F_i)
Not at all confident	407	0.162	0.16
Not too confident	611	0.243	0.40
Somewhat confident	967	0.384	0.79
Very confident	534	0.212	1.00

Simulation of 1,000 Throws of Four Coins

Introduction: In this experiment, we simulate 1,000 independent throws of four fair coins. Each throw results in a certain number of heads (from 0 to 4). We analyze the distribution of the absolute and relative frequencies of these outcomes.

Steps of the Experiment:

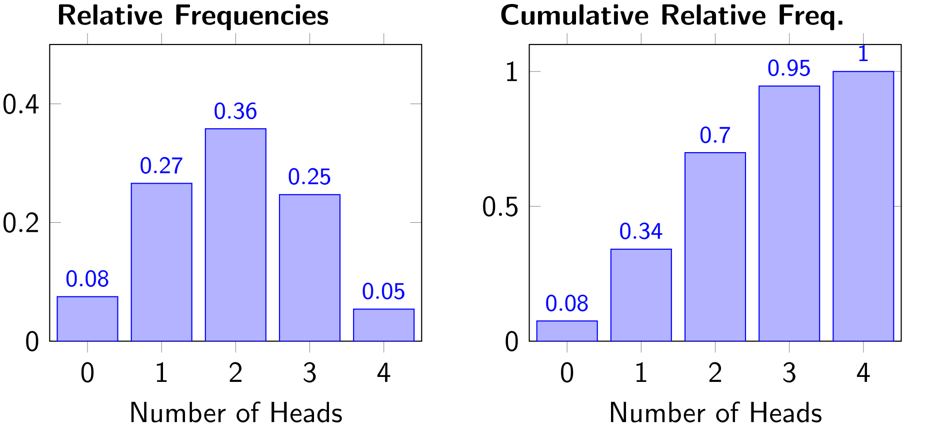
- Each throw consists of flipping 4 independent coins.
- We record the number of heads obtained in each throw.
- The results are summarized in terms of:
  - Absolute frequencies.
  - Cumulative absolute frequencies.
  - Relative frequencies.
  - Cumulative relative frequencies.

Simulation of 1,000 Throws of Four Coins

Results:

# Heads	Absolute Frequency	Cumulative Frequency	Relative Frequency	Cumulative Relative Frequency
0	75	75	0.075	0.075
1	266	341	0.266	0.341
2	358	699	0.358	0.699
3	247	946	0.247	0.946
4	54	1000	0.054	1.0

Relative and Cumulative Frequencies of the Total Number of Heads When Throwing Four Coins



The Weighted Mean

- Sometimes, values in a distribution carry different importance because their relative frequencies vary.
- Therefore, when computing the mean, we want to account for these different relative frequencies. To do so, we use the weighted mean.

Def. **Weighted Mean**. Consider a given value of the variable x_i, and let f_i be its relative frequency. The weighted mean is defined as:

$$\bar{X}_w = \sum_{i=1}^{k} f_i \times x_i$$

where k represents the number of distinct values in the distribution.

Example: The Weighted Mean

- Let's use the previous example using the ATP question to compute the weighted mean.
- We numerically code the ordinal confidence levels as follows:

$$x_i = \begin{cases} 1, & \text{if respondent is ``Not at all confident''} \\ 2, & \text{if respondent is ``Not too confident''} \\ 3, & \text{if respondent is ``Somewhat confident''} \\ 4, & \text{if respondent is ``Very confident''} \end{cases}$$

- Then, we compute the weighted mean of this numeric confidence scale.

Population Mean vs. Sample Mean

Population Mean (μ)

- Mean of all values in the population.
- Formula:

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- **Key:** μ is a fixed value (is the true average of the population at a given point in time).

Sample Mean ($\bar{x}$)

- Mean of values in a sample.
- Formula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- **Key:** $\bar{x}$ varies across samples.

Measures of Position

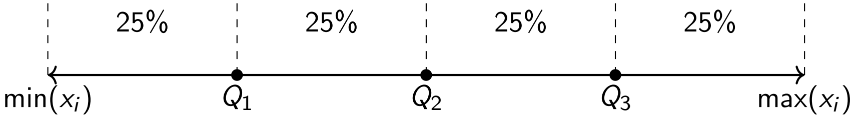
Measure of Position

- **Definition:** A measure of position identifies the location of a specific value within a data set relative to the overall distribution.
- **Purpose:** Helps in understanding how a particular data point compares to the rest of the data.
  - They provide insight into where a data point lies (e.g., near the center, in the tail, or at an extreme) and facilitate comparisons across data sets.
- **Examples:**
  - **Percentiles & Quantiles:** Values that split the data into equal-sized groups.

Quantiles

- **Definition:** Quantiles are values that divide observations into equal-sized intervals, such that a specified proportion of the data lies below each quantile.
- Formally, the p-th quantile Q_p of a variable is a numeric value such that the p proportion of the data is less than or equal to Q_p.
- Common examples include:
  - Quartiles: Divide the data into four equal parts.
  - Deciles: Divide the data into ten equal parts.
  - Percentiles: Divide the data into 100 equal parts.

Quartile Positions Along the Data Range



Quartiles Example: Income Distribution

Assume we have income data from a randomly selected sample. Below are the computed quartiles:

Quartile	Value	Interpretation
$\mathbf{Q_1}$	\$30,000	25% of individuals earn $\leq$ \$30,000.
$\mathbf{Q_2}$ (Median)	\$45,000	50% of individuals earn $\leq$ \$45,000.
$\mathbf{Q_3}$	\$82,000	75% of individuals earn $\leq$ \$82,000.

Note:

- If someone earns \$84,100, they are **above** Q_3 (or "*in* the third quartile"), meaning they earn more than 75% of the sample.
- If someone earns \$23,500, they are **below** Q_1 (or "*below* the first quartile"), meaning they earn less than 75% of the sample (or "are in the lowest 25%").

Deciles and Percentiles Example: Income Distribution

- **Deciles:** Divide the data into ten equal parts.
  - **Ninth Decile Example:**
    - 9th Decile = \$150,000: 90% of individuals earn less than or equal to \$150,000.
- **Percentiles:** Divide the data into one hundred equal parts.
  - **Ninety-Fifth Percentile Example:**
    - 95th Percentile = \$400,000: 99% of individuals earn less than or equal to \$400,000.
- **Interpretation:**
  - Being at the 9th decile means you earn more than 90% of the population.
  - Being at the 99th percentile means you earn more than 99% of the population.

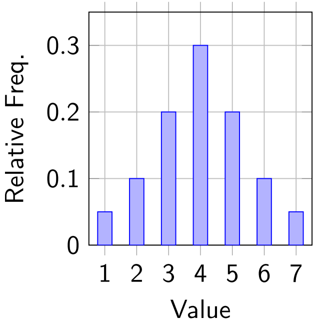
Measures of Dispersion

Introduction to Measures of Dispersion

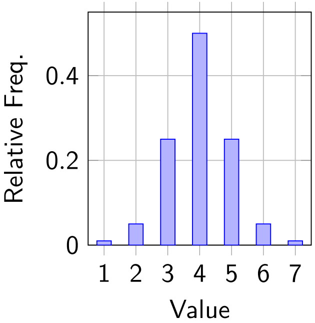
- **Definition:** Measures of dispersion describe a variable's the spread or variability.
- **Purpose:** While measures of central tendency (like the mean or median) tell us about the center of the data, measures of dispersion help us understand the distribution's spread and how much individual data points deviate from the center.

Comparing Two Discrete Distributions with the Same Mean

Distribution 1: Wider Spread



Distribution 2: Narrower Spread



Introduction to Measures of Dispersion

- **Range:** The difference between the maximum and minimum values.

$$\text{Range} = \max\{x_i\} - \min\{x_i\}$$

- **Interquartile Range (IQR):** The range between the first and third quartiles, highlighting the spread of the middle 50% of the data.

$$\text{IQR} = Q_3 - Q_1$$

Introduction to Measures of Dispersion

- **Variance:** The average of the squared deviations from the mean.
- **Standard Deviation:** The square root of the variance; representing the average deviation from the mean.
- **Mean Absolute Deviation:** The average of the absolute deviations from the mean

$$\text{MAD} = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N}$$

  - **Note:** While the MAD can be a useful statistic to describe the variation in the data, we almost always focus on the Standard Deviation instead because of its useful properties for inference.

Intuition for Squaring in Variance

Distance Between Two Points in 2D

The Euclidean distance between two points (x_i, y_i) and (μ_x, μ_y) is given by:

$$d = \sqrt{(x_i - \mu_x)^2 + (y_i - \mu_y)^2}.$$

Why Squaring?

- Squaring ensures all differences are positive, preventing cancellations.
- It naturally arises from the **Pythagorean theorem**, which measures true distance.
- Variance follows the same principle: it measures spread using squared differences from the mean.
- The standard deviation (square root of variance) gives a measure in the original units, just like distance.

Population vs. Sample Variance

Population Variance

- Denoted by: σ^2
- Formula:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- Where:
  - N is the size of the population.
  - x_i is each individual value in the population.
  - μ is the population mean.

Sample Variance

- Denoted by: s^2
- Formula:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Where:
  - n is the size of the sample.
  - x_i is each individual value in the sample.
  - $\bar{x}$ is the sample mean.

Population vs. Sample Standard Deviation

Population Standard Deviation

- Denoted by: σ
- Formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

- Measures the average distance of each data point from the population mean μ.

Sample Standard Deviation

- Denoted by: s
- Formula:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Measures the average distance of each data point from the sample mean $\bar{x}$.

Variance vs. Standard Deviation

Why is Standard Deviation More Useful than Variance?

- **Interpretability:** Standard deviation is in the same units as the original data, making it easier to interpret.
- **Comparison:** It allows for easier comparison between different datasets or distributions.
- **Practicality:** Many statistical methods and models use standard deviation rather than variance for these reasons.

Why the $N - 1$ in the Sample Variance?

- **Correcting for Bias:**
  - To correct for this bias, we divide by $n - 1$ instead of n.
  - This adjustment, known as **Bessel's correction**, increases the variance slightly, providing an unbiased estimate of the population variance.
  - By dividing by $n - 1$, we account for the fact that the sample mean $\bar{x}$ is an estimate and not the true population mean μ.

- **Degrees of Freedom:**
  - The use of $n - 1$ reflects the concept of **degrees of freedom**, which represents the number of values in the final calculation that are free to vary.
  - Since one degree of freedom is "lost" by using the sample mean, only $n - 1$ independent pieces of information remain.

Coefficient of Variation (CV)

- **Definition:** The Coefficient of Variation (CV) is defined as the ratio of the standard deviation (σ) to the mean (μ):

$$\text{CV} = \frac{\sigma}{\mu}$$

- Often expressed as a percentage:

$$\text{CV} = \frac{\sigma}{\mu} \times 100\%$$

- **Relative Measure:** CV provides a standardized measure of dispersion relative to the mean, allowing comparison between datasets with different units or scales.
- **Interpretation:** A higher CV indicates greater relative variability, while a lower CV suggests more consistency around the mean.

Usefulness of the Interquartile Range (IQR)

- **Definition:** The Interquartile Range (IQR) is the difference between the third quartile (Q_3) and the first quartile (Q_1):

$$\text{IQR} = Q_3 - Q_1$$

- **Key Benefits:**
  - **Robustness:** IQR is not affected by outliers or extreme values, making it a robust measure of spread.
  - **Focus on Middle 50%:** IQR provides insight into the spread of the central 50% of the data, highlighting the range within which the bulk of the data lies.
  - **Comparison of Distributions:** IQR allows for easy comparison of variability between different datasets or distributions.
  - **Identifying Outliers:** IQR is used to identify outliers; values that fall below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times \text{IQR}$ are often considered outliers.

Introduction to Statistical Methods in Political Science

Lecture 3: Summarizing Data II - Graphs

Ignacio Urbina

Shape of a Distribution

Distribution of a Variable

A variable's distribution is a description (function) that shows its possible values and the frequency with which they occur.

By examining a variable's distribution, we learn how likely is a given value relative to others.

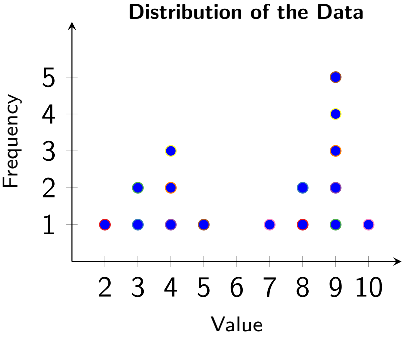
Dot Plot Example with Discrete Numerical Data

Sample:

$\{$ 9, 10, 4, 3, 8, 9, 3, 9, 5, 2, 4, 4, 8, 9, 9, 7$\}$

Frequency Table:

Value	Frequency
2	1
3	2
4	3
5	1
6	0
7	1
8	2
9	5
10	1



Identifying Distribution Shapes

Key Distribution Shapes:

- **Symmetric:** Equal spread on both sides.
- **Right Skewed:** Tail on the right (positive skewness).
- **Left Skewed:** Tail on the left (negative skewness).

Key Modalities:

- **Unimodal:** Single peak.
- **Bimodal:** Two distinct peaks.
- **Multimodal:** More than two peaks.
- **Uniform:** Flat, no peaks.

Symmetry and Skewness

Symmetric Distributions:

- The distribution looks the same on both sides of the mean.

Skewed Distributions:

- **Right skewed:** Long tail on the right.
- **Left skewed:** Long tail on the left.

The direction of skewness indicates where most of the data points lie relative to the tail.

Skewness and the Mean-Median Relationship

- In a **right-skewed distribution**, the mean is typically greater than the median due to the influence of the long tail.
- In a **left-skewed distribution**, the mean is typically less than the median for similar reasons.
- In a **symmetric distribution**, the mean and median are roughly equal.

Summary:

- The relative position of the mean and median can indicate the skewness of a distribution.

Unimodal, Bimodal, Multimodal, and Uniform Distributions

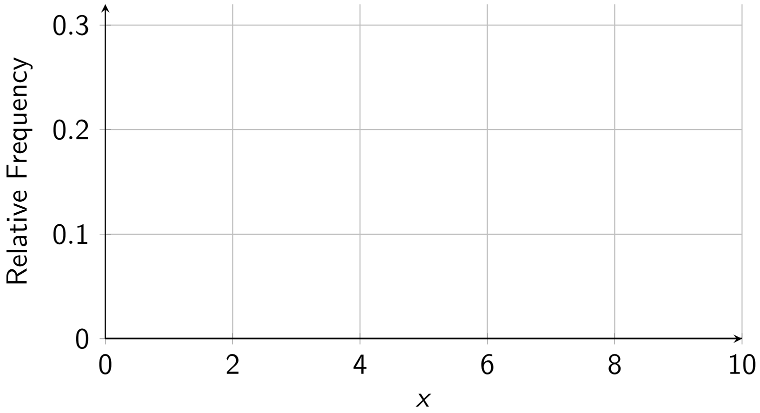
- **Unimodal:** A distribution with one clear peak or mode.
- **Bimodal:** A distribution with two distinct peaks.
- **Multimodal:** More than two peaks, indicating multiple clusters or groups.
- **Uniform:** No peaks; all values have roughly the same frequency.

These characteristics help describe the overall shape of the data and can indicate the presence of subpopulations.

Right-Skewed Distribution Example

Right-Skewed Distribution:

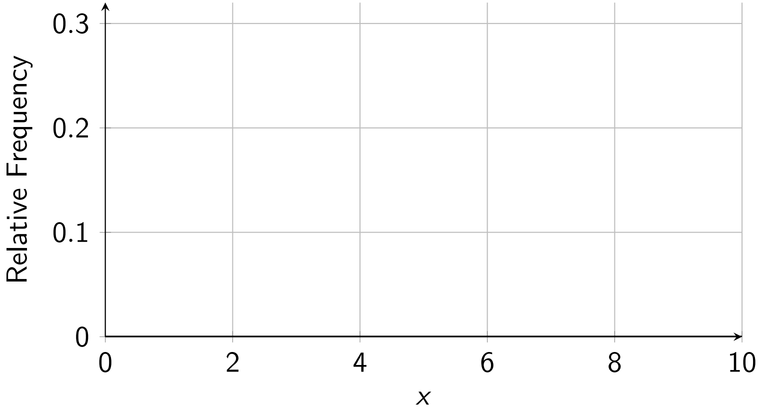
- The tail extends to the right, meaning more data is concentrated on the left.



Left-Skewed Distribution Example

Left-Skewed Distribution:

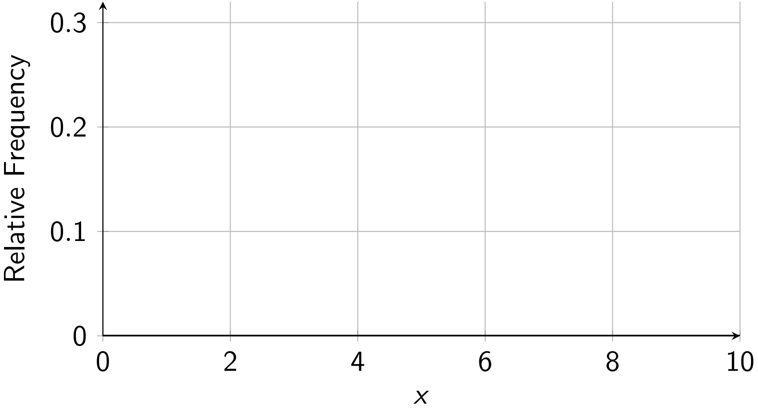
- The tail extends to the left, meaning more data is concentrated on the right.



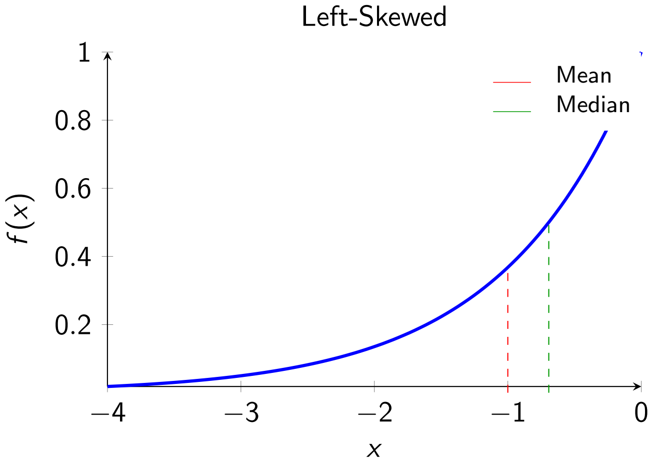
Symmetric Distribution Example

Symmetric Distribution:

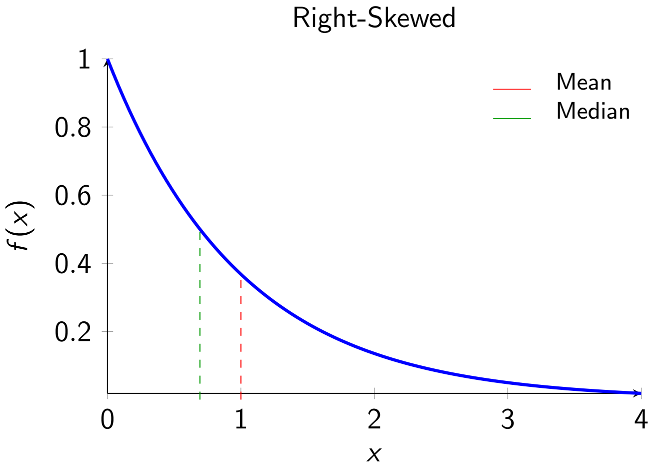
- A symmetric distribution has equal spread on both sides of the mean.



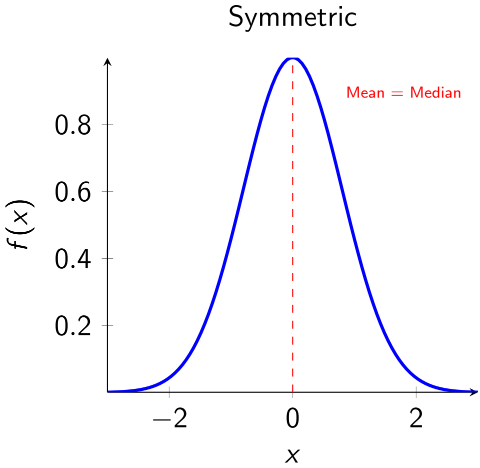
Left-Skewed Distribution



Right-Skewed Distribution



Symmetric Distribution



A symmetric distribution is a distribution in which the shape on the left and right sides of the center are mirror images of each other, meaning that the frequencies change at the same rate and direction as one moves away from the center in both directions.

Histograms

Introduction to Histograms

What is a Histogram?

- A histogram is a graphical representation of the distribution of numerical data.
- It uses bins (or intervals) to group data points and shows the frequency of data points in each bin.
- The height of each bar represents the count or frequency of data within that bin.

Step 1: Understanding Absolute Frequency

Absolute Frequency:

- The absolute frequency of an interval is the number of data points that fall within that interval.
- It's simply a count of occurrences of data points in each interval.

Example:

- Consider the data: $\{3, 5, 7, 9, 11, 5, 7, 9, 3, 5\}$
- The absolute frequency of the interval [3, 6] is 5 (as four data points fall in this interval: 3, 3, 5, 5, 5).

Step 2: Choosing Bins (Intervals)

What are Bins?

- Bins (or intervals) divide the entire range of values into equal-sized chunks.
- Each bin contains a specific range of values, and data points are placed into the bin they fall within.

General Definition:

- *First* bin: $[\text{min}, \text{min} + \text{bin\_size})$
- Bin k: $[\text{min} + (k-1) \cdot \text{bin\_size}, \text{min} + (k) \cdot \text{bin\_size})$
- *Last* bin: $[\text{max} - \text{bin\_size}, \text{max}]$
  - It's normal to close the last bin on the right to capture the maximum.

Step 3: Assigning Data to Bins

Assigning Data to Bins:

- Once bins are defined, we assign each data point to the appropriate bin.
- This process results in an absolute frequency count for each bin.

Example:

- Suppose a discrete variable that goes from 1 to 15.
- Suppose we have 3 bins: [1, 6), [6, 11), [11, 15].
- If the data is $\{3, 7, 9, 12, 3, 5\}$, then:
  - Bin [1, 6): 3 data points $\{3, 3, 5\}$
  - Bin [6, 11): 2 data points $\{7, 9\}$
  - Bin [11, 15]: 1 data point $\{12\}$

Step 4: Constructing the Histogram

Constructing the Histogram:

- Now that we have the counts (absolute frequencies) for each bin, we can construct the histogram.
- The x-axis represents the bin intervals, and the y-axis represents the count of data points in each bin.
- Draw a bar for each bin where the height corresponds to the count (absolute frequency).

Example: Data Table

Step-by-Step Example: Consider the following data set:

3	5	7	9	11	3	4	6	8	10
12	15	1	2	14	5	7	9	11	13

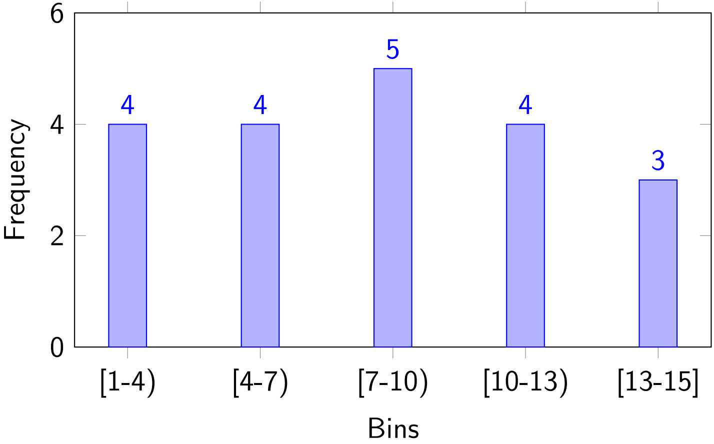
We will construct a histogram with 5 bins.

Example: Histogram with 5 Bins

Step-by-Step Example:

- Range of data: 1 to 15
- Bin size $= \frac{15-1}{5} = 2.8$ (rounded to 3)
- Bins:
  - $[1, 4)$: 4 data points $\{1, 2, 3, 3\}$
  - $[4, 7)$: 4 data points $\{4, 5, 5, 6\}$
  - $[7, 10)$: 5 data points $\{7, 7, 8, 9, 9\}$
  - $[10, 13)$: 4 data points $\{10, 11, 11, 12\}$
  - $[13, 15]$: 3 data points $\{13, 14, 15\}$

Example: Histogram with 5 Bins



Bar Plots

When Do We Use Bar Plots?

Bar Plots:

- Bar plots are used to display the frequency or proportion of categorical data.
- They are especially useful for comparing different categories.
- Each bar represents a category, and the height of the bar represents the value (frequency, proportion, etc.).

Use Cases:

- Visualizing survey responses.
- Comparing the frequency of different groups (e.g., gender, age groups).

What is a Frequency Distribution?

Frequency Distribution:

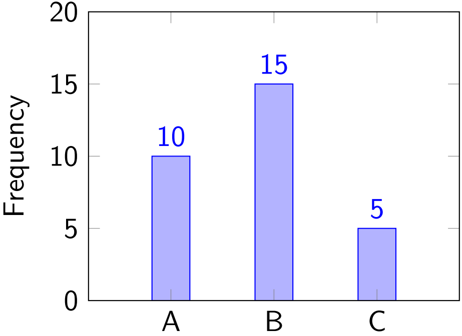
- A frequency distribution is a table that shows the frequency (or count) of each value or category.
- It can be visualized using a bar plot where each category corresponds to a bar.

Example:

Category	Frequency
A	10
B	15
C	5

Basic Bar Plot Structure

- The x-axis represents categories (e.g., different groups).
- The y-axis represents the frequency, proportion, or value corresponding to each category.
- Bars can be vertical or horizontal.



What is Cross-Tabulation?

Cross-Tabulation (Crosstab):

- Cross-tabulation is a method used to analyze the relationship between two categorical variables.
- It creates a matrix (or table) that shows the frequency distribution of the variables across their categories.

Example:

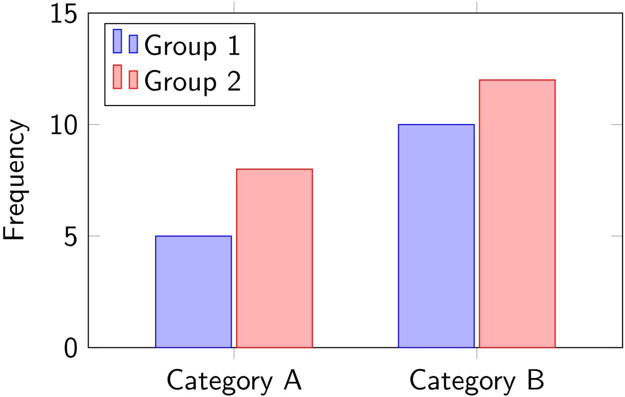
	Group 1	**Group 2**
Category A	5	8
Category B	10	12

Using Bar Plots for Cross-Tabulation

Bar Plots for Two Variables:

- Bar plots can also represent cross-tabulation by plotting grouped bars.
- Each group (e.g., Group 1, Group 2) has its own bar for each category.

