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 $\rightarrow$ 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.

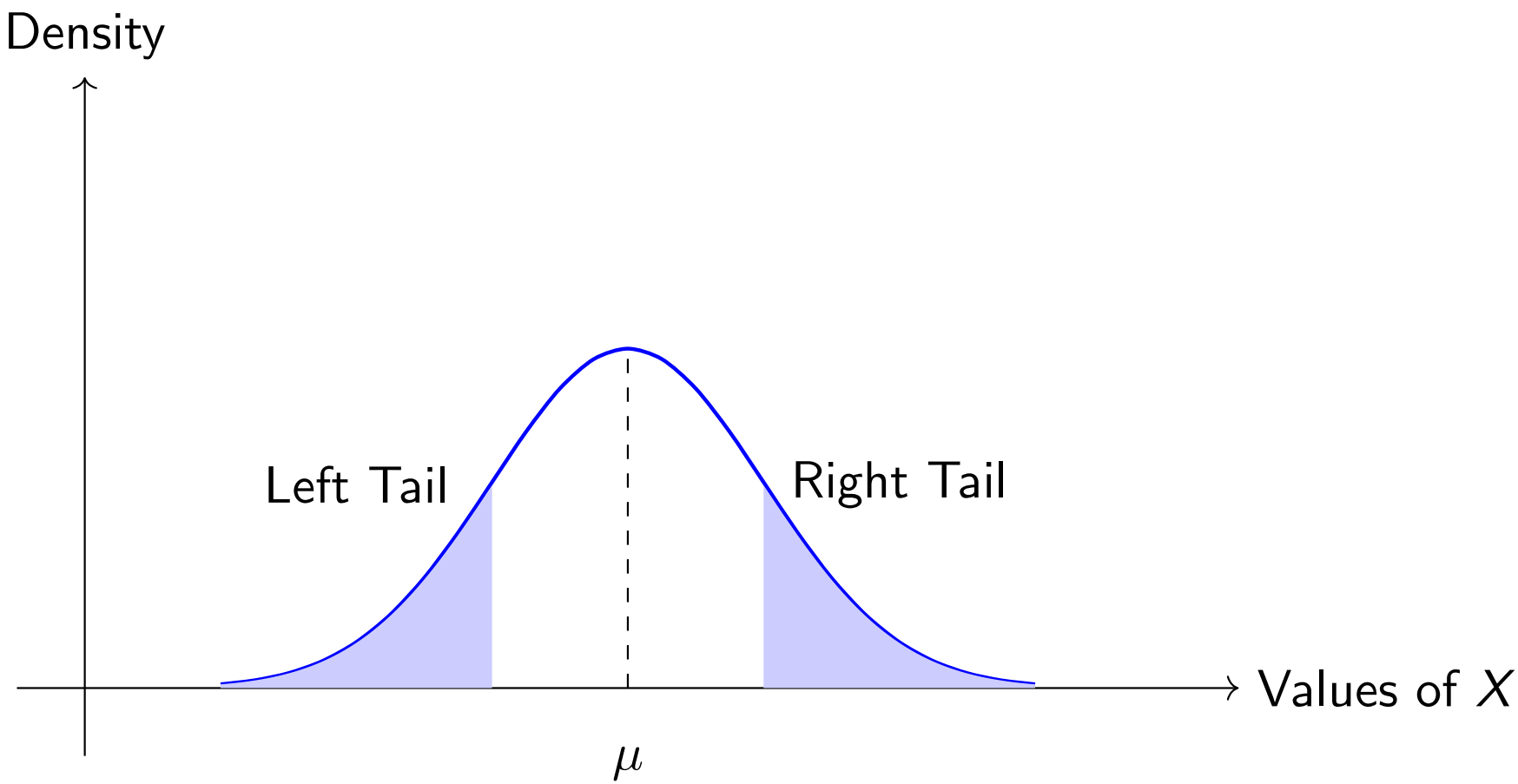
Descriptive Statistics