Example:

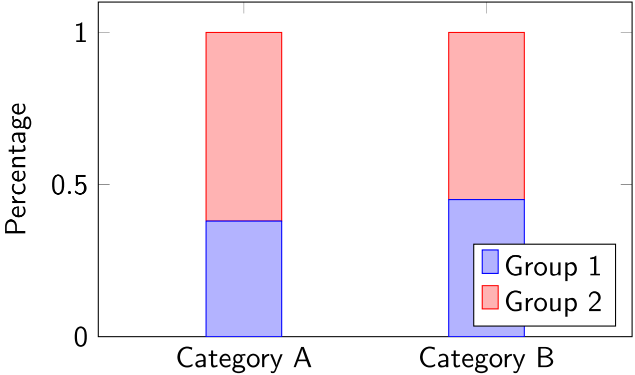


Using Bar Plots for Cross-Tabulation

Stacked Bar Plots for Two Variables:

- Bar plots can also be *stacked* within each category.
- We often do this when we represent percentages in the *y*-axis.

Example:



Box Plots

Introduction to Box Plots

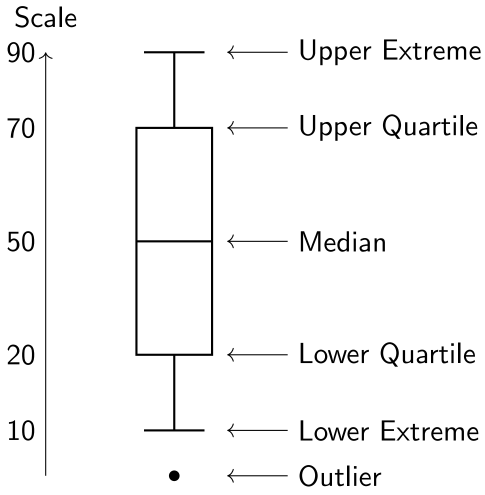
- A box plot (or box-and-whisker plot) displays the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable.
- The box shows the quartiles of the dataset while the whiskers extend to show the rest of the distribution, except for points that are determined to be "outliers" using a method that is a function of the inter-quartile range.
- Box plots give a clear summary of data distribution and variability and are particularly useful for highlighting outliers and for comparing distributions across groups.

Constructing a Box Plot

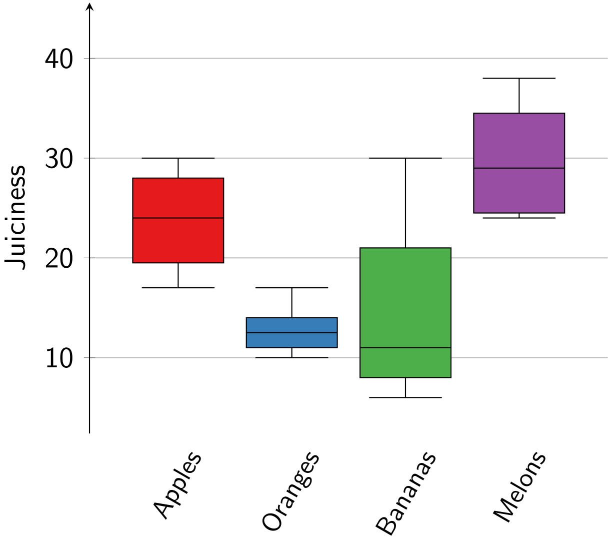
Steps to Construct a Box Plot:

1. Calculate the first (Q1) and third quartiles (Q3).
2. Find the interquartile range (IQR = Q3 - Q1).
3. Determine the “whiskers” which are typically set at 1.5 * IQR above Q3 and below Q1. Data points outside this range are considered outliers.
4. The median (Q2) is marked by a line inside the box. If a distribution is skewed, then the median will not be in the middle of the box, and instead off to the side.

Structure of a Box Plot



Example: Box Plot with Multiple Categories



Dot Charts

Introduction to Dot Charts

What is a Dot Chart?

- A dot chart is a statistical chart consisting of data points plotted on a simple scale.
- It is used to compare frequency, count, or any measure across different categories.
- Similar to bar charts but dots are used instead of bars, making it less cluttered.

Why Use Dot Charts?

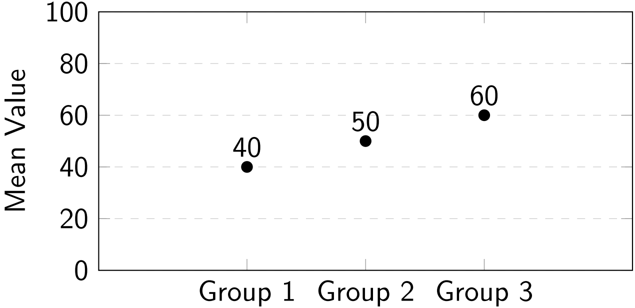
- **Clarity:** Provides a clear and precise representation of data points.
- **Comparison:** Facilitates easy comparison of multiple groups.
- **Space-efficient:** More effective in space usage than bar graphs.

Comparing Means Across Groups Using Dot Charts

How to Use Dot Charts for Comparing Means

- Dot charts are excellent for displaying the mean (or any central tendency) of different groups.
- Each dot represents the mean of a group, aligned along a single axis.
- The position of each dot on the scale directly reflects the value of the mean, making comparisons intuitive.

Example:



Scatter Plot

Introduction to Scatter Plots

What is a Scatter Plot?

- A scatter plot is a type of data visualization that shows the relationship between two numerical variables.
- Each point on the plot represents a pair of values: one on the x-axis and one on the y-axis.
- Scatter plots are useful for identifying correlations, trends, and outliers in data.

Applications of Scatter Plots

- Visualizing correlations between variables.
- Spotting clusters and patterns.
- Detecting outliers and anomalies.

How to Create a Scatter Plot

Steps to Create a Scatter Plot:

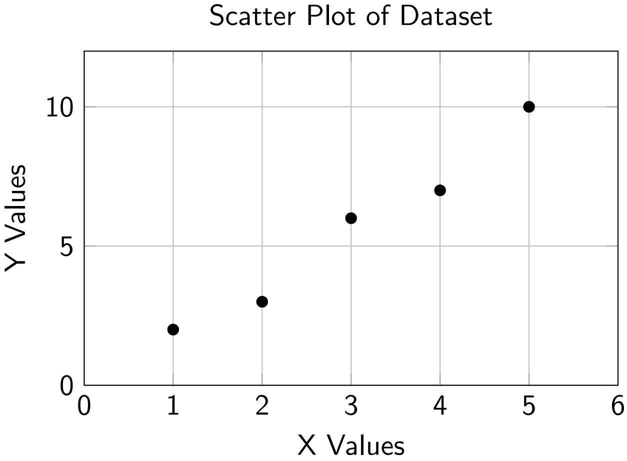
- **Step 1:** Gather your data points. You need two variables, one for the x-axis and one for the y-axis.
- **Step 2:** Plot each pair of values on a coordinate system.
- **Step 3:** Optionally, add labels, grid lines, and color coding to enhance interpretation.

Example:

- Dataset: (1,2), (2,3), (3,6), (4,7), (5,10)
- The scatter plot will show these points in a linear relationship.

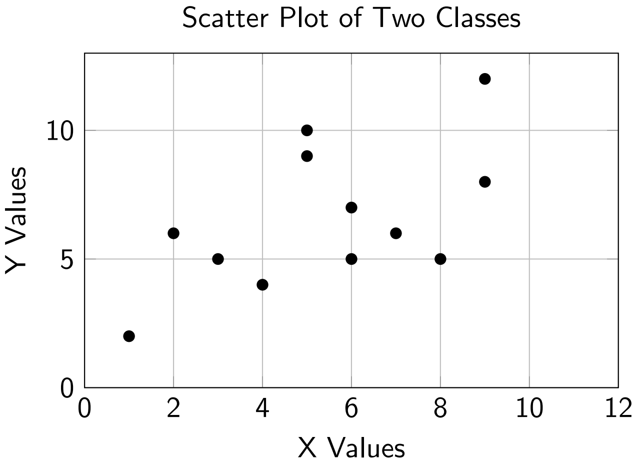
Scatter Plot Example (Basic)

Example of a Scatter Plot:



Scatter Plot Example with Two Classes

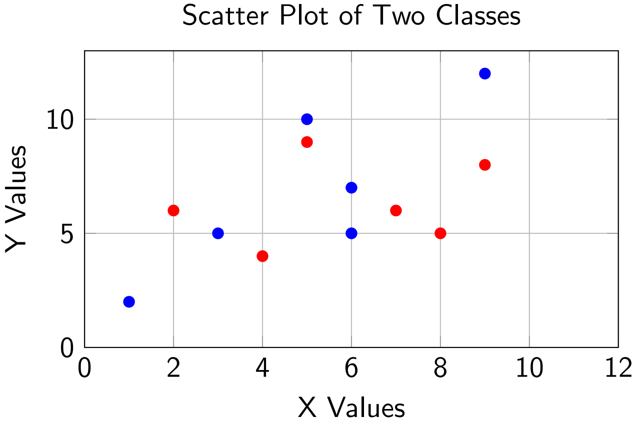
Imagine our dataset has two classes (Class A and Class B).



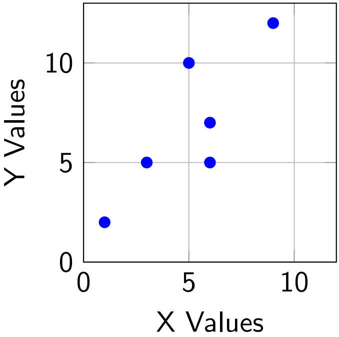
Scatter Plot Example with Two Classes

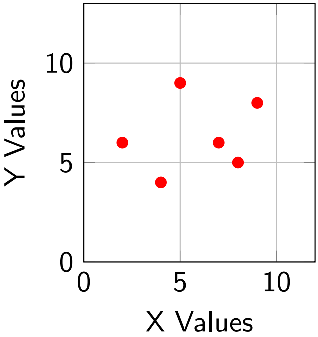
We have two different classes (Class A and Class B), each represented by different colors:

- Blue: Class A
- Red: Class B



Side-by-Side Comparison





Summary and Takeaways

Summary: Distributions and Plots

Key Takeaways:

- **Shape of Distribution:** Understand skewness (right, left) and mode (unimodal, bimodal, multimodal).
- **Histogram:** Useful for visualizing the distribution of intervals or bins of data.
- **Bar Plot:** Best for categorical data comparisons.
- **Box Plot:** Ideal for displaying the distribution's quartiles and identifying outliers.
- **Dot Chart:** Great for comparing multiple categories without the clutter of bars.
- **Scatter Plot:** Essential for identifying relationships and trends between two variables.

Introduction to Statistical Methods in Political Science

Lecture 4: Introduction to Probability

Ignacio Urbina

Ph.D. Candidate in Political Science

Random Processes, Outcomes, and Events

Why Study Probability?

- We want to understand how likely an event is.
- Compare events to say which is more likely.
- When we say an event is more likely, we expect it to happen more than less likely events.
- Likelihood and probability reflect our belief about the chances of events occurring.
- Examples:
  - We can say very confidently that the sun will rise tomorrow or "*there is a very high probability the sun will rise tomorrow*."
  - We can say very confidently that in the middle of summer, it won't snow or "*There is a low probability of snow in the middle of summer*."

A Solid Understanding of Probability

- Provides a framework to deal with uncertainty and randomness.
- Our expectations about likely events influence our actions and planning.
- Examples:
  - Planning an outdoor event based on the weather forecast.
  - Buy a stock based on its price forecast.
  - Settle on a court case based on the probability of losing the lawsuit.
- We study probabilities of **random processes**.

What is a Random Process?

Random Process (*Def.*)

A random process is one where the outcome is uncertain before it happens.

- The outcome is drawn from a set of all possible outcomes.
- Examples:
  - Weather: We cannot predict with certainty if it will rain tomorrow.
  - Throwing a die: The result of a die throw is unknown until it happens.
  - Elections: The winning candidate is uncertain until votes are counted.

What is a Random Process?

- What do we mean by “process”?
- **Process:**
  - A process is a mapping from an initial state to a later state, i.e., a sequence of connected steps.
  - Example: Applying for a scholarship (learning about it, writing the application, submitting it, getting the result).
  - Initial conditions lead to a new condition through a series of steps.
- **Random Process:** A random process involves initial conditions leading to an uncertain outcome.

Example: Election as a Random Process

- Initial conditions: Current political climate, economic conditions, media coverage.
- These factors influence voter preferences in complex ways.
- Outcome: Election result (winning candidate).
- Unpredictable events during the campaign can change voter behavior.
- New information, debates, and events can shift the final outcome.

Understanding Outcomes

Outcome (*Def*.)

An outcome is one realization of a random process.

- Examples:
  - Rain or no rain tomorrow.
  - Rolling a one on a die.
  - "Candidate A" winning the election.

What is an Event?

Event (*Def.*)

An event is a collection of outcomes from a random process.

- Event groups outcomes together so that we can study the probability of broader conditions and scenarios rather than just isolated cases.
- Events can represent conditions that occur in various contexts and times (e.g., different days, multiple experiments).

Event vs. Outcome

- An outcome is a single realization of a random process.
- An event is a set of possible outcomes.
  - **Rolling a die**:
    - *Outcome:* Getting a 6 when throwing a die.
    - *Event:* Getting an even number when rolling a die (2, 4, or 6).
  - **Electoral Results**:
    - *Outcome:* Candidate A wins narrowly.
    - *Event:* Candidate A wins (including both "Wins by a landslide" and "Wins narrowly").

Example: Forecast Scenarios 2024 Election

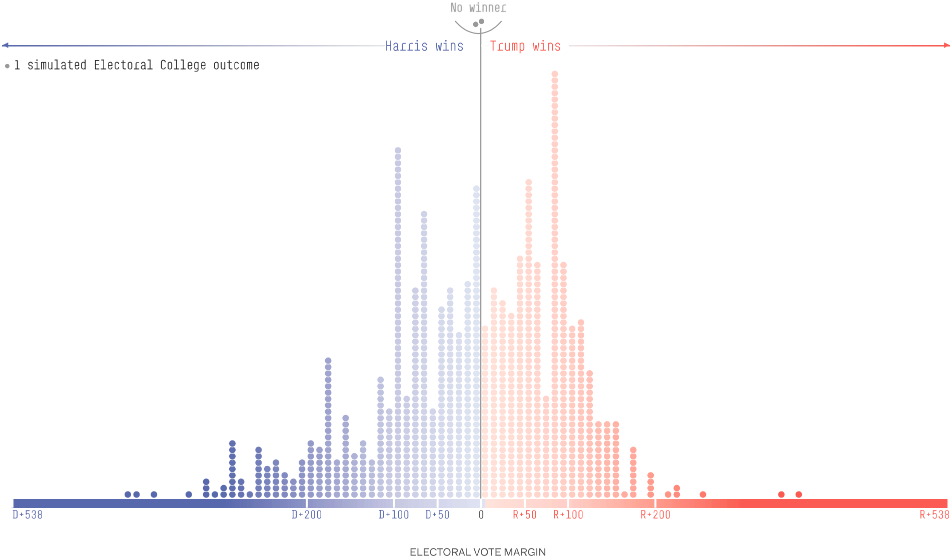


Figure: Source: 538

Disjoint Events

- **Disjoint (Mutually Exclusive) Events:** Two events are disjoint if they do not share any outcomes.
  - Example: Rolling a die
    - Event A: Rolling an even number $\{2, 4, 6\}$
    - Event B: Rolling an odd number $\{1, 3, 5\}$
    - These events are disjoint because they do not share any outcomes.
  - Example: Election (First-past-the-post voting system)
    - Event A: Any candidate from Party X wins $\{X_1, X_2, \cdots\}$
    - Event B: Any candidate from Party Y wins $\{Y_1, Y_2, Y_3, \cdots\}$
    - These events are disjoint because they do not share any outcomes.

Join Events (Not Disjoint)

- **Joint Events (can occur together):** Two events are joint if they share at least one outcome.
  - Example: Weather
    - Event A: It rains tomorrow $\{Rain\}$
    - Event B: It is windy tomorrow $\{Windy\}$
    - These events are joint since the outcome is $\{Rain, Windy\}$ can happen.
  - Example: Rolling two dice
    - Event A: Rolling a 2 on the first die $\{2\}$
    - Event B: Rolling a 4 on the second die $\{4\}$
    - These events are joint because the combined outcome $\{2, 4\}$ can occur.

Set Theory Basics

Set Theory - Basic Definitions

Definition: A set is a collection of well-defined, unordered objects called elements or members.
Explicit Set Definition: A (finite) set can be defined by explicitly specifying all of its elements between curly braces, known as set braces $\{\}$.
For example,

- $A = \{1, 2, 3, 4, 5, 6\}$
- $B = \{a, e, i, o, u\}$
- $C = \{US, UK, FRANCE, CANADA, CHINA, ...\}$
- $D = \{h, t\}$ (outcomes of a coin toss: heads, tails)
- $S = \{\{h, h\}, \{h, t\}, \{t, h\}, \{t, t\}\}$ (outcomes of two coin tosses)

Set Theory - Basic Definitions

Review:

- The null set (or empty set) is denoted by $\emptyset$.
- The union of sets $C \cup D$ includes elements in C, D, or both.
- The intersection of sets $A \cap B$ includes elements common to both A and B.
- The complement of a set $\neg D$ or D^c includes all elements not in D.
- Mutually exclusive (or disjoint) events are sets with no common elements.
- A series of exhaustive events cover all possible outcomes in the sample space.
- The universe set, U, is the set that contains all possible elements.

Set Theory Basics

Define the sets

- $A = \{\text{Banana, Apple, Orange, Watermelon}\}$,
- $B = \{\text{Orange, Plum, Grapes, Apple}\}$.
- Suppose the universal set is $U = \{\text{Banana, Apple, Orange, Watermelon, Plum, Grapes, Lemon}\}$.

Then,

- $A \cup B = \{\text{Banana, Apple, Orange, Watermelon, Plum, Grapes}\}$
- $A \cap B = \{\text{Apple, Orange}\}$
- Complement of $A = A^c = U - A = \{\text{Plum, Grapes, Lemon}\}$
- $A \cup A^c = U$.

Probability Definitions, Axioms, and Probability Distribution

Definitions of Probability

- **Frequentist Definition**

Definition

The probability of an event is the proportion of times the event would occur if we observed the random process an infinite number of times.

- Probability is defined as the long-run frequency of an event occurring.
- Example: Flipping a coin many times and observing the proportion of heads.

Definitions of Probability

- **Classical Definition**

Definition

The probability of an event is the number of ways it can happen divided by the total number of possible outcomes, assuming all outcomes are equally likely.

- Example: Rolling a fair six-sided die, the probability of getting a number lower than 3 (so, either 1 or 2) is $\frac{1+1}{6} = \frac{2}{6}$.
- Note that this definition is only valid when all outcomes are equally likely.
- Therefore, is it useful to calculate probabilities for the outcome of a fair die $\{1, 2, \cdots, 6\}$, but not for the outcome of a single specific election $\{\text{Candidate A Wins}, \text{Candidate B Wins}\}$.

Challenges with One-Shot Events

- Many events cannot be repeated under identical conditions.
  - Example: A specific election outcome.
- Frequentist and classical definitions struggle with one-shot events.
  - What does it mean to repeat an election?
  - How do we count all possible outcomes for a unique event?
- We need additional assumptions in these cases.
- How we define our population becomes crucial in calculating probabilities.
- This often means we need to make our research question more general.

Classical Probability in the Context of Elections

- **Calculating Probabilities in Electoral Contexts**:
  - Collect data from a large number of past elections within a given country or state.
  - Ensure these elections are comparable (e.g., same country/state, similar conditions).
  - Calculate the probability of a candidate with specific characteristics (e.g., gender, age, race) getting elected.
- **Example**:
  - **Suppose** we have data from 200 **comparable elections**. Define this as the **population**.
  - In 55 of these elections, a woman candidate won.
  - Probability of a woman candidate winning *in our population of elections*:

$$\frac{\text{Number of Favorable Cases}}{\text{Total Number of Cases}} = \frac{110}{200} = 0.55$$

Why it's important to learn probability to understand statistics

Suppose that the Office of Student Life claims that 45% of Stony Brook students are registered to vote in the next presidential election. Then, suppose we take a random sample of 100 students and determine that the proportion registered is:

$$\frac{53}{100} = 0.53 \qquad \text{(That is, 53\%)}$$

Now, we need to ask the question:

- *If the actual population proportion is 0.45, how likely is it that we'd get a sample proportion of 0.53?*
- If the answer is "*fairy likely*," the initial claim is reasonable. If not, we reject it.

Learning about probability allows us to answer such questions.

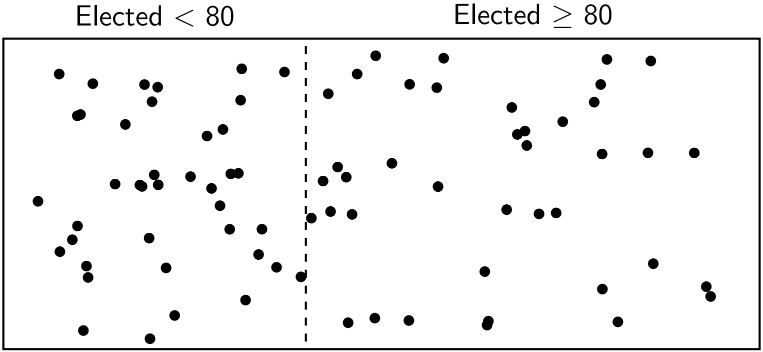
Definitions

Sample Space. The sample space (or outcome space), denoted S, is the collection of all possible outcomes of a random process (random experiment).
Event. Denoted with capital letters $A, B, C, \ldots$ — is just any subset of the sample space S. For example, $A \subset S$, where "$\subset$" denotes "is a subset of."
Probability of an Event. Let A be an event, such that A includes a subset of outcomes taken from S. Then,

$$\text{Probability of A} = \frac{\text{Total Cases in A}}{\text{Total Cases in S}}$$

Classical Probability in the Context of Elections

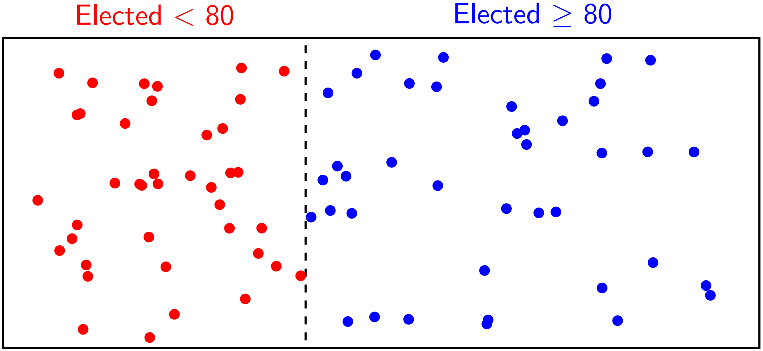
- Population: All elections in the United States in the last 20 years.
- Divided into two parts:
  - Elections in which someone older than 80 was elected.
  - Elections in which someone younger than 80 was elected.
- Dots represent individual cases.



Classical Probability in the Context of Elections

- What is the probability that a candidate younger than 80 years old is elected?

$$P(Elected < 80) = \frac{\#\ (\text{Favorable Cases})}{\#\ (\text{Total Cases})} = \frac{\text{Count of Red Dots}}{\text{Total Number of Dots}}$$



Three Axioms of Probability - Introduction

Probability Function

Probability is a real-valued function P that assigns to each event A in a sample space S a number known as the **probability of the event** A, denoted by $P(A)$.

An axiom in mathematics is a foundational statement assumed to be true, serving as a building block for the branch of mathematics in question. In probability and statistics, our theory is based upon three axioms.

Three Axioms of Probability - Detailed

The function P satisfies the following properties (axioms):

1. $P(A) \geq 0$ for any event A. (Nonnegativity)
2. $P(S) = 1$, where S is the sample space. (Certainty)
3. For any sequence, $A_1, A_2, \ldots A_k$, of mutually exclusive events, i.e., $A_j \cap A_i = \emptyset$, ,

   $$P\left(A_1 \cup A_2 \cup A_3 \cdots \cup A_k\right) = P(A_1) + P(A_2) + P(A_3) + \cdots + P(A_k)$$

   (Additivity)

These axioms form the basis of probability theory, providing a foundation for all subsequent probability rules.

Probability Distributions for Finite Sample Spaces

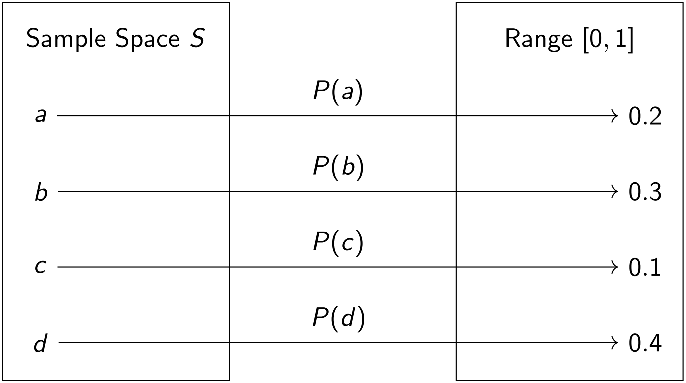
Probability distributions describe how probabilities are assigned across all possible outcomes in a finite sample space.

Definition: A probability distribution assigns a probability $P(a)$ to each outcome a in the sample space such that:

- $P(a) \geq 0$ for each outcome a
- The sum of probabilities for all outcomes is 1, $\sum_{a \in S} P(a) = 1$

The symbol $\in$ is used to construct a statement in set theory notation, and it means "belongs to" or "is an element of." Imagine the set $S = \{a, b, c, d\}$, therefore the statement "$a \in S$" is true because a is an element of S.

Probability Mapping: $P : S \to [0, 1]$



Probability Problem: Number of Heads When Throwing Two Coins

Problem Statement:

- We toss **two fair coins**.
- We define an **event** as a specific **number of heads** observed.
- Our goal is to determine the probability distribution of the different realizations of this outcome.

Sample Space: The possible outcomes of flipping two fair coins are:

$$S = \{HH, HT, TH, TT\}$$

where:

- HH: Both coins land on heads.
- HT and TH: One coin lands on heads, the other on tails.
- TT: Both coins land on tails.

Computing the Probability Distribution of X

Step 1: Count the Outcomes for Each Event

- $X = 0$ (No heads): $\{TT\} \rightarrow 1$ outcome
- $X = 1$ (One head): $\{HT, TH\} \rightarrow 2$ outcomes
- $X = 2$ (Two heads): $\{HH\} \rightarrow 1$ outcome

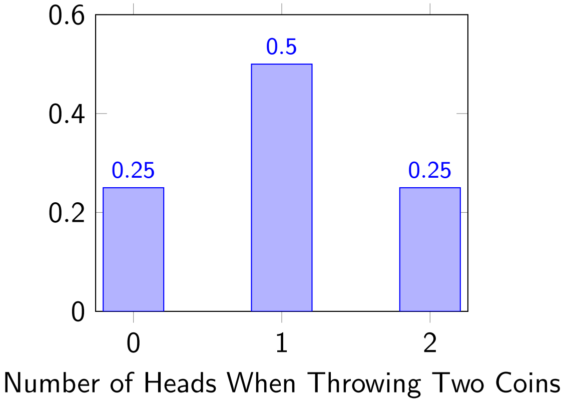
Step 2: Compute Probabilities

X (Number of Heads)	Favorable Outcomes	Probability $P(X)$
0	TT	$1/4 = 0.25$
1	HT, TH	$2/4 = 0.50$
2	HH	$1/4 = 0.25$

Final Answer: The probability distribution of the event is:

$$\text{P(Zero Heads)} = 0.25, \quad \text{P(Just One)} = 0.50, \quad \text{P(Two Heads)} = 0.25$$

Probability Distribution of the Total Number of Heads When Throwing Two Coins



Setting: Probability Distribution of Discrete Outcomes

- We analyze hypothetical (simulated) historical data of country pairs.
- Each combination of variables (Democratic Regimes, Economic Interdependence, War) for a pair of two countries is an outcome.
- We aim to calculate the probability distribution of these discrete outcomes.
- Variables:
  - **Both Countries Are Democratic:** Yes/No (No = if at least one county of the pair is not a democracy)
  - **Economic Interdependence:** High/Low (Do the two countries share strong economic ties and transactions?)
  - **War:** Yes/No (Have the two countries ever been in a war against each other?)

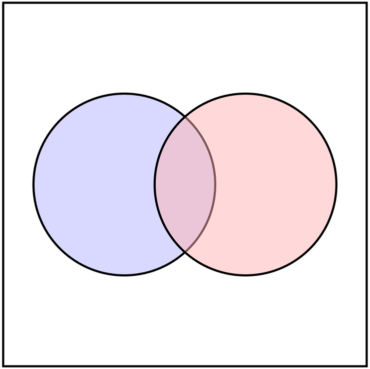
Summary of Dyadic (pairwise) Interactions

Both Democratic	Economic Interdependence	War	Count	Prob.
No	Low	Yes	23	0.12
No	Low	No	82	0.43
No	High	Yes	4	0.02
No	High	No	26	0.14
Yes	Low	Yes	5	0.03
Yes	Low	No	35	0.18
Yes	High	No	15	0.08
Yes	High	Yes	0	0.00

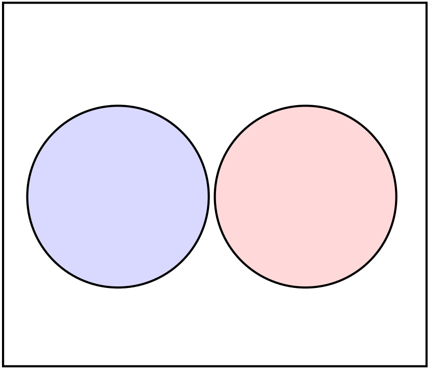
Table: Probability distribution of randomly selecting one country pair combination. Total country pairs = 190.

Venn Diagrams within the Universal Set

Non-empty Intersection



No Intersection



Probability of the Intersection of Two Events

Probability of the Intersection of Two Events (*Def.*)

Following the classical definition, the probability that the event A happens AND the event B happens is given by:

$$P(A \cap B) = \frac{\#(\text{Both } A \text{ and } B \text{ happen})}{\#(\text{Total Cases in } S)}$$

Probability of the Union of Two Events

Probability of the Union of Two Events (*Def.*)

Following the classical definition, the probability that the event A happens OR the event B happens is given by:

$$P(A \cup B) = \frac{\#(A \text{ and not } B) + \#(B \text{ and not } A) + \#(\text{Both } A \text{ and } B)}{\#(\text{Total Cases in } S)}$$

Probability Rules

Addition Rule for Disjoint Events

The addition rule for disjoint (mutually exclusive) events states that if two events A and B cannot occur at the same time, then the probability of A or B occurring is the sum of their probabilities.

$$P(A \cup B) = P(A) + P(B)$$

Example:

- Let A be "At least one country is not a democracy AND They have low economic dependence AND The country pair has never been been at war".
- Let B "Both countries are democratic AND they have low economic dependence AND they have never been at war."
- $P(A \cup B) = P(A) + P(B) = 0.43 + 0.18 = 0.61$

General Addition Rule for Probabilities

The general addition rule is used to find the probability that either of two events occurs.

Formula

If A and B are any two events, then the probability that A or B occurs is given by:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Explanation:

- $P(A \cup B)$ is the probability of either A or B occurring.
- $P(A) + P(B)$ adds the probabilities of A and B occurring.
- Subtracting $P(A \cap B)$ corrects for the double counting of the intersection, where both A and B occur.

Venn Diagrams: Example.

Hypothetical **Dataset:** Comparative sample of countries.

Variables: 1) Regime type (Democratic vs. Not Democratic), 2) Level of Economic Inequality (High vs. Low).

	Inequality		**Row Totals**
Regime	**High**	**Low**	**Total**
Democratic	0.18	0.42	0.60
Not Democ.	0.24	0.16	0.40
Column Totals	0.42	0.58	1.00

Table: Joint Probability Distribution

Question: What is the probability that a randomly selected country is either Democratic or has Low Inequality?

Example

Example:

Let A be “a country is a democracy” and B be “a country has low inequality.” Consider a scenario where:

- $P(A) = 0.60$ (Probability of being a democracy),
- $P(B) = 0.58$ (Probability of having low inequality),
- $P(A \cap B) = 0.42$ (Probability of being both Democratic and having low inequality).

Using the general addition rule:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.60 + 0.58 - 0.42 = 0.76$$

Thus, if we select a random country from the sample, the probability that it is either Democratic or has Low Inequality is 76%.

Introduction to Conditional Probability

Overview:

- Conditional probability is a measure of the probability of an event occurring given that another event has already occurred.
- This type of probability is essential in scenarios where the occurrence of one event affects the likelihood of another.

Concept:

- Instead of considering all possible outcomes, conditional probability focuses only on the outcomes where a specific condition or event has occurred.
- Useful in situations like medical testing, where we might be interested in the probability of a disease given a positive test result.

The Formula for Conditional Probability

Conditional Probability Defined:

- The probability of an event A given that event B has occurred is written as $P(A|B)$.

Formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

- This formula assumes that $P(B) > 0$, i.e., the condition event B has a non-zero probability of occurring.
- Note that $P(A|B) \neq P(B|A)$.

General Multiplication Rule

- Note that this gives a general formula for $P(A \cap B)$

$$P(A \cap B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

Example: Using Conditional Probability

Problem Setup:

- We have a bag with 6 red and 4 blue marbles.
- We draw two marbles **without replacement**.
- We want the probability of drawing a blue marble first, then a red marble.

Calculation:

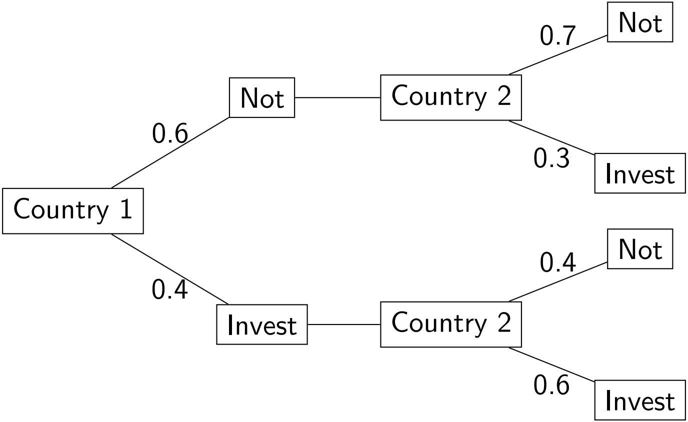
- Probability of drawing a blue marble first: $\frac{4}{10}$.
- Probability of then drawing a red marble (given blue was drawn): $\frac{6}{9}$.
- Combined probability:

$$P(\text{Blue first, then Red}) = \frac{4}{10} \times \frac{6}{9} = \frac{4}{15}.$$

Example: Probability Tree for Nuclear Weapon Investment

We are studying two countries: Country 1 and Country 2. Each country decides whether to invest in nuclear weapons or not.

Probability Tree:



Conditional Probability

What's the probability of the event "Country 2 invests in nuclear weapons (Invest_2), but Country 1 does not (Not_1)"
We can use the following rule: $P(A \cap B) = P(A) \cdot P(B|A)$

Calculation:

- According to the respective branch of the tree: $P(\text{Invest}_2|\text{Not}_1) = 0.3$
- Also, $P(\text{Not}_1) = 0.6$.
- Thus, $P(\text{Invest}_2 \cap \text{Not}_1) = 0.6 \times 0.3 = 0.18$.

Independent Events - Definitions and Properties

Independent Events:

- Events A and B are independent if the occurrence of one does not affect the probability of the occurrence of the other.
- Two events are independent if either:

$$P(B|A) = P(B) \quad \text{(provided that } P(A) > 0\text{) or}$$

$$P(A|B) = P(A) \quad \text{(provided that } P(B) > 0\text{).}$$

- Now, since independence tells us that $P(B|A) = P(B)$, we can substitute $P(B)$ in for $P(B|A)$ in the formula given to us by the multiplication rule:

$$P(A \cap B) = P(A) \times P(B|A) = P(A) \times P(B)$$

Independent Events - *Alternative* Definition

- Given the previous result, events A and B are **independent events** if and only if:

$$P(A \cap B) = P(A) \times P(B)$$

- Otherwise, A and B are called **dependent events**.
- Recall that the "if and only if" in the definition means that the if-then statement works in both directions, in ther words:
    1. If events A and B are independent, then $P(A \cap B) = P(A) \times P(B)$.
    2. If $P(A \cap B) = P(A) \times P(B)$, then events A and B are independent.

Definitions of $P(A \cap B)$ for Independent and Dependent Events

Independent Events

Definition

If events A and B are independent, then:

$$P(A \cap B) = P(A) \times P(B)$$

Dependent Events

Definition

If events A and B are not independent, then:

$$P(A \cap B) = P(A) \times P(B|A)$$

or equivalently:

$$P(A \cap B) = P(B) \times P(A|B)$$

Probability of the Complement

Derivation

Consider an event A and its complement A^c. Since A and A^c are mutually exclusive and exhaustive, the addition rule gives:

$$P(A \cup A^c) = P(A) + P(A^c) - P(A \cap A^c)$$

Because A and A^c are mutually exclusive, $P(A \cap A^c) = 0$. Also, $A \cup A^c$ covers the entire sample space, so $P(A \cup A^c) = P(S) = 1$. Thus:

$$1 = P(A) + P(A^c)$$

Consequently:

$$P(A^c) = 1 - P(A)$$

Example: Law School Applications

Suppose a student estimates a 15% chance of being accepted by any given law school. Assuming acceptance decisions are independent, how many schools should they apply to have a probability of getting accepted to at least one school be higher than 80%?

Let A be the event of being accepted to one particular school, so:

$$P(A) = 0.15, \quad P(A^c) = 0.85.$$

If the student applies to n schools (assuming independent decisions), the probability of being rejected by all schools is:

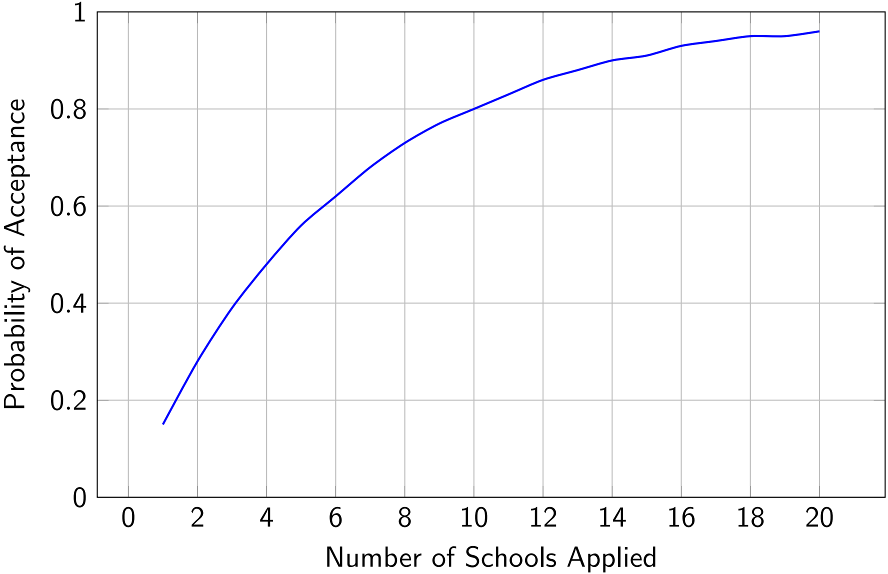
$$P(\text{no acceptances}) = (0.85) \times (0.85) \cdots \times (0.85) = (0.85)^n.$$

Therefore, the probability of at least one acceptance is:

$$P(\text{at least one acceptance}) = 1 - (0.85)^n.$$

We want this probability to be at least 80%. If try for different n we see that 10 is the minimum number of applications such that $P(\text{at least one}) \geq 0.80$

Probability of Acceptance to At Least One School



Law of Total Probability - Intuition

Intuition:

- The Law of Total Probability helps us calculate the probability of an event by considering all possible ways that event can occur.
- It involves breaking down the event into mutually exclusive cases by conditioning on a different event.
- Consider the following example. What's the probability of randomly selecting one Senator and that they are woman?

Law of Total Probability - Example

Define the events B and A. We can split the event B into two cases 1) B and A happen, and 2) B and A^c (not A) happens.

Consider the following example. Assume N_W current senators are woman. Define O as the event "a senator is older than 80 years." And define the event $W \cap O$ as the event "A senator is woman and older than 80 years." Hence, it must be the case that:

- $N_W = N_{W \cap O} + N_{W \cap O^c}$
- Total Woman Senators = # (Woman & Older than 80) + # (Woman & Younger than 80).
- Assume that N is the total number of senators. Then,
- $\frac{N_W}{N} = \frac{N_{W \cap O}}{N} + \frac{N_{W \cap O^c}}{N}$
- $P(W) = P(W \cap O) + P(W \cap O^c)$.

Law of Total Probability - Intuition

Intuition:

- The Law of Total Probability calculates the probability of an event by breaking it down into the mutually exclusive cases the event can happen jointly with another.

Formula:

$$P(B) = P(B \cap A) + P(B \cap A^c)$$

Alternatively, applying the general multiplication rule:

$$P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$$

Explanation:

- $P(B|A)$ is the probability of B given A.
- $P(A)$ is the probability of A.
- $P(A^c)$ is the complement of A.

Law of Total Probability - Example

Example:

- Suppose we have two bags of marbles:
  - Bag One: 70% red marbles, 30% blue marbles.
  - Bag Two: 40% red marbles, 60% blue marbles.
- We randomly choose a bag, with a 50% chance of picking either. What's the probability of drawing a blue marble (B)?

Calculation:

- Let A be the event of choosing Bag One, and A^c be the event of choosing Bag Two.
- $P(B|A) = 0.3$ (Probability of blue marble from Bag One)
- $P(B|A^c) = 0.6$ (Probability of blue marble from Bag Two)
- $P(A) = 0.5$, $P(A^c) = 0.5$

$$P(B) = P(B|A) \cdot P(A) + P(B|A^c) \cdot P(A^c)$$
$$P(B) = 0.3 \cdot 0.5 + 0.6 \cdot 0.5 = 0.15 + 0.3 = 0.45$$

Bayes' Theorem - Intuition and Example

Intuition:

- Bayes' Theorem helps us update our beliefs (probabilities) based on new evidence.
- **Prior:** The initial probability before seeing the new evidence.
- **Posterior:** The updated probability after considering the new evidence.

Bayes' Theorem - Intuition and Example

Example:

- Suppose 2% of politicians are corrupt ($\Pr(C) = 0.02$).
- If a politician is corrupt, there is a 95% chance they are mentioned in the Panama Papers ($\Pr(PP|C) = 0.95$).
- If a politician is not corrupt, there is a 10% chance they are mentioned in the Panama Papers ($\Pr(PP|\neg C) = 0.1$).

Problem: Suppose it is found that a politician was mentioned in the Panama papers. What is the probability that they are corrupt? (i.e., compute $\Pr(C|PP)$).

Bayes' Theorem - General Formula and Law of Total Probability

Bayes' Theorem:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B|A) \cdot P(A) + P(B|\neg A) \cdot P(\neg A)}$$

Terms in Bayesian Nomenclature:

- $P(A|B)$ is the posterior probability of A given B.
- $P(B|A)$ is the likelihood of B given A.
- $P(A)$ is the prior probability of A.
- $P(B)$ is the marginal probability of B, computed using the Law of Total Probability.

Panama Papers and Corruption - Bayes Theorem

Calculation:

$$Pr(P) = Pr(P|C) \cdot Pr(C) + Pr(P|\neg C) \cdot Pr(\neg C)$$

$$Pr(P) = (0.95 \times 0.02) + (0.1 \times 0.98) = 0.019 + 0.098 = 0.117$$

$$Pr(C|P) = \frac{P(PP|C) \times P(C)}{P(PP)} = \frac{0.95 \times 0.02}{0.117} \approx 0.162$$

- The probability that a politician is corrupt, given a mention in the Panama Papers, is approximately 16.24%.

Introduction to Statistical Methods in Political Science

Lecture 5: Random Variables, PMF, PDF, and CDF

Ignacio Urbina

Ph.D. Candidate in Political Science

Random Variables and their Probability Distributions

From Events to Numbers

- Previously, we studied the probability of discrete events, such as:
  - Probability of two countries going to war.
  - Probability that a Senator is a woman.
  - Probability that a randomly sampled person a bachelor's degree.
  - Probability of getting an even number when throwing a die.
- These probabilities help us understand the likelihood of events.
- To perform deeper statistical analysis, we need to quantify these events numerically.

Transforming Educational Attainment into Numbers

- For example, consider educational attainment levels:
  - No education
  - Incomplete high school
  - Complete high school
  - Some college
  - Bachelor's degree
  - Graduate degree
- Instead of discrete categories, we can measure the number of years of schooling.
- This allows us to calculate statistics like the mean number of years of education.

Continuous Numerical Variables: Feeling Thermometers

- Consider the feeling thermometer used in surveys like the American National Elections Study (ANES):
  - Rate feelings toward political figures on a scale from 0 to 100.
  - 0 = Very cold or unfavorable
  - 50 = Neutral
  - 100 = Very warm or favorable
- This allows us to calculate statistics like mean favorability.

Introduction to Random Variables

- We use **random variables** to numerically quantify events.
- A random variable assigns a numerical value to each outcome in a sample space.

Random Variable (*Def.*)

A random variable is a function that maps each outcome in the sample space S to a specific numeric value.

- Allows us to quantify events and analyze them numerically.
- **In practice, we treat Random Variables as numeric variables** and manipulate them using specific algebra rules.
- Example: Tossing a coin, where heads are assigned a 1, and tails a 0.

Types of Random Variables

- **Discrete Random Variables**
  - Take on a finite or countable number of values.
  - Example: Number of children in a family, number of votes received by a candidate.
- **Continuous Random Variables**
  - Take on an infinite number of values within a given range.
  - Example: Income level, time spent on social media.

Random Variables for One Die

- Define the sample space S for throwing one die as:

$$S = \{1, 2, 3, 4, 5, 6\}$$

- Define a random variable X as a function

$$X : S \to \{1, 2, 3, 4, 5, 6\}$$

- X maps each outcome in the sample space to a numerical value.
- Let $X = x$ be a specific value of X. Examples:
  - If the die shows 2, then $x = 2$.
  - If the die shows 5, then $x = 5$.
  - If the die shows 1, then $x = 1$.

Random Variables for Two Dice

- Define the sample space S for two dice as:

$$\begin{aligned} S &= \{(i,j) \mid i,j \in \{1,2,3,4,5,6\}\} \\ &= \{(1,1),(1,2),(1,2),\cdots,(2,1),(2,2),\cdots,(6,5),(6,6)\} \end{aligned}$$

- The sum of the two dice numbers can be represented by a random variable Z.
- Let $Z = z$ be a specific value of Z. Examples:
  - If the dice show (2, 3), then $z = 2 + 3 = 5$.
  - If the dice show (6, 1), then $z = 6 + 1 = 7$.
  - If the dice show (4, 4), then $z = 4 + 4 = 8$.
  - If the dice show (3, 5), then $z = 3 + 5 = 8$.

Random Variables for Two Dice

- Depending on **how we define an event, the function used as a random variable will change**.
- Define the sample space S for two dice as:

$$\begin{aligned} S &= \big\{(i,j) \mid i,j \in \big\{1,2,3,4,5,6\big\}\big\} \\ &= \big\{(1,1),(1,2),(1,3),\cdots,(2,1),(2,2),\cdots,(6,5),(6,6)\big\} \end{aligned}$$

- The **total number of even dice obtained from throwing two dice** can be represented by a random variable V.
- Let $V = v$ be a specific value of V. Examples:
  - If the dice show (2, 3), then $v = 1 + 0 = 1$.
  - If the dice show (1, 6), then $v = 0 + 1 = 1$.
  - If the dice show (4, 4), then $v = 1 + 1 = 2$.
  - If the dice show (3, 5), then $v = 0 + 0 = 0$.

Random Variables for One Coin

- Define the sample space S for throwing a coin as:

$$S = \{H, T\}$$

- Define a random variable W as a function $W : S \rightarrow \{0, 1\}$, where W represents the number of heads.
- W maps each outcome in the sample space to the number of heads.
- Examples:
  - If the coin shows H, then $w = 1$.
  - If the coin shows T, then $w = 0$.

Random Variables for Three Coins

- Define the sample space S for three coins as:

$$S =\big\{(H,H,H),(H,H,T),(H,T,H),(H,T,T),(T,H,H),\\ (T,H,T),(T,T,H),(T,T,T)\big\}$$

- Define a random variable Y as a function $Y : S \to \{0,1,2,3\}$, where $Y = y$ represents the number of heads.
- Y maps each outcome in the sample space to the number of heads.
- Examples:
  - If the coins show (H, H, T), then $y = 2$.
  - If the coins show (T, T, T), then $y = 0$.
  - If the coins show (H, H, H), then $y = 3$.
  - If the coins show (T, H, T), then $y = 1$.

Random Variables for Election Results

- Consider the results of the last Senatorial election in three battleground states. Define the sample space S for these results as:
  $$S = \{(R,R,R),(R,R,D),(R,D,R),(R,D,D),(D,R,R),\\ (D,R,D),(D,D,R),(D,D,D)\}$$
- Define a random variable Z as a function $Z : S \to \{0,1,2,3\}$, where $Z = z$ represents the number of battleground states won by the Republican Party.
- Z maps each outcome in the sample space to the number of states won by the Republican Party.
- Examples:
  - If the results are (R, R, D), then $z = 2$.
  - If the results are (D, D, D), then $z = 0$.
  - If the results are (R, R, R), then $z = 3$.
  - If the results are (D, R, D), then $z = 1$.

Definition of Probability for Discrete Random Variables

Probability for Discrete Random Variables ($Def.$)

When each outcome in the sample space S is equally likely, the probability that a discrete random variable X takes the value x is given by the ratio of the number of outcomes in S where $X = x$ to the total number of outcomes in S.

- For a discrete random variable X and a specific value x:

$$P(X = x) = \frac{\text{Number of favorable outcomes in which } X = x}{\text{Total number of possible outcomes}}$$

- This is based on the classical definition of probability

Properties of the Probability Distribution for a Discrete Random Variable

A function can serve as the probability distribution for a discrete random variable X if and only if its values, $P(X = x)$, satisfy the conditions:

- $P(X = x) \geq 0$ for each value within its domain
- $\sum_{i=1}^{k} P(X = x_i) = 1$, where the summation extends over all the values within its domain ($\mathcal{D}_X = \{x_1, x_2, \cdots, x_k\}$)

Note that a *discrete* random variable's probability distribution is often referred to as its **Probability Mass Function (PMF)**.

Probability Distribution for Number of Heads in Three Coin Tosses

- Sample space S for three coin tosses:

$$S = \{(H,H,H),(H,H,T),(H,T,H),(H,T,T),(T,H,H),\\ (T,H,T),(T,T,H),(T,T,T)\}$$

- Define random variable X as the number of heads.
- Probability Distribution (PMF):

Number of Heads: x	**Probability:** $P(X = x)$
0	1/8
1	3/8
2	3/8
3	1/8

Example: Simplified Feeling Thermometer

- Suppose a survey firm implemented a 10-point feeling thermometer to measure people's feelings towards political leaders or public figures.
- Ratings range from 0 (very unfavorable) to 10 (very favorable).
- Let's say this pollster was interested in the feelings towards the current President among the public.
- Define the random variable X as the response of one subject on this thermometer.

PMF for the 10-Point Feeling Thermometer Towards the Current President (x)

PMF for X

x	$P(X = x)$
0	0.05
1	0.10
2	0.10
3	0.15
4	0.10
5	0.15
6	0.10
7	0.10
8	0.05
9	0.05
10	0.05

Probability Questions

What is the probability that X is...

1. Greater than or equal to 8?

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10)$$
$$= 0.05 + 0.05 + 0.05 = 0.15.$$

2. Less than 5?

$$P(X < 5) =$$
$$= p_X(0) + p_X(1) + p_X(2) + p_X(3) + p_X(4)$$
$$= 0.05 + 0.10 + 0.10 + 0.15 + 0.10 = 0.50.$$

3. Between 3 and 7 inclusive?

$$P(3 \leq X \leq 7) = P(X = 3) + P(4) + \cdots + P(7)$$
$$= 0.15 + 0.10 + 0.15 + 0.10 + 0.10 = 0.60.$$

Common PMF For Discrete Random Variables

Common PMFs for Discrete Random Variables

- There are several common probability mass functions (PMFs) used to model discrete random variables.
- We will discuss three of them:
  - Uniform Distribution
  - Bernoulli Distribution
  - Binomial Distribution
- But note that there are several more: Negative Binomial, Geometric, Poisson, Multinomial, etc (see: Link).
- The following is a strongly recommended reference: `https://uw-statistics.github.io/Stat311Tutorial/discrete-distributions.html`.

Uniform Distribution

- In a uniform distribution, each outcome is equally likely.
- For a discrete random variable X with n possible outcomes:

$$p_X(x) = \frac{1}{n} \quad \text{for all } x$$

- Note that: $p_X(x) \geq 0$ and $\sum_x p_X(x) = \frac{1}{n} + \frac{1}{n} + \cdots + \frac{1}{n} = n \times \frac{1}{n} = 1$
- This means that $p_X(x)$ is a well-defined PMF.

Example: Uniform Distribution

- Consider rolling a fair six-sided die:

$$S = \{1, 2, 3, 4, 5, 6\}$$

- The PMF for each outcome is:

$$p_X(x) = \frac{1}{6} \quad \text{for } x = 1, 2, 3, 4, 5, 6$$

- This means each number has an equal probability of $\frac{1}{6}$.