- 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

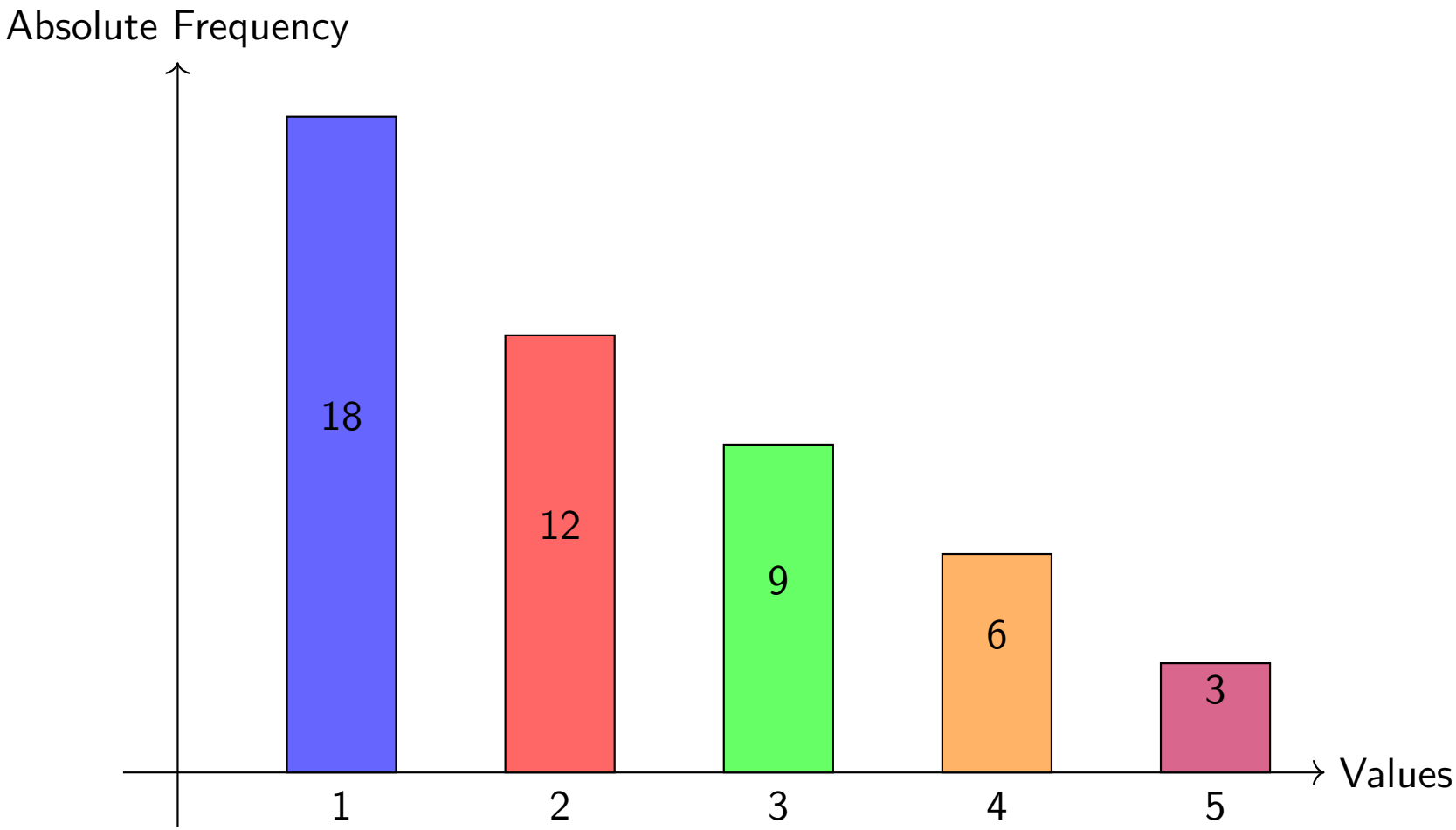


Figure: Categorical Ordinal Distribution

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:

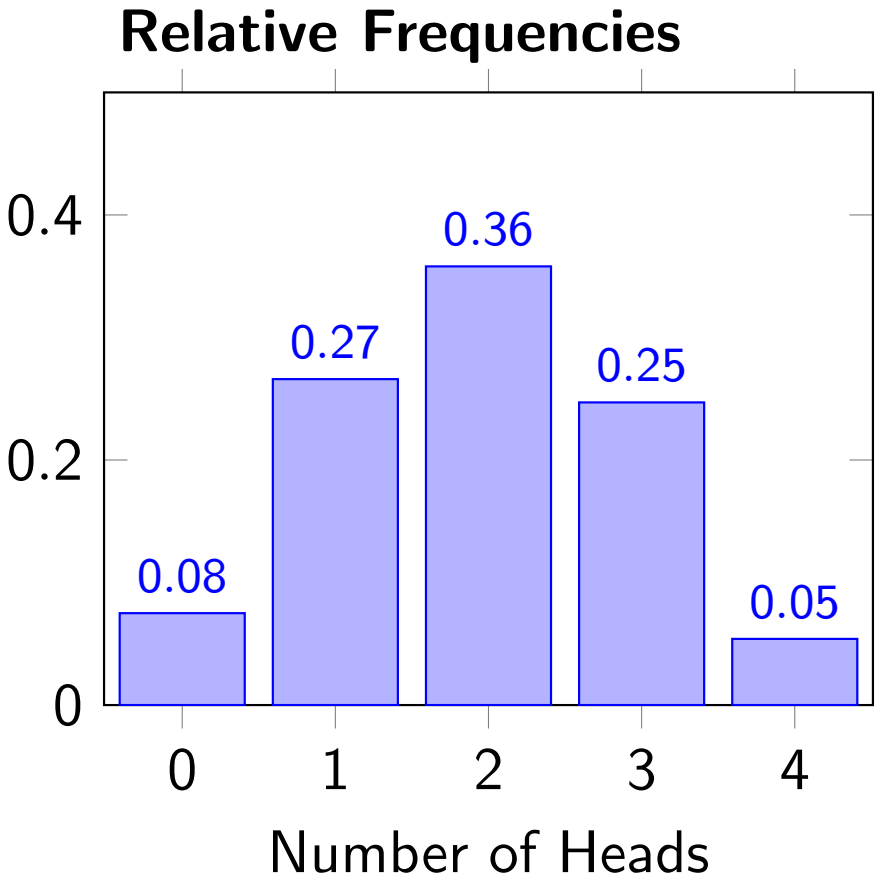
- Each throw consists of flipping 4 independent coins.
- We record the number of heads obtained in each throw.
- The results are summarized in terms of:
  - Absolute frequencies.
  - Cumulative absolute frequencies.
  - Relative frequencies.
  - Cumulative relative frequencies.

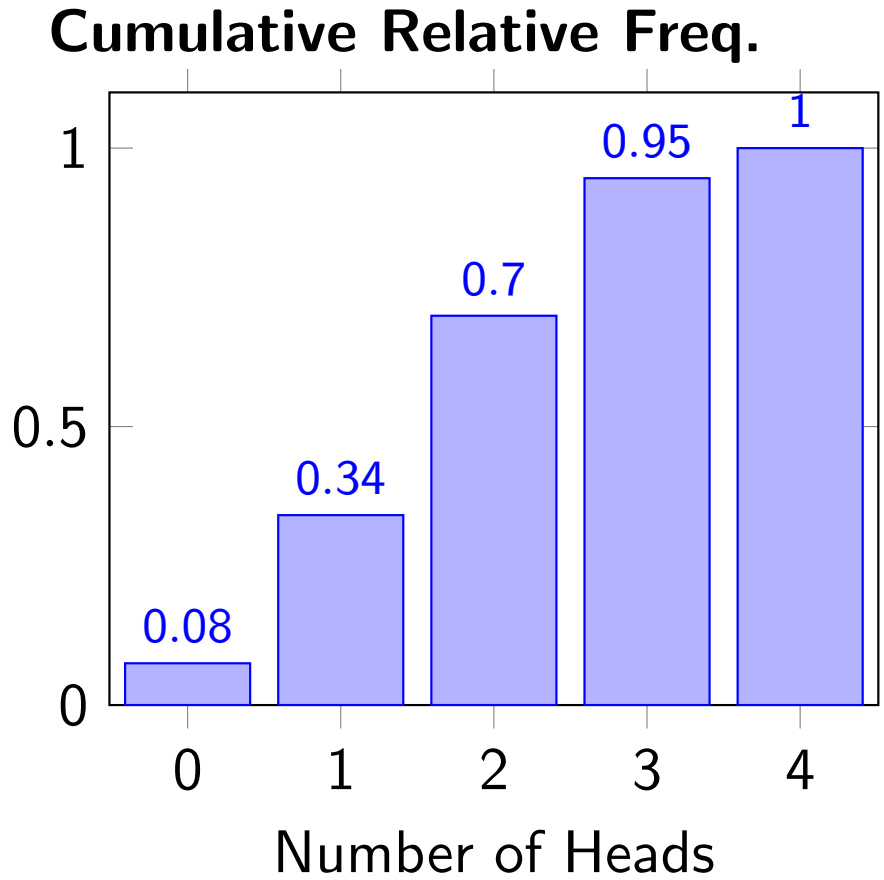
Simulation of 1,000 Throws of Four Coins