Calculating Probabilities for Events Involving a Fair Six-Sided Die

- **Single Value Event:**
  - Probability of rolling a 4:

$$P(X=4)=\frac{1}{6}$$

- **Composite Event:**
  - Probability of rolling an even number (2, 4, or 6):

$$P(x \text{ is even})=P(X=2)+P(X=4)+P(X=6)=\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{3}{6}$$

- **Multiple Disjoint Events:**
  - Probability of rolling a number less than 4 (1, 2, or 3):

$$P(x<4)=P(X=1)+P(X=2)+P(X=3)=\frac{1}{6}+\frac{1}{6}+\frac{1}{6}=\frac{3}{6}=\frac{1}{2}$$

- **Non-occurring Event:**
  - Probability of rolling a 7:

$$P(x=7)=0$$

Bernoulli Distribution

- The Bernoulli distribution models a single trial with two outcomes: success (1) and failure (0).
- For a random variable X with success probability p:

$$p_X(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

- Which of the two outcomes is labeled as a success or failure is generally arbitrarily defined.

Example: Bernoulli Distribution

- Consider flipping a fair coin:

$$S = \{H, T\}$$

- Let X be 1 if heads (H) and 0 if tails (T). If the coin is fair, the PMF is:

$$p_X(x) = \begin{cases} 0.5 & \text{if } x = 1 \\ 0.5 & \text{if } x = 0 \end{cases}$$

Example: Bernoulli Distribution for Sociodemographic Variable

- **Example**: If we randomly select one person among the American public, what is the probability they identify as Latino/a/x?
- Let X be a random variable where $X = 1$ if the person identifies as Latino/a/x and $X = 0$ otherwise.
- According to 2020 US Census data, the probability p that a randomly selected person identifies as Latino/a/x is 0.189.
- We can represent X as a Bernoulli random variable :

$$p_X(x) = \begin{cases} 0.189 & \text{if } x = 1 \\ 0.811 & \text{if } x = 0 \end{cases}$$

Binomial Distribution

- The binomial distribution models the number of successes in a fixed number of independent Bernoulli trials.
- For a random variable X representing the number of successes in n trials with success probability p:

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

- Example: Number of heads in 10 flips of a fair coin.

Example: Binomial Distribution

- Consider flipping a fair coin 10 times:

$$n = 10, \quad p = 0.5$$

- The PMF for the number of heads X is:

$$p_X(x) = \binom{10}{x}(0.5)^x(0.5)^{10-x}$$

- This gives the probability of getting exactly x heads in 10 flips.

Example: Given a random sample of 10 people, what is the probability that two or fewer identify as Latino/a/x?

- Define X as the number of Latinos in a sample of 10. This follows a binomial distribution: $X \sim \text{Binomial}(N = 10, p = 0.189)$.

$$P(X = 0) = \binom{10}{0} \times 0.189^0 \times 0.811^{10} \approx 0.1231$$

$$P(X = 1) = \binom{10}{1} \times 0.189^1 \times 0.811^9 \approx 0.2868$$

$$P(X = 2) = \binom{10}{2} \times 0.189^2 \times 0.811^8 \approx 0.3008$$

- Adding these probabilities:

$$P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2) \approx 0.7107$$

- This formula calculates the total probability of getting 0, 1, or 2 Latinos in the sample, which is approximately 71.07%.

Example: Probability of at least 1 or at least 3 Latinos in a sample of 10

- Define X as the number of Latinos in a sample of 10. This follows a binomial distribution: $X \sim \text{Binomial}(N = 10, p = 0.189)$.
- **Probability of at least 1 Latino:**

$$P(X \geq 1) = 1 - P(X = 0)$$

$$P(X \geq 1) = 1 - 0.1231 \approx 0.8769$$

- **Probability of at least 3 Latinos:**

$$P(X \geq 3) = 1 - P(X \leq 2)$$

$$P(X \geq 3) = 1 - 0.7107 \approx 0.2893$$

- These probabilities show that there is an 87.69% chance of having at least one Latino in the sample and a 28.93% chance of having at least three.

Galton Board and the Binomial Distribution

- Binomial PMF graph: `https://shiny.rit.albany.edu/stat/binomial/`
- Galton Board: `https://www.mathsisfun.com/data/quincunx.html`

Probability for Continuous Random Variables

Concept of Probability for Continuous Random Variables

- Continuous random variables take on infinite possible values within a given range.
- Therefore, the probability of *exactly* obtaining a specific value $X = x$ after a random trial is 0 due to this definition.
- Hence, when dealing with **continuous random variables**, we need to work with **probabilities of intervals**, i.e., $\Pr(X \in [a, b])$.
- So, instead of asking what's the probability that someone's height is precisely 5 feet 7.16 inches, we can ask what is the probability that someone's height is between 5 feet 5 inches and 5 feet 10 inches.

Introduction to Probability Density Function (PDF)

- The **Probability Density Function (PDF)**, denoted as $f_X(x)$, describes the likelihood of a continuous random variable X taking on a particular value.
- Note: the PDF represents the *relative likelihood* of X being near x. $f_X(x)$ **IS NOT** the probability of x.
- To find the probability that X lies within an interval $[a, b]$, we calculate the area under the PDF curve between a and b.
- Mathematically, this is formally expressed using definite integrals:

$$P(X \in [a, b]) = P(a \leq X \leq b) = \int_a^b f_X(x) \cdot dx$$

Notation: Definite Integrals ([illegible])

- A definite integral represents the area under the graph of a function and the x-axis.
- The notation for a definite integral is:

$$\int_a^b f(x) \cdot dx$$

- Here:
  - a is the lower limit of integration.
  - b is the upper limit of integration.
  - $f(x)$ is the function being integrated.
  - dx indicates that the integration is with respect to x.
- The definite integral calculates the net area between the function $f(x)$ and the x-axis from $x = a$ to $x = b$.

Definition of a Well-Defined Probability Density Function (PDF)

Probability Density Function (PDF)

A function $f_X(x)$ is a valid probability density function (PDF) of a continuous random variable X if and only if it satisfies the non-negativity and normalization conditions.

1. **Non-Negativity:** For all possible values of x,

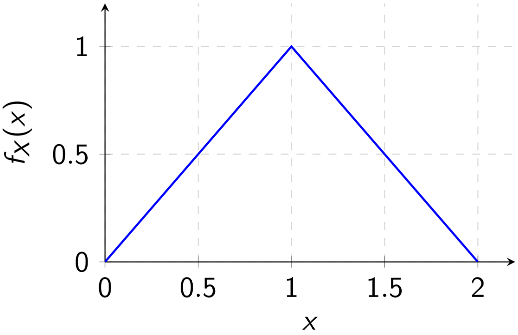
$$f_X(x) \geq 0, \quad \text{for all } x.$$

2. **Normalization:** The total probability over all possible values of X must sum to 1, meaning:

$$\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$$

Connecting the Dots: Example

- Define a random variable X that takes values from 0 to 2.
- The PDF of X follows a *triangular distribution* with mode at $c = 1$:



$$f_X(x) = \begin{cases} x, & 0 \leq x \leq 1 \\ (2 - x), & 1 \leq x \leq 2 \end{cases}$$

Area Under the Curve of the PDF

- The area under the PDF consists of two right triangles.
- The left triangle (from 0 to 1) has:
  - Base $= 1 - 0 = 1$
  - Height $= 1$
- The right triangle (from 1 to 2) has:
  - Base $= 2 - 1 = 1$
  - Height $= 1$
- Using the formula for the area of a triangle:

$$\left(\frac{1}{2} \times 1 \times 1\right) + \left(\frac{1}{2} \times 1 \times 1\right) = 0.5 + 0.5 = 1$$

- This confirms the total area under the curve is 1, making it a valid PDF.

Calculating Probability for an Interval

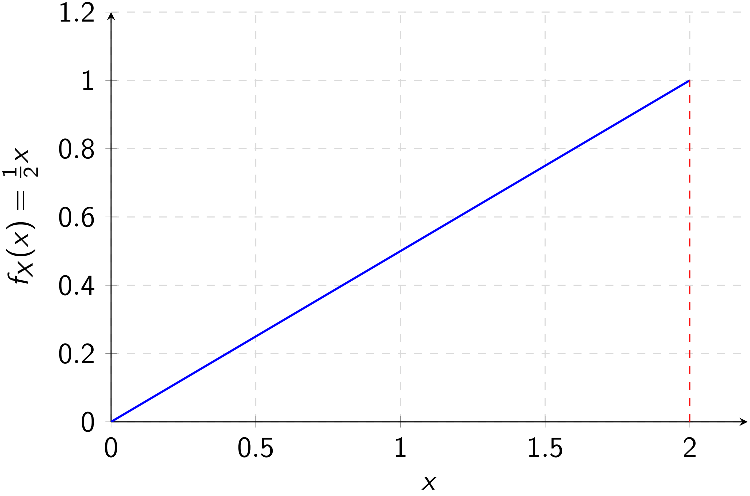
Problem: Calculate the probability that X lies between 0.5 and 1.

- The probability is the area under the PDF between 0.5 and 1, which can be broken down as the area of a small triangle and a rectangle under $f_X(x)$ from 0.5 to 1.0.
- Rectangle area:
  - Base $= 1 - 0.5 = 0.5$
  - Height $= 0.5 - 0 = 0.5$
  - Area $= \mathit{Base} \times \mathit{Height} = 0.5 \times 0.5 = 0.25$
- Small triangle area:
  - Base $= 1 - 0.5 = 0.5$
  - Height $= 1 - 0.5 = 0.5$
  - Area $= \frac{1}{2} \times \mathit{Base} \times \mathit{Height} = \frac{1}{2} \times 0.5 \times 0.5 = 0.125$
- Thus,

$$P(0.5 \leq X \leq 1) = 0.25 + 0.125 = 0.375$$

Example 2

- Define a random variable X that takes on values from 0 to 2.
- Suppose the PDF of X is $f_X(x) = \frac{1}{2}x$ for $0 \leq x \leq 2$.
- Graph of $f_X(x) = \frac{1}{2}x$ from 0 to 2:



Area Under the Curve of the PDF

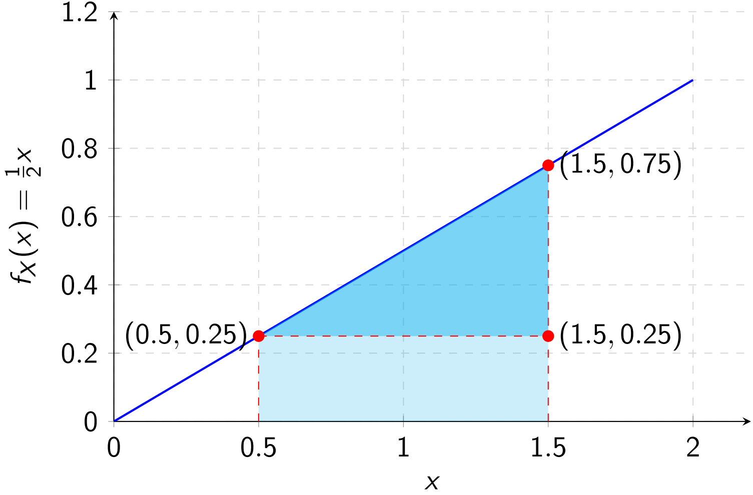
- The area under the curve $f_X(x) = \frac{1}{2}x$ from $x = 0$ to $x = 2$ forms a right triangle.
- The base of the triangle is $2 - 0 = 2$.
- The height of the triangle is $\frac{1}{2} \times 2 = 1$.
- Using the formula for the area of a triangle:

$$\text{Area} = \frac{1}{2} \times \text{base} \times \text{height} = \frac{1}{2} \times 2 \times 1 = 1$$

- This means the total area under the PDF is 1, making it a well-defined PDF.

Calculating Probability for an Interval

- To calculate the probability that X lies between 0.5 and 1.5:
- We use one rectangle and one triangle to represent the area.



Calculating Probability for an Interval

- The area of the rectangle (base = 1.5 - 0.5 = 1, height = 0.25):

$$\text{Area} = \text{base} \times \text{height} = 1 \times 0.25 = 0.25$$

- The area of the triangle (base = 1.5 - 0.5 = 1, height = 0.75 - 0.25 = 0.5):

$$\text{Area} = \frac{1}{2} \times \text{base} \times \text{height} = \frac{1}{2} \times 1 \times 0.5 = 0.25$$

- Total area from 0.5 to 1.5:

$$0.25 + 0.25 = 0.5$$

- This means:

$$P(0.5 \leq X \leq 1.5) = \int_{0.5}^{1.5} \frac{1}{2}x \cdot dx = 0.5$$

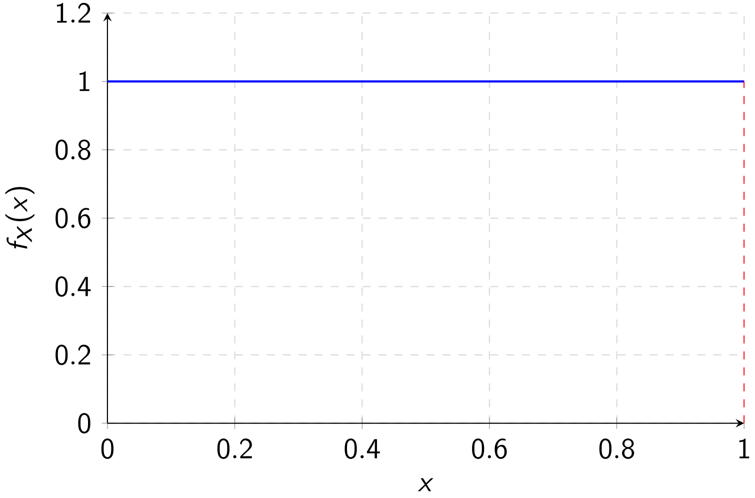
Uniform Distribution for Continuous Random Variables

- Consider the uniform distribution on the interval [0, 1].
- The PDF is:

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- The total area under the PDF curve is 1.

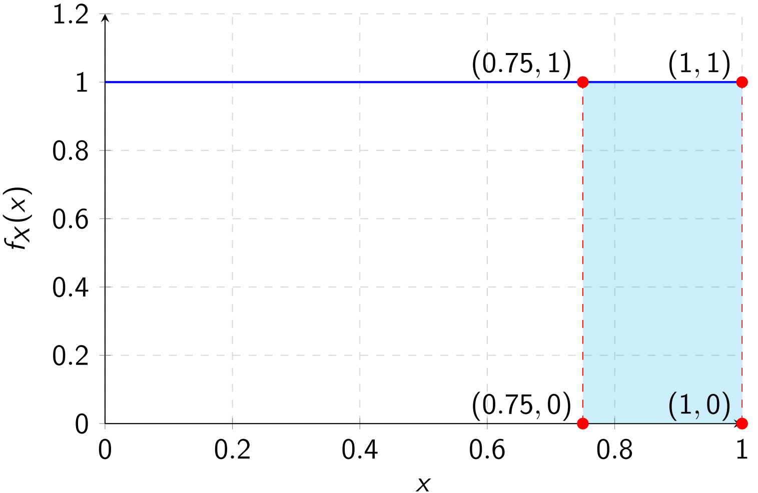
Graph of the Uniform Distribution PDF



Probability Calculation for Uniform Distribution

- The probability that X is greater than 0.75 in a uniform distribution from 0 to 1 is calculated as follows:

$$P(0.75 < X) = 1 \times (1 - 0.75) = 0.25$$



Introduction to the Normal Distribution

Normal Distribution

- The normal distribution is a continuous probability distribution.
- Defined by its mean μ and standard deviation σ.
- The PDF of the normal distribution is:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

- Main features of the normal distribution:
  - It is a symmetrical distribution.
  - Values close to the mean are more likely than values very far from it.

Example: Standard Normal Distribution

- Consider a normal distribution with $\mu = 0$ and $\sigma = 1$ (standard normal distribution).
- The PDF is:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

- The total area under the PDF curve is 1.

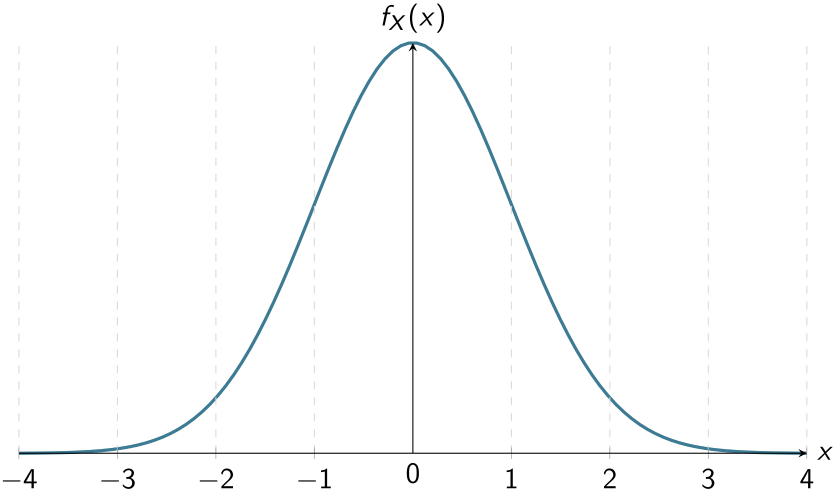


Figure: Standard Normal Distribution

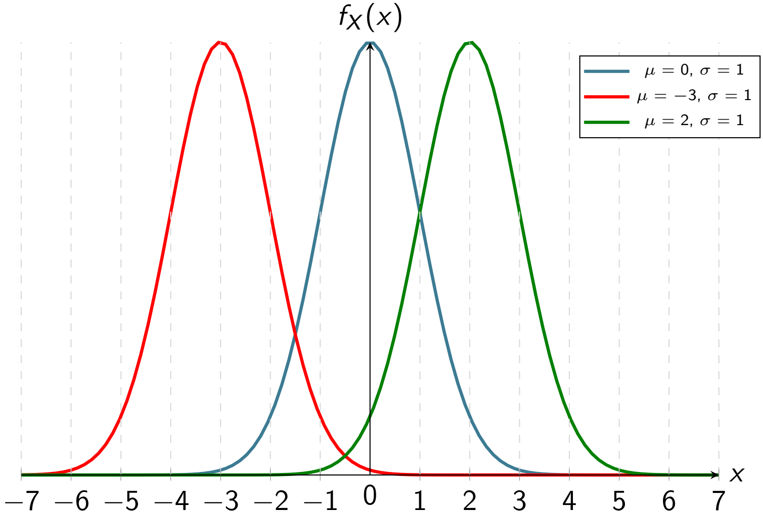


Figure: Comparison of Normal Distributions with Different Means

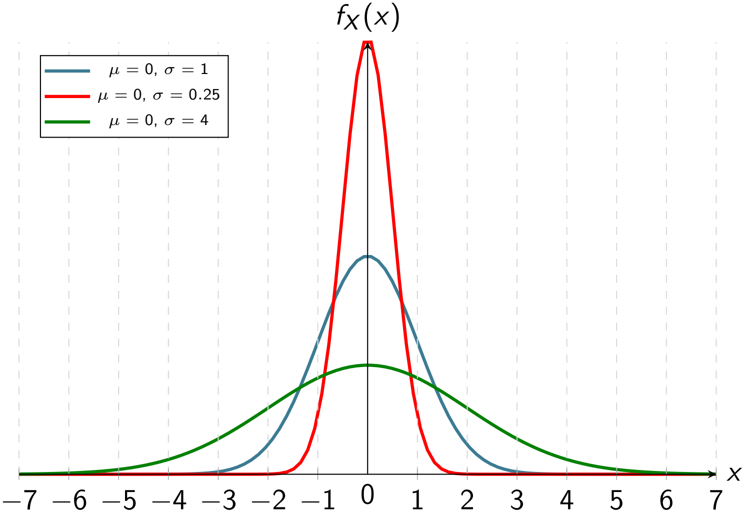


Figure: Comparison of Normal Distributions with Same Mean and Different Standard Deviations

Cumulative Distribution Function

Cumulative Distribution Function (CDF): Concept

- The Cumulative Distribution Function (CDF) is a function that describes the probability that a random variable takes a value less than or equal to a specific value.
- It provides a complete description of the probability distribution of a random variable.
- The CDF is fundamental in probability theory and statistics as it gives an integral overview of the distribution and helps in understanding the likelihood of different outcomes.

Cumulative Distribution Function (CDF): Formal Definition

Discrete Random Variable (RV)

- Definition:

$$F_X(x) = P(X \leq x) = \sum_{k:k \leq x} p_X(k)$$

- Here, $p_X(k)$ represents the probability mass function (PMF) of X at k.
- For here and after we will denote $F_X(x)$ by $CDF_X(x)$.

Continuous Random Variable (RV)

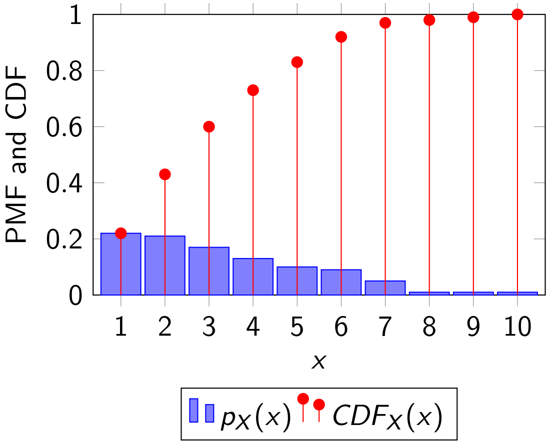
- Definition:

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_X(t)\,dt$$

- $f_X(t)$ is the probability density function (PDF) of X.
- For here and after we will denote $F_X(x)$ by $CDF_X(x)$.

Example of CDF for a Discrete RV

- X = consecutive elections won by current members of the House of Representatives, with values from 1 to 10.
- We define a left-skewed PMF for X where each $p_X(x)$ decreases as x increases, reflecting a higher probability for fewer election wins.



Probability Rules Involving CDFs

- We know that every PMF satisfies:

$$\sum_{x \in D_X} p_X(x) = 1$$

- Let x_1 be the minimum, and x_n the maximum values in the domain of X, D_X. For an $a \in D_X$ such that $x_1 < a < x_n$, the sum of probabilities can be split at a:

$$\sum_{x=x_1}^{x_n} p_X(x) = \sum_{x=x_1}^{a} p_X(x) + \sum_{x=a+1}^{x_n} p_X(x) = 1$$

- This leads to the rule:

$$P(X \leq a) + P(a < X) = CDF_X(a) + P(a < X) = 1$$

- Therefore:

$$\boldsymbol{P(a < X) = 1 - CDF_X(a)}$$

Probability Rules with CDFs and Intervals

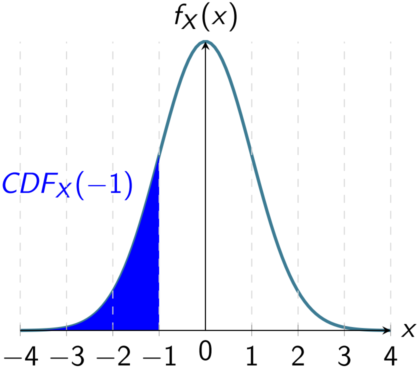
The previous derivation (proof) was shown *just as an illustration*. Proceeding similarly, one can show the following rules apply.

- $1 = P(X \leq a) + P(X > a)$
- $\boldsymbol{P(X > a) = 1 - CDF_X(a)}$
- $1 = P(X \geq a) + P(a > X > b) + P(X \leq b)$
- $P(a > X > b) = 1 - [P(X \geq a) + P(X \leq b)]$
- For $b < a$, $P(X \leq a) = P(X \leq b) + P(b < X \leq a)$
- $\boldsymbol{P(b < X \leq a) = CDF_X(a) - CDF_X(b)}$

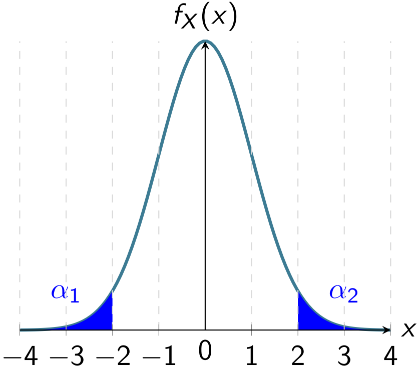
Note: These rules apply to both discrete and continuous random variables. For continuous random variables, the symbols $>$ and $<$, as well as $\geq$ and $\leq$, are used interchangeably because there is no practical distinction between them.

Probabilities with CDF: Example

- Let X be a standard normal random variable. Thus, $X \sim N(0, 1)$.
- Then, $\Pr(X \leq -1) = CDF_X(-1) = \Phi(-1) = 0.1586$.
- The "blue area" corresponds to a probability of 15.9%.
- In other words, the probability that X is lower or equal to -1 is 15.9%.



CDFs for Symmetric Distributions with $\mu_X = 0$



Symmetry and CDF

- Assume $a > 0$. For a symmetric PDF (or PMF) with $\mu_X = 0$, the following holds:

$$P(X \leq -a) = P(X \geq a)$$

- Hence:

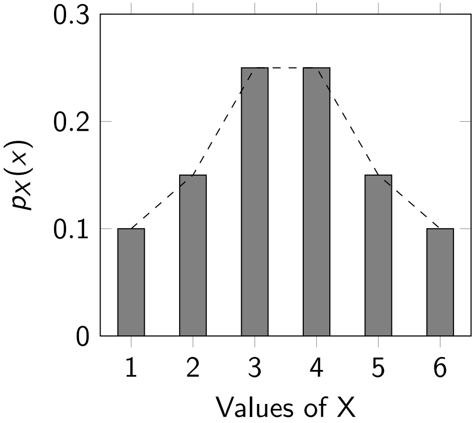
$$CDF_X(-a) = 1 - CDF_X(a)$$

- e.g., if x is a Std. Normal, then:

$$0.023 = \alpha_1 = CDF_X(-2) = 1 - CDF_X(2) = \alpha_2 = 0.023$$

Symmetrical Distribution of a Discrete Random Variable

X	$p_X(x)$	$CDF_X(x)$
1	0.1	0.1
2	0.15	0.25
3	0.25	0.5
4	0.25	0.75
5	0.15	0.9
6	0.1	1.0



Introduction to Statistical Methods in Political Science

Lecture 6: Expected Value and Population Parameters

Ignacio Urbina

Ph.D. Candidate in Political Science

Population Parameters and the Expected Value

Introduction to Population Parameters

- Population parameters (e.g., μ_X and σ_X^2) are numeric, deterministic values that summarize the characteristics of a population's distribution.
- These parameters include measurements of central tendency and spread (i.e., dispersion), providing insights into the general behavior and variability of the population.
- Population parameters are not random variables themselves; they are fixed quantities calculated from the distribution of a random variable.

Expectation: The Expected Value of a Random Variable

- What is the **expected value** or **expectation** of a random variable?
- Simply put, the expected value is the **center** of the distribution.
- But what do we mean by **center**?
  - It is the point that balances the distribution's domain in a precise mathematical sense.
- More formally, the expected value of X is the point where its distribution's **center of mass** lies, balancing the weighted distances to all other points.
- Thus, the **expectation** of X, written as $E[X]$, equals its **population mean**:

$$E[X] = \mu_X.$$

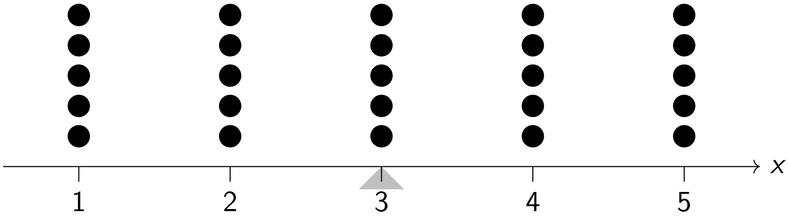
Formal Definition of $E(X)$ - Discrete RV

- Let X be some **discrete random variable (RV)**.
- The *Expected Value* of X is defined as:

$$E(X) = \sum_{x \in D_X} x \cdot p_X(x)$$

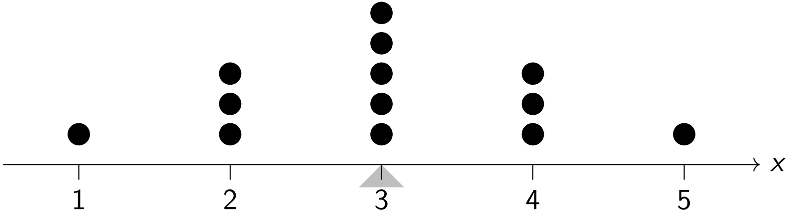
- Here, $p_X(x)$ is the probability mass function (PMF) of X.
- D_X is the domain of X, that is, all the values X can take.
- Note that $E(X)$, at any given point in time, is an unknown constant, not a random variable.

Graphical Example: Symmetric Distribution



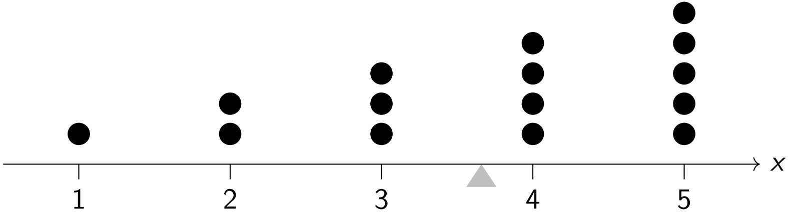
$$\mu_X = 3$$

Graphical Example: Symmetric Distribution



$$\mu_X = 3$$

Graphical Example: Skewed Distribution



$\longrightarrow \mu_X = 3.66$

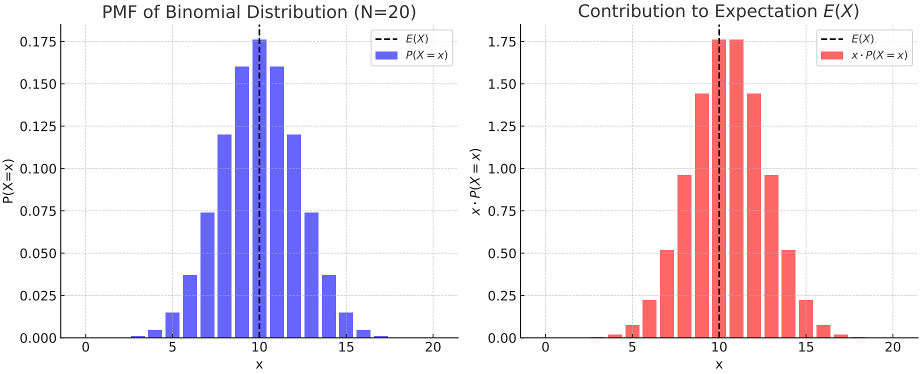
Formal Definition of $E(X)$ - Continous RV

- Let X be a **continuous random variable**.
- The expected value of X is defined as:

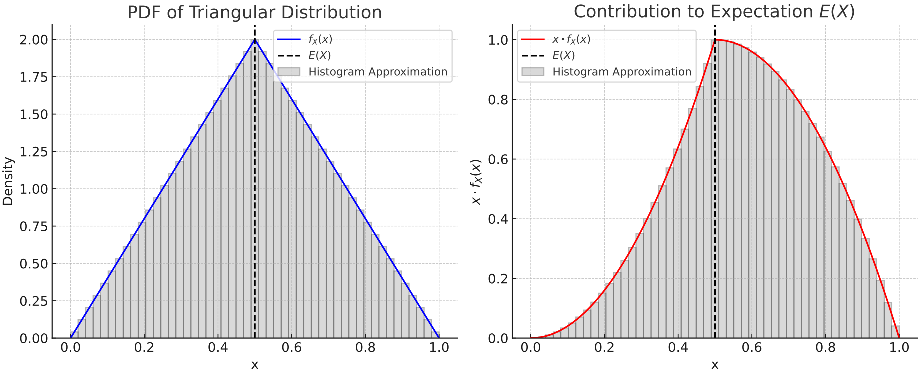
$$E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx$$

- $f_X(x)$ is the probability density function (PDF) of X.
- The integral captures the total weighted area, just like a sum does in the discrete case.
- Just as a sum adds up discrete contributions, the integral accumulates contributions over a continuous range.

Conceptual Illustration of $E(X)$ - Discrete RV



Conceptual Illustration of $E(X)$ - Continuous RV



Example: Discrete RV Representing Ideology

PMF Table:

x (Ideology)	$P(X = x)$
-5 (Far left)	0.05
-4	0.10
-3	0.15
-2	0.20
-1	0.20
0	0.10
1	0.10
2	0.05
3	0.03
4	0.02
5 (Far right)	0.00

Expectation:

- Definition:

$$E(X) = \sum_{\text{For all } x} x \cdot p_X(x)$$

- Calculation:

$$E(X) = (-5)(0.05) + (-4)(0.10)$$

$$+(-3)(0.15)+(-2)(0.20)+(-1)(0.20)$$

$$+(0)(0.10)+(1)(0.10)+(2)(0.05)$$

$$+(3)(0.03)+(4)(0.02)+(5)(0.00).$$

- Result: $E[X] = -1.33$

Some Remarks on Expectation

- Expectation allows us to define fundamental features of the true (unobserved) distribution of a random variable.
- These features are always **centered** by the weights imposed by the probability distribution of the random variable.
- As statistical analysts, our goal is to use samples to estimate these fundamental features, also called **population parameters**.
- This process forms the foundation of statistical inference, enabling us to draw conclusions about the true distribution from observed data.

Population Variance Expressed Using Expectations

- Analogously to the population mean, we can express the population variance using expectations.
- Population variance measures the **expected quadratic deviation from the population mean.**
- Considering the population probability distribution, we ask: what is the expected quadratic deviation from the mean?
- This approach helps us understand how much the values of a *random variable deviate*, on average, from the mean.

Population Variance: Formal Definition Using Expectation

- Formal Definition:

$$\begin{aligned} V(X) = \sigma_X^2 &= E[(X - E(X))^2] \\ &= E[(X - \mu_X)^2]. \end{aligned}$$

- $V(X)$ and σ_X^2 are alternative notations for "*the variance of* X."
- The population variance is a "centered" measure of the squared deviation of X from its population mean.
  - So, σ_X^2 is also an expected value! (only that we take expectation with regard to $(X - E(X))^2$ and not just X)

Population Variance: Discrete Case

- Variance measures the expected squared deviation from the mean.
- For a discrete random variable:

$$V(X) = E[(X - \mu_X)^2] = \sum_{x \in D_X} (x - \mu_X)^2 \cdot p_X(x)$$

- Where $\mu_X = E(X)$ is the expected value of X.
  - **Note 1**: Population SD is given by $\sigma_X = \sqrt{V(X)} = \sqrt{E[(X - \mu)^2]}$
  - **Note 2**: A general property of expectations is, given some function discrete RV, and any continuous function $g(x)$, then $E[g(x)] = \sum_{x \in D_X} g(x) \cdot p_X(x)$.

Population Variance: Continuous Case

- For a continuous random variable, variance is defined as:

$$V(X) = E[(X - \mu_X)^2] = \int_{-\infty}^{\infty} (x - \mu_X)^2 \cdot f_X(x)\, dx$$

- Where $\mu_X = E(X)$ is the expected value of X, and $f_X(x)$ is the PDF of X.

Population Mean, Variance, and SD for Common RVs Distributions

Distrib.	$\boldsymbol{X}$	$\boldsymbol{E(X)}$	$\boldsymbol{V(X)}$	$\boldsymbol{SD(X) = \sqrt{V(X)}}$
Uniform	$X \sim U[a,b]$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2-1}{12}$	$\sqrt{\frac{(b-a+1)^2-1}{12}}$
Bernoulli	$X \sim Ber(p)$	p	$p(1-p)$	$\sqrt{p(1-p)}$
Binomial	$X \sim Bin(n,p)$	np	$np(1-p)$	$\sqrt{np(1-p)}$
Normal	$X \sim N(\mu, \sigma^2)$	μ	σ^2	σ

Example: Mean and Variance for Uniform Distribution (Die Roll)

PMF for a Die Roll

- $p_X(x) = \frac{1}{6}$ for $x = 1, 2, 3, 4, 5, 6$

Population Mean:

- $E(X) = \mu = \sum_{i=1}^{6}(x_i) \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5.$

Population Variance

- $V(X) = \sigma^2 = E[(X - \mu)^2] =$ $= \sum_{i=1}^{6}(x_i - \mu)^2 \cdot \frac{1}{6}$ $= \frac{1}{6}[(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2]$ $= \frac{1}{6}[6.25 + 2.25 + 0.25 + 0.25 + 2.25 + 6.25].$
- $\sigma^2 = \frac{1}{6} \times 17.5 = 2.9167$

Population SD

- $SD(X) = \sigma = \sqrt{2.9167} \approx 1.71$

Properties of Expectations and Variances

Let a, b, and c be numeric constants (i.e., not RVs). Let X be any random variable. Then, the following properties hold:

- $E(aX) = aE(X)$
- $E(c) = c$
- $V(X + c) = V(X)$ (Proof)
- $V(bX) = b^2 V(X)$ (Proof)
- $V(X) = E(X^2) - [E(X)]^2$ (Proof)

Linear Combinations (LCs): Definition and Usefulness

Definition:

- A linear combination of random variables X and Y is an expression of the form $aX + bY + c$, where a, b, and c are constants.

Usefulness:

- Helps in deriving properties of combined distributions.
- Essential for understanding linear regression and other statistical models.
- Facilitates transformations to standardize variables.

Motivation: Total Earnings from Two Jobs

- Alex works two part-time jobs:
  - **Job A:** earns a random amount X per week.
  - **Job B:** earns a random amount Y per week.
- Total weekly earnings:

$$Z = X + Y$$

- Key properties:
  - **True Mean Earnings:** $E[Z] = E[X] + E[Y]$
  - **Variance (if independent):** $\text{Var}(Z) = \text{Var}(X) + \text{Var}(Y)$
- If X and Y are correlated, covariance matters!

Properties of Expectations and Variances of LCs

Let a, b, and c be numeric constants (i.e., not RVs). Let X and Y be any two *independent* random variables. Then,

- $E(aX + bY + c) = aE(X) + bE(Y) + c$
- $V(aX + bY + c) = a^2V(X) + b^2V(Y)$ (when X and Y are independent)
- Let $x_1, x_2, \cdots, x_n$ be a series of independent RVs, which could follow different probability distributions. Then,
  - $E[\sum_{i=1}^{n} x_i] = \sum_{i=1}^{n} E[x_i]$
  - $V[\sum_{i=1}^{n} x_i] = \sum_{i=1}^{n} V[x_i]$

Note: We will discuss the case for non-independent X and Y when we cover linear regression.

Practical Example: Z-score Transformation

Assume Y is normally distributed: $Y \sim N(\mu_Y, \sigma_Y^2)$. Let $Z = \frac{Y-\mu}{\sigma}$ be a linear transformation of Y (this is the *Z-score transformation* to Y).

- $E[Z] = E[\frac{Y-\mu}{\sigma}] = E[\frac{Y}{\sigma}] - E[\frac{\mu}{\sigma}] = \frac{E[Y]}{\sigma} - \frac{\mu}{\sigma} = 0$
- $V[Z] = V[\frac{Y-\mu}{\sigma}] = \frac{1}{\sigma^2} V[Y-\mu] = \frac{1}{\sigma^2} \cdot V[Y] = \frac{1}{\sigma^2} \cdot \sigma^2 = 1$
- Therefore, $E[Z] = 0$ and $V[Z] = 1$

Note that this applies to any random variable, regardless of its distribution.

Why use the Z-score Transformation?

Any r.v. X can be expressed as a Z-score, $Z = \frac{X-\mu}{\sigma}$.

- Why would we want to do this? Because units in Z are expressed as deviations from the mean.
- And its unit of measurement (since we divide by the SD) is in SD units.
- Hence, $Z = 2$ means that the original variable's value (e.g., some X) is 2 standard deviations over the mean.

Z-scoring normal random variables

- Teach how to use the Z score and the standard normal to find probabilities of normally distributed RVs

Practical Example: Sample Average as a Linear Combination

- Assume there is a series of independent draws of X, denoted by x_i, all following a normal distribution $X \sim N(\mu, \sigma^2)$.
- Construct a linear combination representing the sample average:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Calculate $E[\bar{X}]$:

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n}\sum_{i=1}^{n} E[x_i] = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}\cdot n\mu = \mu$$

Practical Example: Sample Average as a Linear Combination

- Calculate $V[\bar{X}]$:

$$V\left[\bar{X}\right] = V\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2} V\left[\sum_{i=1}^{n} x_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} V[x_i] = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

- Therefore:
  - $E[\bar{X}] = \mu$
  - $V[\bar{X}] = \sigma^2/n$
  - $SD(\bar{X}) = \sqrt{V[\bar{X}]} = \sigma/\sqrt{n}$

Lecture 7: Estimators and Sampling Distributions

Ignacio Urbina

Introduction & Motivation

- **Objective**: Explore estimators and sampling distributions to assess the probability of obtaining specific estimates, given known distribution assumptions.
- **Why is this important?**
  - To quantify the likelihood of observing a particular estimate or range of estimates.
  - Analyze the behavior of sample estimates relative to population parameters and underlying assumptions.
- This forms the foundation of **statistical inference**, enabling us to draw conclusions about populations from samples.

Population Parameters vs. Sample Estimators

- **Population Parameters**:
  - Fixed values that describe the characteristics of the entire population (e.g., mean μ, variance σ^2).
  - These values are typically unknown and constant (at a given point in time).
- **Sample Estimators**:
  - **Estimator**: A function applied to sample data to estimate population parameters.
  - **Estimate**: A specific value calculated from a sample that serves as an approximation of the population parameter.
  - **Examples**: Sample mean ($\bar{X}$) estimates μ, sample proportion ($\hat{p}$) estimates population proportion, and sample variance (s^2) estimates σ^2.

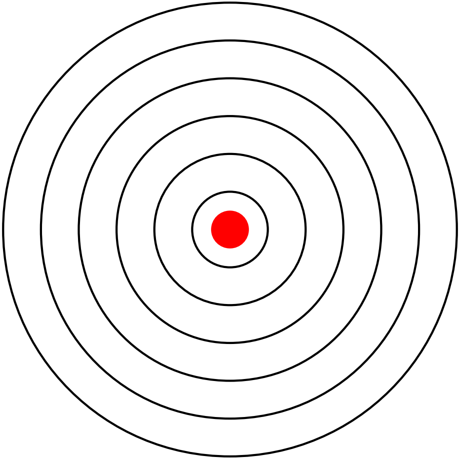
Basic Properties of Estimators

- **Unbiasedness**: The expected value of the estimator equals the population parameter (i.e., $E[\hat{\theta}] = \theta$).
- **Consistency**: As the sample size increases, the estimator approaches the true population parameter.
- **Efficiency**: The estimator has the smallest variance among all unbiased estimators.

Bias and Efficiency

- We define the bias of an estimator $\hat{\theta}$ as:
  - Bias $= E[\hat{\theta}] - \theta$
- Assume two different estimators $\{\hat{\theta}_1, \hat{\theta}_2\}$ of the same population parameter θ.
  - We say that the estimator $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if $V(\hat{\theta}_1) < V(\hat{\theta}_2)$
- Ideally, we want to employ an estimator that is unbiased and efficient to approximate a given population parameter of interest.

Bias vs Variance - A Graphical Illustration



Using Estimators to do Inference

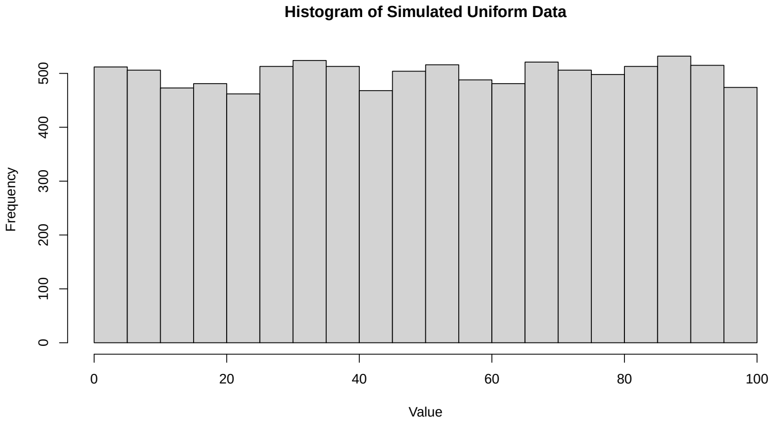
- Say we decide to use an estimator with good properties (i.e., the sample mean). Further, we collect a sample and then compute a specific estimate.
- How can we then **do inferences about the population parameter** (i.e., the true population mean)?
- How can we assess our **degree of confidence** that the specific sample estimate is a good approximation of the true population mean?

Sample Variability - A Simulated Example

- The issue here is **sample variability** or uncertainty. To what extent if we were to collect the sample again we would likely obtain a similar value of the estimate?
- We will do a simulation to illustrate the issue of sample variability.
- **Simulating Uniform(0, 100)**:
  - Generate data from a uniform distribution between 0 and 100.
  - Population size: 10,000.

```
set.seed(789)  # Set seed for reproducibility
pop_data <- runif(10000, min = 0, max = 100)  # Generate 10,000
     random values from a uniform distribution
# Plot histogram of the uniform data
hist(pop_data, breaks = 30,
  main = "Histogram of Simulated Uniform Data",
  xlab = "Value")
```

Plotting the *Population* Data



Random Sampling from Simulated Uniform Data

- Wel'll draw **three random samples from the simulated population (Uniform(0, 100))**:
  - Sample sizes: $N = 20$ (three different samples).
  - Compute the sample mean for each case.

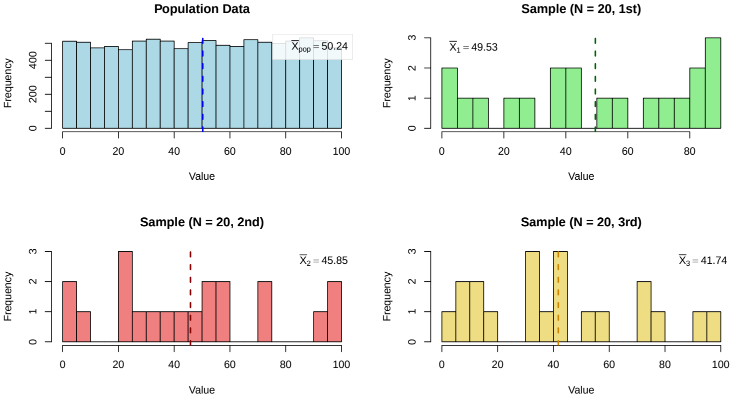
```
# Population mean
true_mean <- mean(pop_data)  # Calculate the true mean of the
     population

# Random samples from the population
set.seed(4321)  # Set seed for reproducibility
sample_20_1 <- sample(pop_data, size = 20)  # First random sample of
      size 20
sample_20_2 <- sample(pop_data, size = 20)  # Second random sample
     of size 20
sample_20_3 <- sample(pop_data, size = 20)  # Third random sample of
      size 20

# Compute sample means
mean_20_1 <- mean(sample_20_1)  # Mean of the first sample
mean_20_2 <- mean(sample_20_2)  # Mean of the second sample
mean_20_3 <- mean(sample_20_3)  # Mean of the third sample
```

Computing the True Mean and Sample Means

- **True mean of the population**:
  - 50.24
- **Sample means**:
  - $N_1 = 20$, $\bar{X}_1 = 49.53$
  - $N_1 = 20$, $\bar{X}_2 = 45.85$
  - $N_1 = 20$, $\bar{X}_3 = 41.74$

Visualizing the Sampling Process

- We can observe differences in the shape of the distribution across the samples and compared to the population distribution.



Understanding Sampling Variability

- **Imagine we had only done the process once** and obtained a sample mean equal to 49.53. How much confidence should we place on that estimate?
- Our small exercise shows some **variability**: taking the sample again changes the estimate.
- So, given the sample size, how likely is it that we would get a wide range of different values or a small one?
- To answer this, we will introduce the concept of **sampling distributions**.

Estimators and Sampling Distributions

- **Sampling Distribution**:
  - The probability distribution of an estimator.
  - **Key Concept**: Sampling distributions allow us to understand the variability of estimators.
- **Example**: Distribution of the sample mean $\bar{X}$ when repeatedly sampling from the population.

Example: Sampling Distribution for a Normal Population

- Assume a continuous random variable $X \sim N(\mu, \sigma^2)$ – aka "*a normal population*." Assume we draw a sample of size n.
- What is the sampling distribution of the sample mean $\bar{X}$?
- We can start answering this question by deriving $\bar{X}$ true mean and standard deviation (also known as standard error).
- Following the derivation included in last week's lecture:
  - $E[\bar{X}] = \mu$
  - Standard Error: $SE(\bar{X}) = SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$

Standard Error vs Standard Deviation

- **Standard Deviation (σ)**:
  - Measures the spread of a population.
- **Standard Error (SE)**:

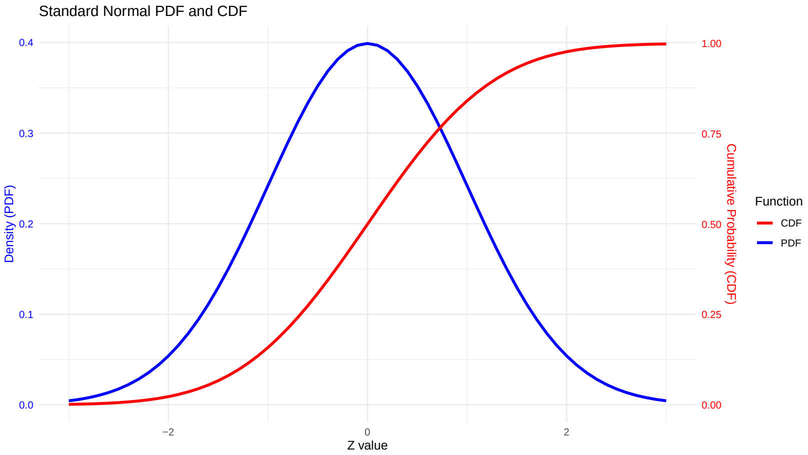
$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

  - Measures the spread of an estimator (e.g., the sample mean).
  - Simply put, the Standard Error is the standard deviation of an estimator (recall that estimators *are random variables*)
- **Difference**: SE decreases as the sample size increases, reflecting more precise estimates.

Detour: Quick Review of the Normal Distribution

- **Probability Density Function (PDF)**:
  - $$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
  - Where μ is the mean and σ is the standard deviation.
- **Properties**:
  - The mean, median, and mode are all equal.
  - Symmetrical PDF.
  - Approximately 68% of the data falls within 1 standard deviation of the mean, 95% within 2, and 99.7% within 3 (aka, **the empirical rule**).
- **Standard Normal Distribution**: A normal distribution with mean 0 and standard deviation 1, often denoted by $Z \sim N(0, 1)$.

Plotting PDF and CDF of Standard Normal Distribution



Cumulative Distribution Function (CDF) of Standard Normal

z	CDF	z	CDF	z	CDF	z	CDF
-3.00	0.00	-1.40	0.08	0.20	0.58	1.80	0.96
-2.80	0.00	-1.20	0.12	0.40	0.66	2.00	0.98
-2.60	0.00	-1.00	0.16	0.60	0.73	2.20	0.99
-2.40	0.01	-0.80	0.21	0.80	0.79	2.40	0.99
-2.20	0.01	-0.60	0.27	1.00	0.84	2.60	0.99
-2.00	0.02	-0.40	0.34	1.20	0.88	2.80	1.00
-1.80	0.04	-0.20	0.42	1.40	0.92	3.00	1.00
-1.60	0.06	0.00	0.50	1.60	0.94		

Table 1: CDF of Standard Normal Distribution, $CDF(z) = \Pr(Z < z)$

Computing Probabilities Using the Standard Normal

- **Computing the Probability** $P(Z < 1) =$ **0.84** :

```
P_Z_less_than_a <- pnorm(1)  # pnorm(a) is a R function that
     computes CDF(a) for the std. normal distr.
P_Z_less_than_a
```

```
## [1] 0.8413447
```

- **Computing the Probability** $P(Z < -0.8) =$ **0.21**:

```
P_Z_less_than_b <- pnorm(-0.8)  # Compute the prob. that Z<-0.8
P_Z_less_than_b
```

```
## [1] 0.2118554
```

- **Probability** $P(-0.8 < Z < 1) =$ **0.63**:

```
P_between_a_and_b <- pnorm(1) - pnorm(-0.8)  # Compute the prob.
     that Z is between -0.8 and 1
P_between_a_and_b
```

```
## [1] 0.6294893
```

Z-Scores & Probability Computations

- **Z-Score Formula**:

$$Z = \frac{X - \mu}{\sigma}$$

- **Theorem 1**: If $X \sim N(\mu, \sigma^2)$, then Z follows a standard normal distribution $N(0, 1)$.
  - **Takeaway**: If we can **assume** $X \sim N(\mu, \sigma^2)$, then we can standardize any value of X using the Z-score to compute probabilities under the standard normal distribution.
- **Example**: Given $X \sim N(50, 25)$, compute $P(X < 55)$ using the standard normal CDF table:
  - Note $\sigma = \sqrt{\sigma^2} = \sqrt{25} = 5$.
  - Convert $X = 55$ to $Z \longrightarrow Z = \frac{X-50}{5} = \frac{55-50}{5} = 1$
  - Use a Z-table to find the corresponding probability, $P(X < 55) = P(Z < 1) = 0.84$.

Z-Scores & Probability Computations

- Given $X \sim N(\mu, \sigma^2)$, how do we use the standard normal CDF table if we want to compute $P(X > b)$?
- **Example 2**: Given $X \sim N(50, 25)$, compute $P(X > 42)$ using the standard normal CDF table:
  - Note $\sigma = \sqrt{25} = 5$.
  - Convert $X = 42$ to Z: $Z = \frac{42-50}{5} = -1.6$
  - Use a Z-table to find the probability of the complement of $X > 42$: $P(X < 42) = P(Z < -1.6) = 0.06$,
  - Then use the complement rule: $P(X > 42) = 1 - P(X < 42) = 1 - P(Z < -1.6) = 1 - 0.06 = 0.94$

Wrapping Up Normal CDFs $\longrightarrow$ Back to Sampling Distributions

- We took a short **detour** to review the **standard normal distribution** and how to:
  - Compute probabilities using the **CDF of** $Z \sim N(0,1)$
  - Convert values from $X \sim N(\mu, \sigma^2)$ into **Z-scores**
  - These tools are essential for doing inference with **normally distributed estimators**.
- Now, let's return to our core question:
  - What is the distribution of the **sample mean** $\bar{X}$ when each $X_i \sim N(\mu, \sigma^2)$?
- Key observation:
  - $\bar{X}$ is a **linear combination (or scaled sum)** of the individual X_i's.
  - So, to fully understand $\bar{X}$, we first need to understand: The distribution of **sums of normal random variables**.

Distribution of a Sum of Normal RVs

- Assume there is a sequence of independent draws of Y, denoted by y_i, all following a normal distribution $Y \sim N(\mu, \sigma^2)$. Define the sum of these draws as:

$$S = \sum_{\forall i} y_i = y_1 + y_2 + \cdots + y_N$$

- What is the distribution of S?
- **Theorem 2**: If $Y_1, Y_2, \ldots, Y_N$ are independent and identically distributed random variables with $Y_i \sim N(\mu, \sigma^2)$, then the sum S also follows a normal distribution:

$$S \sim N(E[S], V[S])$$

  - This applies to any linear combination of the form: $S = \sum_{\forall i} a_i \cdot y_i$ in which a_i corresponds to fixed coefficient (not RVs).

Sampling Distribution of $\bar{X}$ when $x_i \sim N(\mu, \sigma^2)$

- Assume there is a series of independent draws of X, denoted by x_i, all following a normal distribution $X \sim N(\mu, \sigma^2)$.
- Define the sample mean as a linear combination:

$$\bar{X} = \frac{\sum_{\forall i} x_i}{N} = \frac{x_1}{N} + \frac{x_2}{N} + \cdots + \frac{x_N}{N}$$

- What is the distribution of $\bar{X}$?
- **Theorem 3**: Applying the previous theorem: $\bar{X} \sim N(E[\bar{X}], V[\bar{X}])$. Thus,

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{N}})$$

Sampling Distribution of the Sample Mean for a Normal Population

- Assume $X \sim N(\mu = 50, \sigma^2 = 25)$. Then, **compute the probability** $P(\bar{X} > 51)$ assuming a sample size of $n = 100$ using a standard normal CDF table:

```
# Given mean (mu), standard deviation (sigma), and sample size (n)
mu <- 50  # Population mean
sigma <- 5  # Population standard deviation
n <- 100  # Sample size
b <- 51  # Threshold value for sample mean
# Standard error
SE <- sigma / sqrt(n)  # Calculate the standard error of the sample
     mean
# Z-score
Z_b <- (b - mu) / SE  # Calculate the Z-score for the sample mean
     threshold
# Probability P(bar(X) > b)
P_X_greater_than_b <- 1 - pnorm(Z_b)  # Calculate the prob. that the
      sample mean > b
P_X_greater_than_b
```

```
## [1] 0.02275013
```

Recap, Sampling Distributions

- Starting with a known distribution for a random variable under examination, we can derive the exact sampling distribution of the sample mean.
- The sampling distribution can be used as a tool to measure the underlying uncertainty of a given estimate, which allows us to **perform statistical inference**.
- Yet, in most cases, we do not know the underlying distribution of a random variable we are studying (i.e., the population distribution is unknown).
- *How can we then perform statistical inference with a sampling estimate when the population distribution is unknown*?

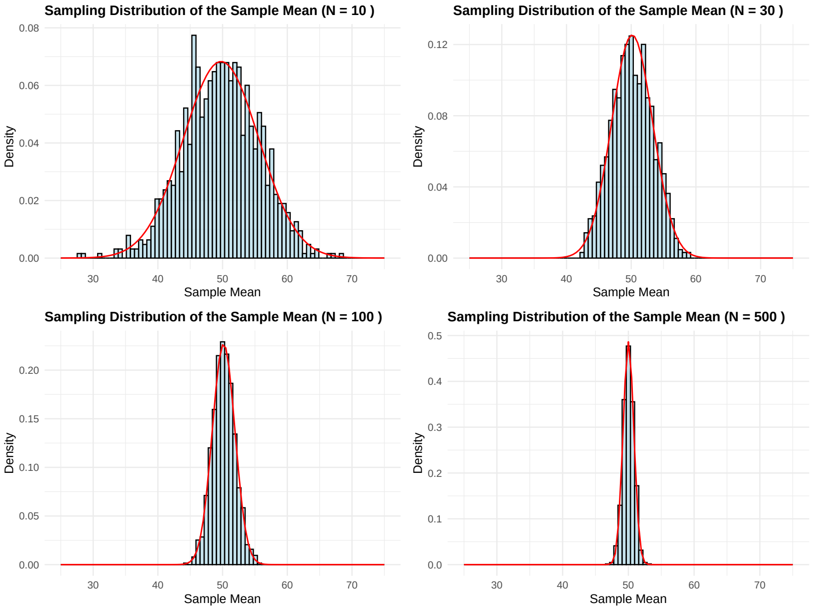
The Central Limit Theorem (CLT)

- **Theorem**: For a sufficiently large sample size, the sampling distribution of the sample mean $\bar{X}$ is approximately normal, *regardless* of the population's distribution.
- **Key Implication**: This enables us to use *Z-score transformations* to approximate probabilities for $\bar{X}$, even when the original variable X is not normally distributed.
- **Why It Matters**: The CLT justifies inference on population parameters using sample data, making it foundational for hypothesis testing and confidence intervals.

Assumptions of the Central Limit Theorem (CLT)

- For the CLT to hold, the following conditions must be met:
  - The sample consists of **independent** observations
    - □ Simple Random Sampling (SRS) supports the independence assumption
  - The observations are drawn from the **same distribution**
  - The underlying distribution has a **finite mean** and **finite variance**
  - The **sample size is sufficiently large**
- Notes:
  - "Sufficiently large" depends on the **shape of the population**:
    - □ For populations with symmetric distributions: $n \geq 30$ often works well
    - □ Skewed or heavy-tailed distributions may require larger n

Simulated Example of CLT (Population is *Uniform*$(20, 80)$)



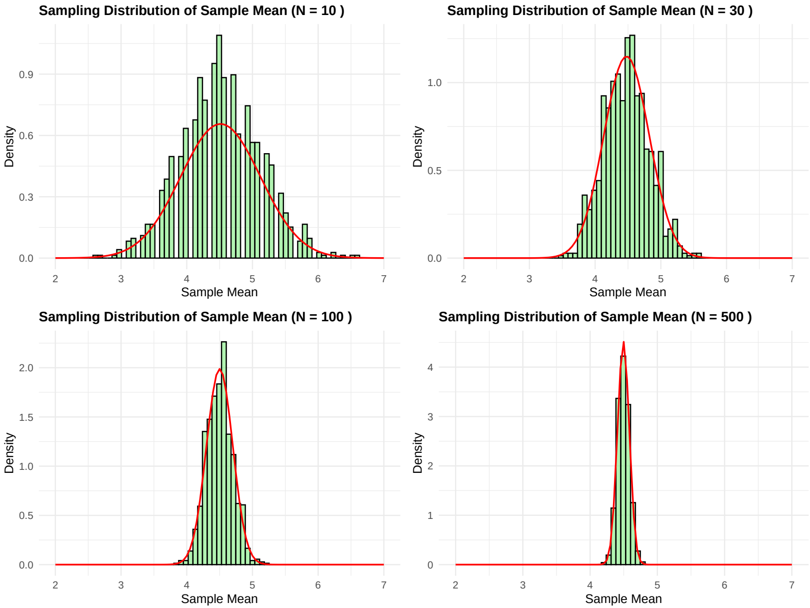
Sampling Distribution of the Sample Mean (N = 10)

Sampling Distribution of the Sample Mean (N = 30)

Sampling Distribution of the Sample Mean (N = 100)

Sampling Distribution of the Sample Mean (N = 500)

CLT Example 2 (Pop. is $Binomial(n = 20, p = 0.15)$)



Z-Scores, the CLT, & Probability Computations

- Invoking the CLT **assume** $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$:
  - Use the Z-score formula to standardize the sample mean $\bar{X}$ for probability computations.
  - **Z-Score Formula**:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- **Example**:
  - Given $\bar{X} \sim N(50, \frac{25}{100})$ (where $\mu = 50, \sigma^2 = 25, n = 100$), compute $P(\bar{X} > 51)$.
  - Convert to Z:

$$Z = \frac{51 - 50}{\frac{5}{\sqrt{100}}} = 2$$

  - Use a Z-table to find the probability corresponding to $Z = 2$.

From Theory to Practice: The Plug-in Principle & the CLT

- For a large n, the Central Limit Theorem assumes:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

where μ and σ are **population parameters** (usually unknown).

- In practice, we don't know these parameters.
  - We apply the **plug-in principle**: use sample statistics to estimate unknown parameters. This is justfied by the Law of Large Numbers (LLN).
- Replace:
  - μ with $\bar{Y}$ (sample mean)
  - σ with s (sample standard deviation)
- This gives the approximation:

$$\bar{Y} \approx N\left(\bar{Y}, \frac{s^2}{n}\right)$$

The Law of Large Numbers (LLN)

- **Law of Large Numbers** (LLN): As the sample size n increases, the *sample mean* $\bar{X}_n$ converges to the *population mean* μ.
  - More formally: If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with finite mean μ, then:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu \quad \text{as } n \to \infty$$

    (Convergence in probability)
- **Key assumptions**. The observations must be:
  - *Independent*
  - *Identically distributed (i.i.d.)*
  - With a *finite expected value* $\mathbb{E}[X] = \mu$
- **Why it matters**: With enough independent data, the sample mean gets arbitrarily close to the true mean.

Using the CLT - Practical Example.

Suppose we have a random sample of a random variable y_i of sample size $N = 400$. Further, suppose $\bar{Y} = 23.5$ and $s = 20$ (sample sd). What is the probability that the sample mean is higher than 25?

```
# Given values
mean_y <- 23.5  # Sample mean
s <- 20  # Sample standard deviation
N <- 400  # Sample size
# Calculate the standard error of the mean
se_y <- s / sqrt(N)
# Calculate the probability directly
prob_calc <- 1 - pnorm(25, mean = mean_y, sd = se_y, lower.tail =
    TRUE)  # lower.tail = TRUE gives P(X < x)
# Display the probability
prob_calc
```

```
## [1] 0.0668072
```

- **Result**: *If we draw a new sample of same size*, the probabiltiy that the sample mean is greater than 25 is app. 6.7%.

Conclusion

- **Key Takeaways**:
  - Estimators help us understand population parameters using sample data.
  - Sampling distributions allows to measure uncertainty in a given estimate.
  - The **Central Limit Theorem** is foundational for statistical inference.
  - Z-scores are useful for probability calculations under normal assumptions or when we invoke the CLT.

Introduction to Statistical Methods in Political Science

Lecture 8: Statistical Inference for One Proportion

Ignacio Urbina

Ph.D. Candidate in Political Science

Learning Objectives

- **Understand:**
    - Confidence Intervals for p
    - Hypothesis Testing for p

Sample Proportion - Sampling Distribution

Inference with Proportions

- **Key Terms:**
  - **Population Proportion (p)**:

$$p = \frac{\text{Number of successes in population}}{\text{Total population size}} = \frac{S}{N}$$

  - **Sample Proportion ($\hat{p}$)**:

$$\hat{p} = \frac{S_n}{n}$$

$$S_n = \text{Number of successes in the sample}, \quad n = \text{Sample size}$$

- **Purpose:** Infer p from $\hat{p}$

Intuition Behind Inference

- **Concept:** Statistical inference involves estimating an unknown population parameter based on sample data.
- **Sample Variability:** Measuring the same quantity multiple times with slight variations each time.

$$\text{Repeated sampling} \Rightarrow \hat{p}_1, \hat{p}_2, \ldots, \hat{p}_k$$

- **Example**: Assume you collect data from a sample of size n, and you compute the proportion of people vaccinated against some virus, $\hat{p}$.

Visualizing Sample Variability in Vaccination Rates

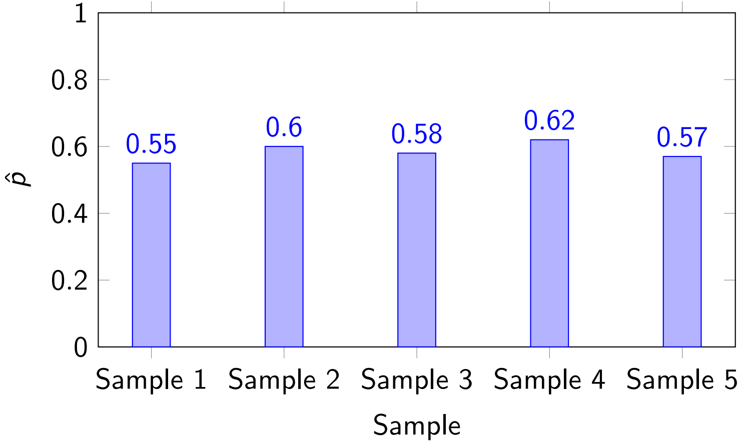


Figure: Variability in Sample Proportions ($\hat{p}$)

Sample Proportion

Let:

$\{X_i\}_{i=1}^{N}$ be a sample collected via simple random sampling

$$X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Note that any X_k and X_j:

- Are identically distributed. So, $E[X_k] = E[X_j] = p$, and $V[X_k] = V[X_j] = p(1-p)$.
- Are independent. So, this implies the following variance of a linear combination of two values, $V((X_k + X_j)/2) = (1/4) \cdot V(X_k + X_j) = (1/4) \cdot [p(1-p) + p(1-p)] = \frac{p(1-p)}{2}$

Expectation and Variance

Population mean:

$$\mathbb{E}[\hat{p}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = p$$

Variance:

$$\mathrm{Var}(\hat{p}) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)$$

$$= \frac{1}{n^2}\cdot n \cdot p(1-p) = \frac{p(1-p)}{n}$$

Sample Variability and Sampling Distribution

- **Repeated Sampling:** Each sample yields a different $\hat{p}$.
- **Objective:** Quantify the variability of $\hat{p}$.

$$\text{Variance of } \hat{p} = \frac{p(1-p)}{n}$$

$$\text{Standard Error (SE)} = \sqrt{\frac{p(1-p)}{n}}$$

- **Sampling Distribution:** If the sample size is large, we can use the CLT to approximate the sampling distribution of $\hat{p}$:

$$\hat{p} \approx N\left(p, \frac{p(1-p)}{n}\right)$$

Confidence Intervals

What is a Confidence Interval?

- A confidence interval (CI) provides a range of values within which we are fairly certain the true population parameter (e.g., a proportion or mean) lies.
- CIs give us an idea of how reliable our estimate is, based on our sample data.
- They quantify the uncertainty around our estimate, offering a range likely to contain the true value.

Confidence Interval Formula

- For an unknown population proportion p, in large samples the sample proportion $\hat{p} = \frac{S_n}{n}$ is approximately normally distributed:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx Z \sim N(0,1)$$

- To calculate a 95% confidence interval, we utilize:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

Calculating the Confidence Interval Range

- Translating the standardization back to our interval:

$$P(-1.96 \leq Z \leq 1.96) = 0.95$$

$$P(-1.96 \leq \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq 1.96) \approx 0.95$$

$$P\left(\hat{p} - 1.96\sqrt{\frac{p(1-p)}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{p(1-p)}{n}}\right) \approx 0.95$$

- We are 95% confident that the true population parameter p lies within this calculated interval.

Logic Behind the Confidence Interval Derivation

- We start by making a reasonable assumption about the sampling distribution for $\hat{p}$.
- Then we standardize $\hat{p}$ into a Z-score, and estimate the variability due to sampling using the standard normal.
- When we say $P(-1.96 \leq Z \leq 1.96) = 0.95$, we mean that 95% of possible outcomes will fall within this ± 1.96 range under the normal distribution.
- Then, by using the properties of the normal distribution, we define a range within which the true proportion p likely lies.

The Anatomy of a Confidence Interval

- Is a procedure that depends on realized values for a random variables, in this case, $\hat{p}$.
- Thus, it is subject to sample variability.
- Note that for each given confidence interval, either the true proportion is or isn't contained. The true proportion is a fixed quantity.
- We ask, what percentage of the time will the confidence interval capture the true proportion? That is the confidence level, or one minus alpha $(1 - \alpha)$. Alpha is the significance level.
- Over the long run (note this thought exercise is informed by the sampling distribution of $\hat{p}$), we expect that p will be included $(1 - \alpha)$% of the time.

Intuition Behind Confidence Intervals

Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward 95%.

Estimating the Standard Error (SE)

- The standard error for the proportion is given by:

$$\mathrm{SE} = \sqrt{\frac{p(1-p)}{n}}$$

- Since p is unknown, by the plugin principle, we use the sample proportion $\hat{p}$ to estimate it:

$$\mathrm{SE} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Margin of Error (MOE)

- The margin of error (MOE) is a statistic expressing the amount of random sampling error in the results of a survey.
- The margin of error is found by multiplying the SE by the critical value (1.96 for a 95% confidence level):

$$\text{MOE} = 1.96 \times \text{SE}(\hat{p})$$

- For a general confidence interval level of $(1-\alpha)$:

$$\text{MOE} = z_{1-\frac{\alpha}{2}} \times \text{SE}(\hat{p})$$

where $z_{1-\frac{\alpha}{2}}$ is the critical value from the standard normal distribution.

Constructing the Confidence Interval

- Finally, the confidence interval is given by:

$$[\hat{p} - \text{MOE}, \; \hat{p} + \text{MOE}]$$

- For a 95% CI:

$$\text{MOE} = 1.96 \times \text{SE}(\hat{p}) = 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- This interval provides an estimated range that, with 95% confidence, contains the true population proportion p.

General Formula for Confidence Intervals for p

- Invoking the CLT:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$

- Thus, the Z-score standardized form:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$$

- To construct a confidence interval with confidence level $1 - \alpha$, we use the critical value $z_{1-\alpha/2}$ such that:

$$P\left(Z < z_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}$$

- The general confidence interval for p becomes:

$$\hat{p} \pm z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Steps to Calculate a 99% Confidence Interval

1. **Determine the Confidence Level and Critical Value:**
   - For a 99% confidence interval, the critical value from the normal distribution is approximately $Z = 2.576$.
2. **Estimate the Standard Error (SE):**
   - Use the sample proportion $\hat{p}$ to estimate SE:

$$\text{SE} \approx \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

3. **Calculate the Margin of Error (MOE):**
   - Multiply the SE by the critical value:

$$\text{MOE} = 2.576 \times \text{SE}$$

4. **Construct the Confidence Interval:**
   - Add and subtract the MOE from the sample proportion $\hat{p}$:

$$[\hat{p} - \text{MOE}, \ \hat{p} + \text{MOE}]$$

   - This interval provides a 99% confidence range for the true population proportion p.

Applied Example: Computing 95% and 99% Confidence Intervals

Problem Statement:

- A public health survey finds that out of a sample of 400 people, 120 are vaccinated.
- We wish to calculate the 95% and 99% confidence intervals for the true proportion of vaccinated individuals in the population.

Given Data:

- Sample size $(n) = 400$
- Sample proportion $(\hat{p}) = \frac{120}{400} = 0.300$

Step 1: Calculate Standard Error (SE)

- **Formula:** $\text{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- **Calculation:**

$$\begin{aligned} \text{SE} = \sqrt{\frac{0.300 \times (1 - 0.300)}{400}} = \sqrt{\frac{0.300 \times 0.700}{400}} &= \sqrt{\frac{0.210}{400}} \\ &= \sqrt{0.000525} \\ &\approx 0.023 \end{aligned}$$

Step 2: Calculate 95% Confidence Interval

- **Critical Value:** For a 95% confidence level, $Z_{1-0.05/2} = Z_{0.975} = 1.96$
- **Margin of Error (MOE):**

$$\mathrm{MOE} = 1.96 \times \mathrm{SE} = 1.96 \times 0.023 \approx 0.045$$

- **95% CI Calculation:**

$$\mathrm{CI} = \hat{p} \pm \mathrm{MOE} = 0.300 \pm 0.045$$

$$95\%\ \mathrm{CI} = [0.255, 0.345]$$

Step 3: Calculate 99% Confidence Interval

- **Critical Value:** For a 99% confidence level (i.e., $\alpha = 0.01$), $Z_{1-0.01/2} = Z_{0.995} = 2.576$
- **Margin of Error (MOE):**

$$\text{MOE} = 2.576 \times \text{SE} = 2.576 \times 0.023 \approx 0.059$$

- **99% CI Calculation:**

$$\text{CI} = \hat{p} \pm \text{MOE} = 0.300 \pm 0.059$$

$$99\% \text{ CI} = [0.241, 0.359]$$

Interpretation of 95% and 99% Confidence Intervals

- **95% CI (0.255, 0.345):**
  - We are 95% confident that the true proportion of vaccinated individuals is between 25.5% and 34.5%.
- **99% CI (0.241, 0.359):**
  - We are 99% confident that the true proportion of vaccinated individuals is between 24.1% and 35.9%.
- **Observations:**
  - The 99% CI is wider than the 95% CI, reflecting greater confidence and a broader range for the estimate.

Which Percent of Future Sample Proportions Will Fall in Our CI?

- A 95% confidence interval captures the **true parameter** p in 95% of repeated samples.
- But the interval is built around one observed sample proportion $\hat{p}_{\text{orig}}$.
- **Future sample proportions** $\hat{p}_{\text{new}}$ follow a distribution centered at p, not at $\hat{p}_{\text{orig}}$.
- Therefore, the long-run probability that future $\hat{p}_{\text{new}}$ falls within the **fixed CI** from the past will be most likely lower than 95%.

But note this is a **moot question**. We really care about p, not $\hat{p}$!

Hypothesis Testing

Introduction to Hypothesis Testing

- **Why Test Hypotheses?**
  - Often, we want to know if a certain belief or assumption about a population is likely to be true based on our sample data.
  - For example, we might wonder, “Is the vaccination rate really 30% in the general population, or is it different?”
- **Hypothesis Testing:** A systematic way to check if our data supports or refutes our initial belief.
  - Imagine having a statement about a population—our hypothesis—and then using data to evaluate if there’s enough evidence to challenge that statement.

How Hypothesis Testing Works

- **Formulating Two Opposing Claims:**
  - We start with two possible claims about a population measure, such as a proportion.
  - One claim (the *null hypothesis*) represents a baseline assumption, often suggesting "no change" or "no effect."
  - The other claim (the *alternative hypothesis*) suggests there's a meaningful difference from the null.
- **Gathering Evidence:**
  - Using sample data, we evaluate if there's enough evidence to support or refute our baseline assumption.
  - Just as in a courtroom, we start with a presumption (the null hypothesis is true) and only reject it if the evidence is strong.
- **Making a Decision:**
  - If our sample data aligns well with the null hypothesis, we "fail to reject" it.
  - If the data strongly contradicts the null hypothesis, we "reject" it in favor of the alternative hypothesis.

Introducing the Test Statistic

- **What is a Test Statistic?**
  - A test statistic is a number we calculate from our sample data to help us decide if our sample provides enough evidence to challenge the null hypothesis.
  - Think of it as a "score" that tells us how far our sample result is from what we would expect if the null hypothesis were true.
- **The Formula for the Test Statistic (for Proportions):**

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- **Defining Each Term:**
  - $\hat{p}$: The **sample proportion**.
  - p_0: The **null hypothesized proportion**, or the value of the proportion we are assuming is true under the null hypothesis.
  - n: The **sample size**, or the number of individuals in our sample.
  - $\sqrt{\frac{p_0(1-p_0)}{n}}$: The **standard error (SE)** of $\hat{p}$ assuming the null hypothesis.

Understanding the Test Statistic and its Interpretation

- **What Units Does the Test Statistic Use?**
  - The test statistic Z is measured in **standard error units**.
  - It represents how many standard errors the sample proportion $\hat{p}$ is away from the null hypothesis proportion p_0.
- **Why Large Absolute Values of Z Suggest Evidence Against the Null:**
  - If Z is large (either positive or negative), it means $\hat{p}$ is far from p_0 in terms of *expected variability under the assumption of the null hypothesis*, which may indicate that the null hypothesis is unlikely.

Formal Setup of a Hypothesis Test

- **Defining Hypotheses:**
  - We write the **null hypothesis** as: $H_0 : p = p_0$
  - The **alternative hypothesis** represents what we want to test against H_0. For example:
    - □ $H_a : p \neq p_0$ (two-tailed)
    - □ $H_a : p > p_0$ (one-tailed, right)
    - □ $H_a : p < p_0$ (one-tailed, left)
- **Significance Level (α):**
  - α is the threshold probability for deciding when to reject H_0.
  - Common values for α are 0.05 or 0.01, indicating a 5
  - If the probability of observing our test statistic (or more extreme) under H_0 is less than α, we reject H_0.
  - α is a standard, carefully chosen rule that guides us on when to be confident enough to reject H_0 based on how unlikely our observed data is under the null hypothesis.

Intuition of the Significance Level

- **Understanding α as a Tolerance for Error:**
  - The significance level α defines how much evidence we need to reject H_0.
  - It represents our tolerance for being wrong—a boundary for the probability of making a Type I error (rejecting H_0 when it's actually true).
  - Typical values (e.g., $\alpha = 0.05$) mean we accept up to a 5% risk of mistakenly rejecting H_0.
- **The Rejection Region:**
  - The rejection region is determined by α and lies at the "extreme ends" of the distribution under H_0.
  - If our test statistic falls in this region, it suggests that our observed result is too rare under H_0 for us to retain it with confidence.
  - Thus, if our result is in the rejection region, we are willing to reject H_0 because it fits our set threshold for "unusual" results.

Example: Hypothesis Test for Voting Preference

- **Context:** Pew Research surveyed 1,000 participants on voting preference.
  - 52% of the sample say they will vote for Trump.
  - We want to test if the true proportion who will vote for Harris is 51% (i.e., the true proportion of Trump voters is 49%).
- **Hypotheses:**
  - Null hypothesis $H_0 : p = 0.49$
  - Alternative hypothesis $H_a : p \neq 0.49$ (two-tailed test)
- **Significance Level:** Set $\alpha = 0.05$

Solution: Test Statistic and Conclusion

- **Sample Proportion:** $\hat{p} = 0.52$
- **Standard Error (SE):**

$$\text{SE} = \sqrt{\frac{p_0(1 - p_0)}{n}} = \sqrt{\frac{0.49 \times (1 - 0.49)}{1000}} \approx 0.0158$$

- **Test Statistic:**

$$Z = \frac{\hat{p} - p_0}{\text{SE}} = \frac{0.52 - 0.49}{0.0158} \approx 1.898$$

- **Decision:**
  - Since $Z = 1.898$ is within the range of $[-1.96, 1.96]$, we **fail to reject** H_0 at $\alpha = 0.05$.
  - Conclusion: There is not enough evidence to suggest that the true proportion differs from 49%.

Rejection Region, Alpha, and Z Test Statistic

- **Defining the Rejection Region:**
  - The **rejection region** is determined by the significance level (α) and represents the values of the test statistic for which we reject H_0.
  - It depends on whether the test is **one-sided** or **two-sided**.
- **Critical Values and Non-Rejection Region:**
  - For a **two-sided test** at significance level α:

    $$\text{Non-Rejection Region: } -Z_{\alpha/2} \leq Z \leq Z_{\alpha/2}$$

  - For a **one-sided test** (right-tailed) at significance level α:

    $$\text{Non-Rejection Region: } Z \leq Z_{\alpha}$$

  - For a **one-sided test** (left-tailed):

    $$\text{Non-Rejection Region: } Z \geq -Z_{\alpha}$$

  - The critical value Z_{α} is found from standard normal tables corresponding to the chosen α.
- **Decision Rule:**
  - If the calculated Z falls **inside** the rejection region, we **reject** H_0.

Understanding the P-value

- **Definition of P-value:**
  - The P-value is the probability, under the null hypothesis H_0, of obtaining a test statistic as extreme as, or more extreme than, the observed value.
- **Relation to Z Test Statistic:**
  - For a **two-sided test**:

    $$\text{P-value} = 2 \times P(Z \geq |Z_{\text{obs}}|)$$

  - For a **one-sided test** (right-tailed):

    $$\text{P-value} = P(Z \geq Z_{\text{obs}})$$

  - For a **one-sided test** (left-tailed):

    $$\text{P-value} = P(Z \leq Z_{\text{obs}})$$

- **Interpreting the P-value:**
  - A small P-value indicates strong evidence against H_0.
  - A large P-value suggests that the observed data is consistent with H_0.

P-value and Different Significance Levels

- **Assessing Significance with P-value:**
  - The P-value allows us to determine at which significance levels H_0 would be rejected.
  - By comparing the P-value to various α levels, we can see the minimum α for which we would reject H_0.
- **Decision Making:**
  - If P-value $\leq \alpha$, we **reject** H_0.
  - If P-value $> \alpha$, we **fail to reject** H_0.
- **Example:**
  - If P-value $= 0.03$, H_0 would be rejected at $\alpha = 0.05$ but not at $\alpha = 0.01$.

Example: Computing P-value for a One-Sided Test

Problem Statement:

- A political poll indicates that 52% of a sample of 1,000 voters support Candidate A.
- We wish to test if there is evidence that the true proportion supporting Candidate A is greater than 50%.

Hypotheses:

- Null hypothesis $H_0 : p = 0.50$
- Alternative hypothesis $H_a : p > 0.50$ (one-sided test)

Calculations:

- Sample proportion $\hat{p} = 0.52$
- Standard Error SE $= \sqrt{\frac{p_0(1-p_0)}{n}} = \sqrt{\frac{0.50 \times 0.50}{1000}} \approx 0.0158$
- Test Statistic $Z = \frac{\hat{p}-p_0}{\text{SE}} = \frac{0.52-0.50}{0.0158} \approx 1.265$
- P-value $= P(Z \geq 1.265) = 1 - \Phi(1.265) \approx 0.103$

Conclusion:

- At $\alpha = 0.05$, since P-value $= 0.103 > 0.05$, we **fail to reject** H_0.
- There is insufficient evidence to reject the hypothesis that $p_0 = 0.50$

Introduction to Statistical Methods in Political Science

Lecture 9: Inference for Two Proportions

Ignacio Urbina

Ph.D. Candidate in Political Science

Motivation: Inference for Categorical Differences

- Categorical variables represent data sorted into distinct groups or categories.
- Often, we are interested in comparing proportions of a categorical outcome between two groups.
- Example: In a survey, respondents might be grouped by region to see if a higher proportion of people in one region favor a particular policy option compared to another.
- These comparisons help answer questions such as:
  - Does the proportion of people supporting stricter environmental regulations differ between urban and rural areas?
  - Is there a difference in preferred government spending options between younger and older age groups?

Introduction to Comparing Proportions

- In statistics, we often compare two independent groups to understand differences in proportions.
- Examples include comparing vaccination rates between cities or customer satisfaction across products.
- Proportions are used in medicine, marketing, public health, and social sciences.
- These comparisons help identify significant differences in behaviors, treatments, or characteristics.

Example Scenarios

- Medical studies: Comparing success rates of medications.
- Marketing: Evaluating customer satisfaction between products.
- Social sciences: Analyzing behavior differences between demographic groups.

Example: Public Opinion

One common application of comparing two sample proportions is in analyzing public opinion across different demographic groups:

- Consider two groups: those under 30 and those over 60.
- Survey both groups to assess their support for the current president.
- Let $\hat{p}_1$ be the proportion of support among those under 30 and $\hat{p}_2$ among those over 60.

This analysis helps us understand how support varies with age, crucial for policy making and election strategies.

Example: Psychological Experiment

In psychological research, comparing sample proportions evaluates the effect of different treatments:

- Study testing two interventions on anxiety reduction.
- One group receives cognitive behavioral therapy (CBT); another receives mindfulness-based stress reduction (MBSR).
- Let $\hat{p}_1$ be the proportion reporting significant anxiety reduction with CBT, and $\hat{p}_2$ with MBSR.

This comparison provides insights into which intervention is more effective.

Notation and Definitions

- Let p_1 and p_2 represent the population proportions for two independent groups.
- $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions from each group.
- Sample sizes: n_1 for Group 1 and n_2 for Group 2.

Introduction to Sample Proportions

Consider two independent samples where:

- Sample 1: n_1 observations with proportion $\hat{p}_1$.
- Sample 2: n_2 observations with proportion $\hat{p}_2$.

We are interested in the statistic $\hat{p}_1 - \hat{p}_2$, the difference between two sample proportions.

- Our goal is to estimate the difference $p_1 - p_2$ and determine if it is significantly different from zero.
- This analysis helps assess if observed differences could be due to chance or represent a true population difference.

Review of Expectations and Variances

Expectations:

- For random variables X and Y, and constants a, b:

$$E(aX + bY) = aE(X) + bE(Y)$$

Variances:

- For independent random variables X and Y, and constants a, b:

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

Expectation of the Statistic

Using linearity of expectations:

$$E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$$

Where p_1 and p_2 are the true population proportions.

Variance of the Statistic

For independent random variables:

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \mathrm{Var}(\hat{p}_1) + \mathrm{Var}(\hat{p}_2)$$

Given:

$$\mathrm{Var}(\hat{p}_1) = \frac{p_1(1-p_1)}{n_1}, \quad \mathrm{Var}(\hat{p}_2) = \frac{p_2(1-p_2)}{n_2}$$

Thus:

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

Standard Error of the Statistic

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\mathrm{Var}(\hat{p}_1 - \hat{p}_2)}$$

Substituting the variances:

$$SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Sampling Distribution of Difference in Proportions

- The distribution of $\hat{p}_1 - \hat{p}_2$ is approximately normal for large sample sizes.
- Under the Central Limit Theorem:

$$\hat{p}_1 - \hat{p}_2 \approx N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

- This allows us to perform hypothesis testing and construct confidence intervals.

Derivation of the Sampling Distribution

- Variance of a sample proportion $\hat{p}$: $\mathrm{Var}(\hat{p}) = \frac{p(1-p)}{n}$
- Variance of the difference:

$$\mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- Standard error (SE):

$$SE = \sqrt{\mathrm{Var}(\hat{p}_1 - \hat{p}_2)}$$

Confidence Interval for Difference in Proportions

- Confidence interval:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE$$

- Where:

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- z^* is the critical value for the desired confidence level (e.g., 1.96 for 95% confidence).

Constructing the Confidence Interval

1. Calculate sample proportions: $\hat{p}_1$ and $\hat{p}_2$.
2. Compute the standard error (SE):

$$SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

3. Determine critical value z^* for the desired confidence level.
4. Multiply z^* by SE to find the margin of error (MOE).
5. Construct the confidence interval:

$$(\hat{p}_1 - \hat{p}_2) \pm MOE$$

Applied Example: Comparing Vaccination Rates - Setup

- Sample data:
  - Group 1: $n_1 = 400$, $\hat{p}_1 = 0.30$
  - Group 2: $n_2 = 500$, $\hat{p}_2 = 0.25$
- We will calculate the standard error and confidence interval.

Applied Example: Comparing Vaccination Rates - Calculation

- Calculate SE:

$$SE = \sqrt{\frac{0.30 \times 0.70}{400} + \frac{0.25 \times 0.75}{500}} \approx 0.032$$

- 95% Confidence Interval:

$$(0.30 - 0.25) \pm 1.96 \times 0.032$$

$$0.05 \pm 0.063 \implies [-0.013, 0.113]$$

Interpretation of Example Results

- Confidence interval includes zero (-0.013 to 0.113), suggesting no significant difference.
- If the interval excluded zero, it would indicate a statistically significant difference.

Common Misconceptions

- **Misconception 1**: A confidence interval that includes zero proves there is no difference between the two population proportions.
  - **Reality**: It simply suggests we lack sufficient evidence of a difference at the specified confidence level.
- **Misconception 2**: A wider confidence interval means the difference is less likely.
  - **Reality**: A wider interval indicates greater uncertainty, often due to smaller sample sizes.
- **Misconception 3**: A 95% confidence level means the interval has a 95% chance of containing the true difference.
  - **Reality**: A 95% confidence level means that, in the long run, 95% of intervals from multiple independent samples of the same size will contain the true difference.

Summary and Key Takeaways

- Inference for difference in proportions helps compare two independent groups.
- Large samples allow the sampling distribution to be approximately normal.
- Confidence intervals offer a range of plausible values for the true difference.
- The examples illustrated the application and interpretation of these concepts.

Hypothesis Testing for Difference in Proportions

- Beyond estimating the confidence interval, we often want to test whether the observed difference between two proportions is statistically significant.
- Hypothesis testing allows us to determine if there is enough evidence to support a claim about the difference in population proportions.
- Common applications include testing the effectiveness of a new treatment compared to a control or comparing preferences between two groups.

Formulating Hypotheses

- **Null Hypothesis (H_0)**: Assumes no difference between the population proportions.

$$H_0 : p_1 - p_2 = 0$$

- **Alternative Hypothesis (H_a)**: Proposes that there is a difference.

$$\text{Two-tailed test: } H_a : p_1 - p_2 \neq 0$$

$$\text{One-tailed test: } H_a : p_1 - p_2 > 0 \text{ or } p_1 - p_2 < 0$$

- The choice between one-tailed and two-tailed tests depends on the research question.

Test Statistic

- The test statistic measures how far the sample difference is from the null hypothesis, relative to the standard error.
- Under H_0, the test statistic is:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_0}$$

- Where SE_0 is the standard error calculated under the assumption that H_0 is true.

Standard Error Under Null Hypothesis

- Under $H_0 : p_1 = p_2 = p$, we pool the sample proportions to estimate the common population proportion p.
- The pooled sample proportion $\hat{p}$ is:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

- Where x_1 and x_2 are the number of successes in each sample.
- The standard error under H_0 is:

$$SE_0 = \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

Calculating the Test Statistic

1. Compute the pooled sample proportion $\hat{p}$:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

2. Calculate the standard error SE_0 under H_0:

$$SE_0 = \sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

3. Compute the test statistic z:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{SE_0}$$

Decision Rule and p-value

- **Decision Rule**:
  - Compare the test statistic z to critical values from the standard normal distribution.
  - For a two-tailed test at $\alpha = 0.05$, reject H_0 if $|z| > 1.96$.
- **p-value**:
  - The p-value is the probability of observing a test statistic as extreme as, or more extreme than, the one calculated, assuming H_0 is true.
  - For a two-tailed test:

$$p\text{-value} = 2 \times P(Z > |z|)$$

- **Conclusion**:
  - If the p-value is less than α, reject H_0.
  - Otherwise, fail to reject H_0.

Applied Example: Hypothesis Testing - Setup

- Continuing the vaccination rate example:
  - Group 1: $n_1 = 400$, $\hat{p}_1 = 0.30$, $x_1 = 120$
  - Group 2: $n_2 = 500$, $\hat{p}_2 = 0.25$, $x_2 = 125$
- Test whether there is a significant difference in vaccination rates between the two groups at the 5% significance level.
- Formulate hypotheses:

$$H_0 : p_1 - p_2 = 0 \quad \text{vs.} \quad H_a : p_1 - p_2 \neq 0$$

Applied Example: Hypothesis Testing - Calculations

1. Compute pooled proportion:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{120 + 125}{400 + 500} = \frac{245}{900} \approx 0.272$$

2. Calculate standard error under H_0:

$$SE_0 = \sqrt{0.272 \times 0.728 \left(\frac{1}{400} + \frac{1}{500} \right)}$$
$$\approx \sqrt{0.198 \times 0.0045} \approx 0.0299$$

3. Compute test statistic:

$$z = \frac{0.30 - 0.25}{0.0299} \approx \frac{0.05}{0.0299} \approx 1.675$$

Applied Example: Decision and Conclusion

- Critical value for two-tailed test at $\alpha = 0.05$ is $z_{\alpha/2} = 1.96$.
- Since $|z| = 1.675 < 1.96$, we fail to reject H_0.
- **p-value**:

$$\begin{aligned} p\text{-value} &= 2 \times P(Z > 1.675) \\ &= 2 \times (1 - \Phi(1.675)) \\ &\approx 2 \times 0.04697 = 0.094 \end{aligned}$$

- Since p-value $= 0.094 > 0.05$, we fail to reject H_0.
- **Conclusion**:
  - There is insufficient evidence at the 5% significance level to conclude a difference in vaccination rates between the two groups.

Interpretation of Example Results

- Although the sample proportions differ (30% vs. 25%), the difference is not statistically significant at the 5% level.
- This suggests that the observed difference could be due to random sampling variability.
- It's important to consider sample sizes and variability when interpreting results.

Common Misconceptions

- **Misconception 1**: A non-significant result means there is no difference.
  - **Reality**: It means we do not have sufficient evidence to conclude a difference exists.
- **Misconception 2**: A small p-value indicates a large effect size.
  - **Reality**: The p-value measures evidence against H_0, not the magnitude of the effect.
- **Misconception 3**: Failing to reject H_0 proves H_0 is true.
  - **Reality**: We can never prove H_0; we can only fail to reject it.

Summary and Key Takeaways

- Hypothesis testing for the difference in proportions assesses whether an observed difference is statistically significant.
- The test involves calculating a test statistic under the assumption that the null hypothesis is true.
- Understanding the standard error and using the pooled proportion are critical steps.
- The p-value helps determine the strength of evidence against the null hypothesis.
- Always interpret results in context and be cautious of common misconceptions.

Introduction to Statistical Methods in Political Science

Lecture 10: Sampling Distributions for Estimators of Continuous Variables

Ignacio Urbina

Ph.D. Candidate in Political Science

Sampling Distribution of the Sample Mean $\bar{x}$

Our goal is often to estimate the unknown population mean μ using the sample mean $\bar{x}$ calculated from a random sample $X_1, ..., X_n$.

The sample mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} X_i$ is itself a random variable, as its value depends on the particular sample drawn.

The **sampling distribution of** $\bar{x}$ describes the probability distribution of the possible values of $\bar{x}$ if we were to repeatedly draw samples of size n from the same population.

Key properties of this distribution are its mean $E(\bar{x})$ and its variance $\text{Var}(\bar{x})$ (or standard error $SE(\bar{x})$).

Case 1: Normal Population (Known σ^2)

Assume the underlying population follows a normal distribution, $X_i \sim N(\mu, \sigma^2)$, and the population variance σ^2 is known.

- The sample mean $\bar{x}$ is **exactly** normally distributed.
- Mean of $\bar{x}$: $E(\bar{x}) = \mu$ (unbiased estimator).
- Variance of $\bar{x}$: $\text{Var}(\bar{x}) = \frac{\sigma^2}{n}$.
- Standard Error of $\bar{x}$: $SE(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$.

Distribution:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Standardized Statistic (Z-score):

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Case 2: Large Sample Size (CLT)

What if the population distribution is not normal, or unknown?
Central Limit Theorem (CLT): If the sample size n is sufficiently large ($n \geq 30$), the sampling distribution of $\bar{x}$ will be **approximately** normal, regardless of the shape of the population distribution.

- Mean of $\bar{x}$: $E(\bar{x}) = \mu$.
- Variance of $\bar{x}$: $\text{Var}(\bar{x}) = \frac{\sigma^2}{n}$.

Approximate Distribution:

$$\bar{x} \approx N\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{for large } n$$

The CLT is fundamental because it allows us to use normal distribution methods for inference on μ in many practical situations.

The Plug-In Principle (Large Sample)

Usually, the population variance σ^2 is **unknown**.
Plug-In Principle: Estimate σ^2 using the sample variance s^2:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Estimate the standard error of $\bar{x}$ using s:

$$\text{Estimated } SE(\bar{x}) = \frac{s}{\sqrt{n}}$$

For **large samples** ($n \geq 30$), combining CLT and plug-in:

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} \approx N(0, 1)$$

This justifies Z-procedures for μ with large samples when σ is unknown.

Case 3: Small Sample Size (Unknown σ^2)

What if n is small $(n < 30)$ **and** σ^2 is unknown?
If we assume the population is **normal**, we use the **t-distribution**:

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

The t-distribution accounts for the extra uncertainty from estimating σ^2 with s^2.
Details of inference using the t-distribution for small samples will be covered separately.

Why Compare Two Means? Examples

Comparing two sample means helps answer questions across various fields:

- **Business (Job Satisfaction):** Is average job satisfaction (μ_1) in the IT industry different from that in finance (μ_2)? Compare sample means $\bar{x}_1$ and $\bar{x}_2$.
- **Health Science (Physical Activity):** Does a high-intensity exercise regimen (μ_1) lead to a greater mean decrease in cholesterol than a moderate-intensity one (μ_2)? Compare sample mean decreases $\bar{x}_1$ and $\bar{x}_2$.

The goal is to use the sample difference $\bar{x}_1 - \bar{x}_2$ to infer about the population difference $\mu_1 - \mu_2$.

Setup for Comparing Two Means

Consider two **independent** samples:

- Sample 1: Size n_1, mean $\bar{x}_1$, from population with mean μ_1, variance σ_1^2.
- Sample 2: Size n_2, mean $\bar{x}_2$, from population with mean μ_2, variance σ_2^2.

We focus on the sampling distribution of the statistic:

$$\text{Difference in sample means: } \bar{x}_1 - \bar{x}_2$$

Review: Properties of E and Var

Recall fundamental properties: **Expectations (Linearity):**

$$E(aX + bY) = aE(X) + bE(Y)$$

Variances (for Independent X, Y):

$$\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$$

These rules are key to deriving the properties of $\bar{x}_1 - \bar{x}_2$.

Expectation of the Difference $\bar{x}_1 - \bar{x}_2$

Using linearity of expectation $(a = 1, b = -1)$:

$$E(\bar{x}_1 - \bar{x}_2) = E(\bar{x}_1) - E(\bar{x}_2)$$

Since $E(\bar{x}_1) = \mu_1$ and $E(\bar{x}_2) = \mu_2$:

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

The difference in sample means is an unbiased estimator of the difference in population means.

Variance and SE of the Difference $\bar{x}_1 - \bar{x}_2$

Assuming the two samples are **independent**: Using the variance rule $(a = 1, b = -1)$:

$$\mathrm{Var}(\bar{x}_1 - \bar{x}_2) = \mathrm{Var}(\bar{x}_1) + \mathrm{Var}(\bar{x}_2)$$

Substitute known variances of sample means:

$$\mathrm{Var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

The Standard Error (SE) is the square root of the variance:

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Sampling Distribution of $\bar{x}_1 - \bar{x}_2$ (Large Samples)

If n_1, n_2 are large (CLT), or populations normal (σ's known):

- $\bar{x}_1 \approx N(\mu_1, \sigma_1^2/n_1)$
- $\bar{x}_2 \approx N(\mu_2, \sigma_2^2/n_2)$

Since samples are independent, the difference is also (approx.) normal:

$$\bar{x}_1 - \bar{x}_2 \approx N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

This allows Z-procedures for $\mu_1 - \mu_2$ in these cases.

The Plug-In Principle (Two Means, Large Samples)

When σ_1^2, σ_2^2 unknown, but n_1, n_2 large: Estimate SE using sample variances s_1^2, s_2^2:

$$\text{Estimated } SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The test statistic for $H_0 : \mu_1 - \mu_2 = \Delta_0$ (often $\Delta_0 = 0$):

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \approx N(0, 1)$$

Small Samples: t-Distribution (Two Means - Brief Mention)

If either n_1 or n_2 is small, **and** populations assumed normal, **and** σ_1^2, σ_2^2 unknown:
Use the **t-distribution**. The statistic has the form:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Calculating the correct degrees of freedom (df^*) requires specific methods (e.g., Welch-Satterthwaite) unless variances are assumed equal.
Detailed procedures for the two-sample t-test will be covered separately.

Summary and Key Assumptions

- **Sampling Distributions:** Describe the behavior of statistics ($\bar{x}$, $\bar{x}_1 - \bar{x}_2$) over repeated sampling.
- **CLT:** Crucial for large samples, allows using Normal approx. even for non-normal populations.
- **Plug-in Principle:** Use sample variance(s) s^2 when population variance(s) σ^2 are unknown.
- **Independence:** Formulas for $\text{Var}(\bar{x}_1 - \bar{x}_2)$ require independent samples.
- **Large vs. Small Samples:** Use Z-procedures (based on CLT/Normal) for large samples; use t-procedures (based on t-distribution, requires population normality assumption) for small samples when σ's are unknown.
- **Convergence:** For large n, the t-distribution approaches the N(0,1) distribution.

Introduction to Statistical Methods in Political Science

Lecture 11: Large-Sample Inference for Means

Ignacio Urbina

Ph.D. Candidate in Political Science

Confidence Intervals for Means

Motivation: From Proportions to Means

Last week, we focused on inference for **population proportions (p)**.

- Confidence intervals and hypothesis tests for p using large-sample Z-procedures.

Now, we shift to **quantitative (continuous) data** where the parameter is the **population mean (μ)**.

- Examples: Average height, mean income, time.

Goal: Develop inference methods for μ using CLT and the sampling distribution of $\bar{x}$.

Foundation: The Central Limit Theorem (Recap)

For large samples ($n \geq 30$), the sampling distribution of $\bar{x}$ is approximately normal:

$$\bar{x} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Approximate normality holds regardless of the original population distribution.

Handling Unknown σ & The Z-Statistic

When σ is unknown, we estimate it using the sample standard deviation s (plug-in principle). The resulting test statistic:

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Large-Sample CI for μ: Formula & Structure

Confidence Interval formula:

$$\bar{x} \pm z_{1-\frac{\alpha}{2}} \left(\frac{s}{\sqrt{n}} \right)$$

Structure:

- Point Estimate: $\bar{x}$
- Margin of Error: $z_{1-\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$

Margin of Error for μ CI

Depends on:

1. Critical Value $z_{1-\frac{\alpha}{2}}$
2. Standard Error $SE = \dfrac{s}{\sqrt{n}}$

Thus, the Margin of Error (ME) is:

$$ME = z_{1-\frac{\alpha}{2}} \times SE$$

Interpretation of a Confidence Interval for μ

Example: 95% CI = (5.2, 6.8) Correct Interpretation:

- **"We are 95% confident that μ lies within (5.2, 6.8)."**

Example: CI for Average Commute Time (Problem Statement)

Problem: What is the average daily commute time for workers in a city?
Data: $n = 100$ workers, sample mean commute time $\bar{x} = 32.5$ minutes, sample standard deviation $s = 7.0$ minutes.
Task: Construct a 95% confidence interval for the true mean commute time.

Example: CI for Average Commute Time (Calculation)

Given:

- $n = 100$, $\bar{x} = 32.5$, $s = 7.0$
- 95% confidence level $(\alpha = 0.05) \Rightarrow z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96$

Standard error:

$$SE = \frac{7.0}{\sqrt{100}} = 0.7$$

Margin of error:

$$ME = 1.96 \times 0.7 = 1.372$$

Confidence Interval:

$$(32.5 - 1.372,\ 32.5 + 1.372) = (31.128,\ 33.872)$$

Foundation: Distribution of $\bar{x}_1 - \bar{x}_2$ (Recap)

For large independent samples:

$$\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Handling Unknown Variances & The Z-Statistic (Two Means)

Estimate population variances with sample variances:

$$s_1^2 \quad \text{and} \quad s_2^2$$

Standard error for $\bar{x}_1 - \bar{x}_2$:

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Use z-statistic for large samples.

Large-Sample CI for $\mu_1 - \mu_2$: Formula & Structure

Confidence Interval:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\frac{\alpha}{2}} \times SE$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Interpretation of CI for $\mu_1 - \mu_2$

- Positive Interval: $\mu_1 > \mu_2$
- Interval Contains 0: No significant difference
- Negative Interval: $\mu_1 < \mu_2$

Example: CI for Comparing Study Methods (Problem Statement)

Problem: Do two different study methods lead to different average exam scores among university students?
Data: Method 1 ($n_1 = 50$, $\bar{x}_1 = 85.2$, $s_1 = 8.5$); Method 2 ($n_2 = 60$, $\bar{x}_2 = 81.5$, $s_2 = 9.1$).
Task: Construct a 99% confidence interval for $\mu_1 - \mu_2$.

Example: CI for Comparing Study Methods (Calculation)

Given:

- $n_1 = 50$, $\bar{x}_1 = 85.2$, $s_1 = 8.5$
- $n_2 = 60$, $\bar{x}_2 = 81.5$, $s_2 = 9.1$
- 99% confidence level $(\alpha = 0.01) \Rightarrow z_{1-\frac{\alpha}{2}} = z_{0.995} = 2.576$

Standard error:

$$SE = \sqrt{\frac{8.5^2}{50} + \frac{9.1^2}{60}} \approx 1.68$$

Margin of error:

$$ME = 2.576 \times 1.68 \approx 4.33$$

Confidence Interval:

$$(85.2 - 81.5) \pm 4.33 = 3.7 \pm 4.33$$

$$(-0.63,\ 8.03)$$

Hypothesis Testing for Means

Motivation: Hypothesis Tests for Means

Apply hypothesis testing to population **means (μ)**:

- Steps: Hypotheses, α, Test Statistic, Decision, Conclusion

Example: Average Hours of Sleep (Problem Statement)

Problem: Do college students sleep an average of 7 hours per night?
Data: $n = 64$, $\bar{x} = 6.8$ hours, $s = 0.6$ hours.
Task: Test $H_0 : \mu = 7$ against $H_a : \mu \neq 7$ at $\alpha = 0.05$.

Example: Average Hours of Sleep (Calculation)

Calculate standard error:

$$SE = \frac{0.6}{\sqrt{64}} = 0.075$$

Test statistic:

$$Z = \frac{6.8 - 7.0}{0.075} = -2.67$$

Critical value for a test with $\alpha = 0.05$: ± 1.96.
Since $|-2.67| > 1.96$, reject H_0.

Example: New Drug Effectiveness (Problem Statement)

Problem: Is a new drug more effective at reducing blood pressure than the current standard drug?
Data: Drug 1 ($n_1 = 120$, $\bar{x}_1 = 15.5$, $s_1 = 5.0$); Drug 2 ($n_2 = 100$, $\bar{x}_2 = 13.8$, $s_2 = 4.5$).
Task: Test $H_0 : \mu_1 - \mu_2 = 0$ vs. $H_a : \mu_1 - \mu_2 > 0$ at $\alpha = 0.01$.

Example: New Drug Effectiveness (Calculation)

Calculate standard error:

$$SE = \sqrt{\frac{5.0^2}{120} + \frac{4.5^2}{100}} \approx 0.641$$

Test statistic:

$$Z = \frac{15.5 - 13.8}{0.641} \approx 2.65$$

Critical value: $z_{0.99} \approx 2.33$. Since $2.65 > 2.33$, reject H_0.

Hypothesis Tests with Nonzero Null Values

Previously, we tested $H_0 : \mu = 0$ or $H_0 : \mu_1 - \mu_2 = 0$.
Now, we allow the null hypothesis to state a nonzero value:

- $H_0 : \mu = \mu_0$ where $\mu_0 \neq 0$
- $H_0 : \mu_1 - \mu_2 = \Delta_0$ where $\Delta_0 \neq 0$

Interpretation: We are testing whether the population mean or the difference between two means equals a specific number.

Adjusted Z-Statistic Formulas

One Mean:

$$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where μ_0 is the null hypothesized mean (not necessarily 0).

Two Means:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where Δ_0 is the hypothesized difference (often 0, but not always).

Note: Subtract μ_0 or Δ_0 when forming the numerator.

Example: Average Calories Consumed (Problem Statement)

Problem: Do individuals consume an average of 2500 calories per day?
Data: $n = 49$, $\bar{x} = 2550$ calories, $s = 150$ calories.
Task: Test $H_0 : \mu = 2500$ against $H_a : \mu \neq 2500$ at $\alpha = 0.05$.

Example: Average Calories Consumed (Calculation)

Calculate standard error:

$$SE = \frac{150}{\sqrt{49}} = 21.43$$

Test statistic:

$$Z = \frac{2550 - 2500}{21.43} \approx 2.33$$

Critical value: ± 1.96.
Since $2.33 > 1.96$, reject H_0.

Hypothesis Tests for Two Means: Nonzero Difference

In many situations, we want to test whether the difference between two population means equals a value other than zero.
Examples:

- Testing if a new product improves scores by 5 points compared to an old product.
- Checking if two treatments differ by a clinically significant margin (e.g., 2 mmHg in blood pressure).

Null hypothesis: $H_0 : \mu_1 - \mu_2 = \Delta_0$ where $\Delta_0 \neq 0$.

Adjusted Z-Statistic: Two Means

For large independent samples, the test statistic becomes:

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where:

- Δ_0 is the hypothesized difference between means.
- $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is the standard error.

Key change: Subtract Δ_0 from the observed difference $(\bar{x}_1 - \bar{x}_2)$ in the numerator.

Example: Comparing Manufacturing Processes (Problem Statement)

Problem: Is there a difference of 5 units in mean output between two manufacturing processes?
Data: Process 1 ($n_1 = 40$, $\bar{x}_1 = 105$, $s_1 = 10$); Process 2 ($n_2 = 50$, $\bar{x}_2 = 98$, $s_2 = 12$).
Task: Test $H_0 : \mu_1 - \mu_2 = 5$ against $H_a : \mu_1 - \mu_2 \neq 5$ at $\alpha = 0.05$.

Example: Comparing Manufacturing Processes (Calculation)

Calculate standard error:

$$SE = \sqrt{\frac{10^2}{40} + \frac{12^2}{50}} \approx 2.32$$

Test statistic:

$$Z = \frac{(105 - 98) - 5}{2.32} \approx 0.43$$

Critical value: ± 1.96.
Since $0.43 < 1.96$, fail to reject H_0.

Large Sample Inference with Paired Samples

Paired Samples: When and Why?

Paired data arise when observations are naturally matched:

- Before-and-after measurements (e.g., pre- and post-treatment)
- Measurements on the same individual under two conditions. Hence, before and after data *are not independent*.

Key idea: Reduce the two measurements to a single difference score:

$$d_i = x_{i,1} - x_{i,2}$$

Then apply inference methods to the sample of differences. Key assumption: *Independence across individuals*.

Paired Samples: CI and Hypothesis Test

Confidence Interval for μ_d

For large samples of paired differences ($n \geq 30$), apply CLT:

$$\bar{d} \sim N\left(\mu_d, \frac{s_d^2}{n}\right)$$

CI formula:

$$\bar{d} \pm z_{1-\frac{\alpha}{2}} \times \frac{s_d}{\sqrt{n}}$$

- $\bar{d}$: Mean of differences
- s_d: Std. dev. of differences
- $z_{1-\frac{\alpha}{2}}$: Critical value

Hypothesis Test for μ_d

Test the null hypothesis:

$$H_0 : \mu_d = \mu_{d,0} \quad \text{vs.} \quad H_a : \mu_d \neq \mu_{d,0}$$

Z-statistic:

$$Z = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}}$$

- Compare Z to critical value or use p-value
- Reject H_0 if evidence is strong

Example: Effect of a Sleep Intervention

Problem: Does a cognitive-behavioral sleep intervention improve sleep duration among adults with mild insomnia?
Study Design: A sample of $n = 200$ individuals participated in a 4-week cognitive-behavioral therapy (CBT) program targeting sleep hygiene, routines, and stress reduction. Each participant recorded their average nightly sleep duration *before and after* the program.

Data: Mean difference $\bar{d} = 0.75$ hours, $s_d = 1.2$ hours, $n = 200$.
Task: Construct a 95% confidence interval for the mean increase in sleep duration, μ_d.

$$SE = \frac{1.2}{\sqrt{200}} \approx 0.085, \quad z_{0.975} = 1.96$$

$$ME = 1.96 \times 0.085 \approx 0.17, \quad CI = (0.75 \pm 0.17) = (0.58,\ 0.92)$$

Conclusion: We are 95% confident the CBT program increased sleep duration by 0.58 to 0.92 hours per night.

Summary: Key Formulas and Concepts

Summary Table

Scenario	Confidence Interval (CI)	Hypothesis Test (Z-Statistic)	Notes
One Mean (μ)	$\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	$Z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	μ_0 can be 0 or nonzero
Two Means ($\mu_1 - \mu_2$)	$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$Z = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Δ_0 can be 0 or nonzero
Paired Samples (μ_d)	$\bar{d} \pm z_{1-\frac{\alpha}{2}} \frac{s_d}{\sqrt{n}}$	$Z = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}}$	Analyze differences $d_i = x_{i,1} - x_{i,2}$.

Important:

- Subtract μ_0, Δ_0, or $\mu_{d,0}$ when performing a hypothesis test.
- $\mu_d = 0$ equals no difference for the average individual difference.
- For confidence intervals, center at the sample statistic.
- $z_{1-\frac{\alpha}{2}}$ depends on the confidence level (e.g., 1.96 for 95%).

Introduction to Statistical Methods in Political Science

Lecture 12: Small-Sample Inference for Means: Student-t Toolbox

Ignacio Urbina

Ph.D. Candidate

Why Do We Need New Tools?

Motivating example: Tiny Exit Poll

A survey team intercepts **12** early voters leaving a rural precinct. They record time-in-booth (minutes) to study wait-time equity. Longer time-in-booth may signal inefficiencies, understaffing, or barriers to quick voting (e.g., confusing ballots, slow machines).

- Population SD σ is *unknown*.
- Sample histogram shows minor right-skew and an outlier at 14 min.
- Question: Can we still make a reasonably justified inference about the true mean wait time?

Z procedures from last week assume:

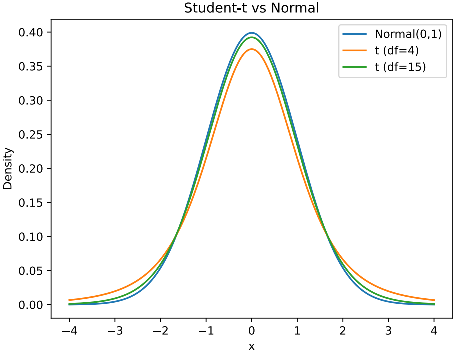
$$\text{Either } n \geq 30 \quad \textbf{or} \quad \sigma \text{ is known.}$$

Neither is true here $\implies$ enter Student-t.

What actually changes when n is small?

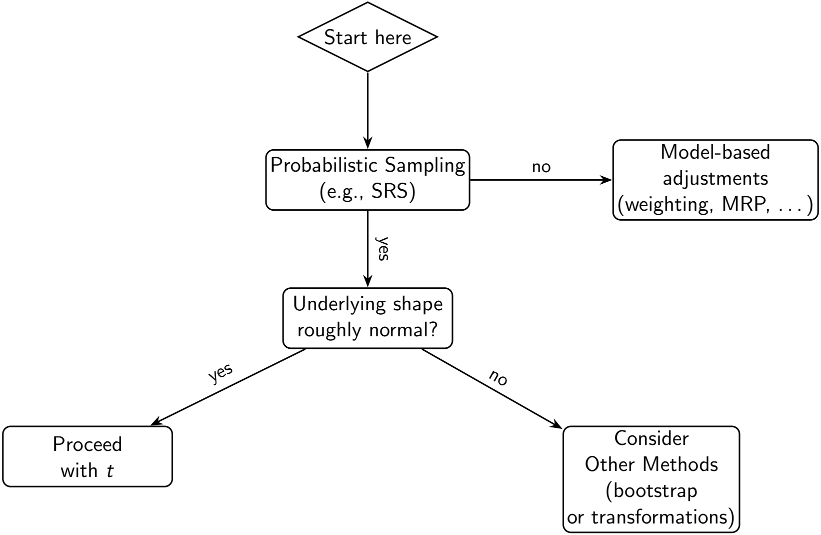
- We swap σ for the noisier estimate s.
- That extra "plug-in" noise fattens the tails of our test statistic.
- We can't rely on the "plug-in principle" (Law of Large Numbers doesn't hold).
- Student-t distribution captures this inflation with **degrees of freedom**:

$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{df=n-1}$$



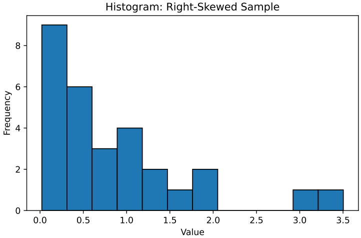
(visual: t_4, t_{15}, and $N(0,1)$)

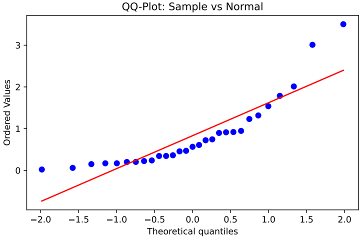
Checklist before using a small-sample *t* method


Start here
Probabilistic Sampling (e.g., SRS)
no
Model-based adjustments (weighting, MRP, . . .)
yes
Underlying shape roughly normal?
yes
no
Proceed with *t*
Consider Other Methods (bootstrap or transformations)

Data-shape diagnostics come first

- With $n < 30$ a single high outlier can wreck validity.
- Always inspect a **histogram** and **QQ-plot**.
- Rule of thumb: mild skew is tolerable for $n \geq 15$; heavy skew/outliers call for non-parametrics or resampling.



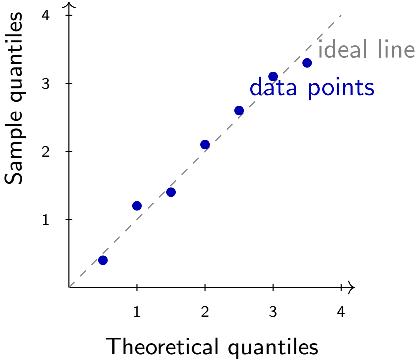


Intuition Behind a Q–Q Plot

What is a Q–Q plot?

- Compares your data's quantiles (y-axis) to theoretical quantiles (x-axis).
- If data follow the chosen distribution, points lie roughly on the 45° line.
- Deviations highlight skew, heavy tails, or outliers.
- Think of "lining up" your sample against the ideal.

QQ-Plot



One-Sample CI for a Mean

Sampling Distribution of Standardized $\bar{x}$ (Small Sample)

Goal: Understand the behavior of $\bar{x}$ as an estimator of μ when sample size is small ($n < 30$).

Common assumptions:

- Data are collected via SRS.
- The population distribution is approximately Normal (key for small n inference).
- Population SD (σ) is unknown.

Result:

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{df=n-1}$$

- Use s to estimate σ → introduces extra variability.
- The t distribution accounts for this with heavier tails than Normal.

Confidence Interval for μ in Small Samples – The recipe

$$\boxed{\bar{x} \ \pm \ t^*_{df=n-1,\,1-\alpha/2} \times \frac{s}{\sqrt{n}}}$$

- **Degrees of freedom** $df = n - 1$.
- **Critical value** t^* comes from a table or software. It depends on both df and α.
- **Interpretation** follows the familiar “We are 95% confident . . . ”.

Question – How does t^* compare with z^*?

Suppose $\alpha = 0.05$.

A. t^*_{19} is **smaller** than $z^* = 1.96$

B. t^*_{19} is **equal to** z^*

C. t^*_{19} is **larger** than z^*

D. It cannot be determined because t^* depends on the sample's standard deviation, whereas z^* does not

(Answer: C; heavier tails)

Worked example: Local campaign donors

Goal: assess mean contributions, based on our sample of 18 local donors, by constructing a 90% confidence interval for the mean donation.

- $n = 18$, $\bar{x} = \$42.8$, $s = \$9.2$.
- 90% confidence wanted ($\alpha = 0.10$).

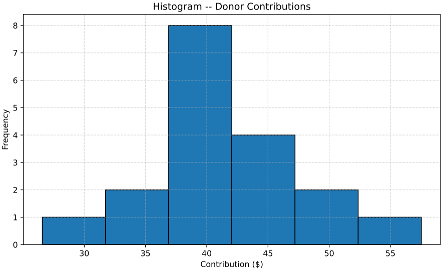
Solution:

$$t^*_{n-1,\,1-\frac{\alpha}{2}} = t^*_{17,\,0.95} = 1.740$$

$$SE = \frac{9.2}{\sqrt{18}} = 2.17$$

$$ME = 1.740 \times 2.17 = 3.77$$

CI:
$\$42.8 \pm 3.8 = (\$39.0,\ \$46.6)$.



Hypothesis Testing for Means

Five–Step Road Map

1. **State H_0 and H_a** (parameter language, direction – one or two tailed)
2. **Choose** α (tolerable Type I risk)
3. **Compute test statistic** $T = \dfrac{\text{estimate} - \text{null}}{SE}$

   $T \sim t_{df}$ if checklist passes.
4. **Decision rule** — compare either $|T|$ to a critical value t^*_{df} *or* use a p-value.
5. **Conclusion in context**. (plain-English, mention evidence strength)

Worked Example — Average Time in Booth

Rural exit-poll revisit.

In response to concerns about unequal voting experiences, electoral officials assert that average time spent in the voting booth should not exceed **6 minutes**. They claim that rural polling stations are operating efficiently and equitably.

Activists, skeptical of this claim, conduct an informal audit by collecting a small **simple random sample** of $n = 12$ early voters at a rural precinct. Each voter is asked how long they spent in the booth, from entry to casting their ballot.

- Sample mean: $\bar{x} = 7.8$ minutes
- Sample standard deviation: $s = 3.1$ minutes
- Officials' claim: $\mu = 6$ minutes (true average time)

Goal: Test whether the true mean booth time μ differs from 6 minutes. Use a two-sided t test at significance level $\alpha = 0.05$.

Worked Example – 5-Step Procedure

Step 1: State hypotheses

$H_0 : \mu = 6$ minutes (official claim)
$H_a : \mu \neq 6$ minutes (activists suspect difference)

Step 2: Set significance level

$\alpha = 0.05$ (two-tailed test). Hence, Critical value: $t^*_{0.975,\,11} = 2.201$.

Step 3: Compute test statistic

Sample mean: $\bar{x} = 7.8$, sample SD: $s = 3.1$, $n = 12$
Standard error: $SE = \frac{3.1}{\sqrt{12}} = 0.90$
Test statistic: $T = \frac{7.8-6}{0.90} = 2.00$

Step 4: Make decision

Degrees of freedom: $df = 12 - 1 = 11$. Critical value: $t^*_{0.975,\,11} = 2.201$
Since $|T| = 2.00 < 2.201$, we **fail to reject** H_0

Step 5: Conclusion in context

Evidence is insufficient (at the 5% level) to conclude that the true average booth time differs from the official 6-minute claim.

Inference for Means in Paired-Samples

Why Use Paired Measurements?

Context: In many research settings, it's hard to detect a treatment effect when individual baseline differences are large.

Solution: Pairing allows each subject (or unit) to serve as their own control.

Common examples of pairing:

- **Before vs. after** a treatment or policy change (e.g., turnout before/after voter ID law)
- **Twin studies** in medical or behavioral research (genetically matched units)
- **Matched groups or regions —** e.g., similar counties, classrooms, or districts

Why Use Paired Measurements?

Why it works:

- Controls for *individual-level variability* (age, baseline attitudes, income, etc.)
- Focuses analysis on the **within-pair difference** d_i
- Turns the problem into a simpler one-sample inference on $\mu_d =$ mean change
- Usually improves precision and statistical power

Sampling Distribution of the Mean Difference

Data: For each of n units we observe a before/after (or matched) pair and compute the difference $d_i = x_{\text{after},i} - x_{\text{before},i}$.

Assumptions:

- Differences d_i are independent draws.
- Distribution of d_i is approximately Normal (key when $n < 30$).

Estimator	**Sampling Distribution**	**Sample SD**
$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$	$\frac{\bar{d} - d}{s_d/\sqrt{n}} \sim t_{df=n-1}$	$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$

- s_d estimates the unknown σ_d, inflating tail thickness.
- Degrees of freedom $df = n - 1$ adjust for that extra noise.

Sampling Distribution of the Mean Difference

Data: For each of n units we observe a before/after (or matched) pair and compute the difference $d_i = x_{\text{after},i} - x_{\text{before},i}$.

Assumptions:

- Differences d_i are independent draws.
- Distribution of d_i is approximately Normal (key when $n < 30$).

Estimator	**Sampling Distribution**	**Sample SD**
$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$	$\frac{\bar{d}-d}{s_d/\sqrt{n}} \sim t_{df=n-1}$	$s_d = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2}$

- s_d estimates the unknown σ_d, inflating tail thickness.
- Degrees of freedom $df = n - 1$ adjust for that extra noise.

CI & Test for Mean Difference μ_d

$$\text{Confidence Interval:} \quad \bar{d} \pm t^*_{df=n-1,\,1-\alpha/2} \frac{s_d}{\sqrt{n}}$$

$$\text{Test statistic:} \quad T = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}} \sim t_{df=n-1}$$

- $\mu_{d,0}$ is the *null hypothesized* mean difference (often 0 for "no change").

Key assumptions

- Differences d_i are *independent*.
- Sample size for *df* is the count of *pairs*, not raw observations.
- The distribution of d_i is *approximately Normal* (check histogram/QQ-plot).
- For CI: same as test, plus choice of confidence level $1 - \alpha$.

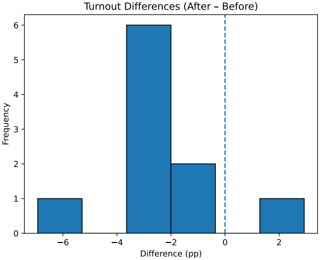
Example: Turnout before vs after voter-ID law

Ten matched counties → $n = 10$. $\bar{d} = -1.8$ pp, $s_d = 2.7$ pp; **Task:** Compute 95 % confidence interval.

$$t^*_{9,\,0.975} = 2.262,$$
$$SE = \frac{2.7}{\sqrt{10}} \approx 0.85,$$
$$ME = 2.262 \times 0.85 \approx 1.9$$

CI: $-1.8 \pm 1.9 = (-3.7,\ 0.1)$ pp

Take-away: 95% CI = (-3.7 pp, +0.1 pp): a zero increase can't be excluded.



Example – Interpreting the 95% CI

- CI for mean change: (-3.7 pp, 0.1 pp) (pp.=percentage points).
- All values in this range are equally compatible with the data at the 5% significance level
- **You cannot** assign greater “plausibility” to negative vs. positive values
- Conclusion: Data support a decrease, no change, or a small increase — nothing beyond this interval is consistent at 95%

Practice Problem 2 – Paired Data, Small Sample

$$T = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}} \sim t_{n-1}$$

Context

- $n = 9$ individuals measured **before** and **after** a training program.
- Goal: Test if mean improvement μ_d differs from 0.
- Use a **paired** t test — small sample, assume differences $\approx$ normal.

Quick practice (think–pair–share):
$\bar{d} = 4.1,\ s_d = 5.4,\ \alpha = 0.10$ (two-sided).
Reject H_0? $\rightarrow T = 2.27 > t^*_8 = 1.86 \rightarrow$ **Yes**.

Addendum – Inference for Paired Differences (Large n)

Scenario: You observe two measurements on each unit (e.g., before vs after treatment) and compute the difference $d_i = x_{\text{after},i} - x_{\text{before},i}$.

When n is large, we invoke the Central Limit Theorem:

$$\bar{d} \sim N\left(\mu_d,\ \frac{\sigma_d^2}{n}\right) \quad \text{(approximately, by CLT).}$$

Use Z procedures if:

- $n \geq 30$ (number of *pairs*),
- Independence: pairs are randomly sampled or randomly assigned,
- You estimate σ_d with sample SD (s_d).

$$\textbf{CI for } \mu_d : \bar{d} \pm z^*_{1-\alpha/2} \cdot \frac{s_d}{\sqrt{n}} \qquad \text{Test stat: } Z = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}}$$

Comparing Two Small Samples

Comparing Means Using Small Samples

Core Question: Do two distinct groups differ *on average* in some key outcome?

- Do 12 rural precincts using new machines have shorter average wait times than 12 using the old ones?
- Do honors students in a pilot class (n = 14) outperform regular students (n = 15) on a civics quiz?
- Are turnout rates different across two small counties in a special election (n = 10 precincts each)?

Comparing Means Using Small Samples

What makes this different from one-sample inference?

- Two samples = two sources of variability
- Independence between groups is critical
- Assumptions about spread (equal vs unequal variance) influence the method

Our goal: Infer whether population means μ_1 and μ_2 are different, using small samples.

Why Variance Matters More with Small Samples

In large samples, we relied on the Law of Large Numbers:

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad \text{(plug-in with } s_1, s_2\text{)}$$

But with small samples:

- Estimates s_1 and s_2 are noisy — not reliable stand-ins for σ_1, σ_2
- This extra uncertainty fattens the tails of our test statistic
- We need to adjust using a **t distribution** — with carefully chosen degrees of freedom

When & How to Pool Variances

Equal-variance assumption $\sigma_1^2 = \sigma_2^2 = \sigma^2$ in the populations. A quick screen: variance (or SD) ratio < 2 *and* similar histograms.

Pooled estimate of the common SD

$$s_{pooled}^2 = s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}, \qquad s_p = \sqrt{s_p^2}.$$

$$SE_{\text{pooled}} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \qquad df = n_1 + n_2 - 2.$$

Use pooled t only when:

- Boxplots / histograms show comparable spread;
- Sample sizes are not wildly unequal;
- A formal test (e.g. Levene) does *not* reject equal variances.

Otherwise, default to Welch's unpooled procedure.

Welch **t** statistic (safer default)

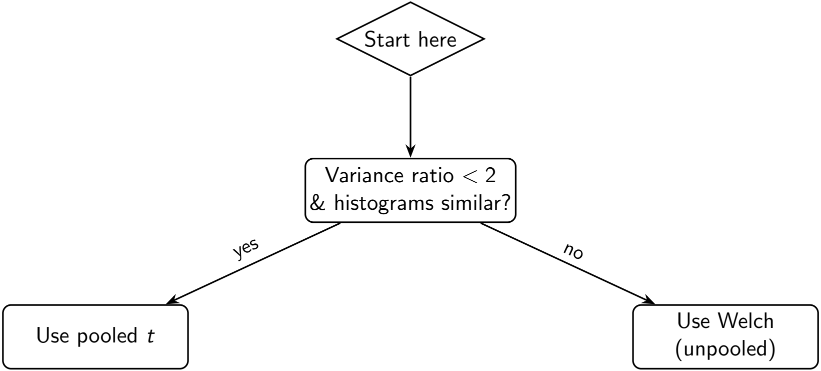
$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \quad \sim \quad t_{df_{\text{Welch}}}$$

Welch degrees of freedom (software reports this)

$$df = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}.$$

Use pooled-SD version only when diagnostics support equal population variances.

Pooled vs. Welch decision tree



Example — Civics Quiz Scores: Honors vs Regular

- **Research context:** Instructor wants to know if an enriched honors curriculum leads to higher civics-quiz performance than the standard curriculum
- **Data:** 20-question multiple-choice quiz (0–100 scale), administered simultaneously to two sections in Spring term
- **Goal:** Estimate and test the difference in true mean scores between honors vs. regular students
- Honors class: $n_1 = 12,\ \bar{x}_1 = 77.3,\ s_1 = 8.4$
- Regular class: $n_2 = 15,\ \bar{x}_2 = 70.1,\ s_2 = 7.1$
- Hypothesis test: $H_0 : \mu_1 - \mu_2 = 0$ vs. two-sided H_a at $\alpha = 0.05$

Calculations

$$SE = \sqrt{\frac{8.4^2}{12} + \frac{7.1^2}{15}} = 3.29, \quad T = \frac{77.3 - 70.1}{3.29} = 2.19$$

$$df_{\text{Welch}} = \frac{(8.4^2/12 + 7.1^2/15)^2}{\frac{8.4^4}{12^2 \cdot 11} + \frac{7.1^4}{15^2 \cdot 14}} \approx 20.7$$

Two-tailed critical value: $t^*_{0.975,\,20} = 2.086$. Since $2.19 > 2.086$ we **reject** H_0.

Conclusion: Honors students score significantly higher ($\approx$ 7 pts).

Common traps with two-sample t

- **Heteroskedasticity**: ignoring unequal variances shrinks SE.
- **Imbalanced** n: smaller group sample size drives df; watch power.
- **Multiple testing**: comparing many sub-groups inflates Type I error (adjust α or use FDR).

Key Formulas and Takeaways

Cheat-Sheet – Small Sample Inference

Scenario	**Confidence Interval**	**Test Statistic** (Δ_0 or μ_0 in numerator)
One mean μ	$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$	$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \quad (df = n - 1)$
Paired mean μ_d	$\bar{d} \pm t^*_{n-1} \frac{s_d}{\sqrt{n}}$	$T = \frac{\bar{d} - \mu_{d,0}}{s_d/\sqrt{n}} \quad (df = n - 1)$
Two means $\mu_1 - \mu_2$ (Welch)	$(\bar{x}_1 - \bar{x}_2) \pm t^*_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$	$T = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad (df = \text{Welch})$
Two means $\mu_1 - \mu_2$ (Pooled)	$(\bar{x}_1 - \bar{x}_2) \pm t^*_{n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$	$T = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \quad (df = n_1 + n_2 - 2)$

- Always check independence *and* approximate Normal shape.
- Use Welch unless equal-variance assumption is defensible.
- Report *df* and *p*-value to two decimals.