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

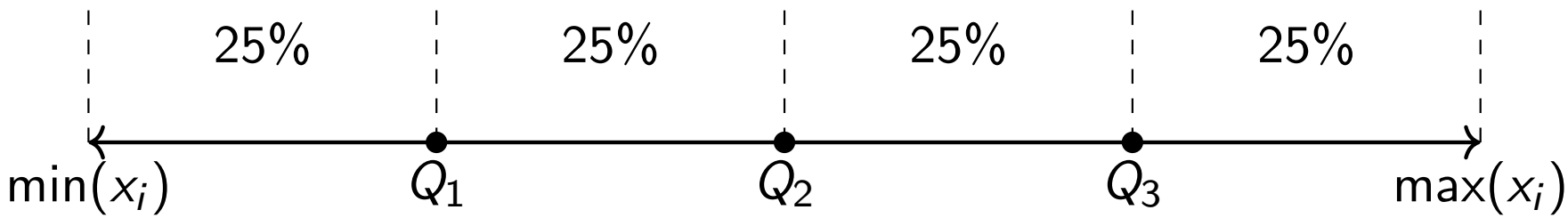
Measure of Position

- **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

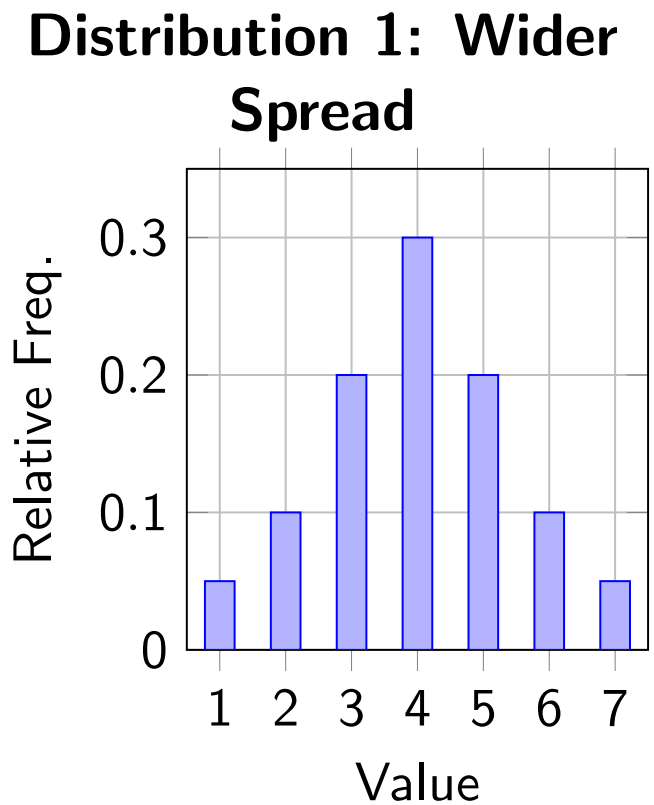
- **Deciles:** Divide the data into ten equal parts.
  - **Ninth Decile Example:**
    - 9th Decile = \$150,000: 90% of individuals earn less than or equal to \$150,000.
- **Percentiles:** Divide the data into one hundred equal parts.
  - **Ninety-Fifth Percentile Example:**
    - 95th Percentile = \$400,000: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

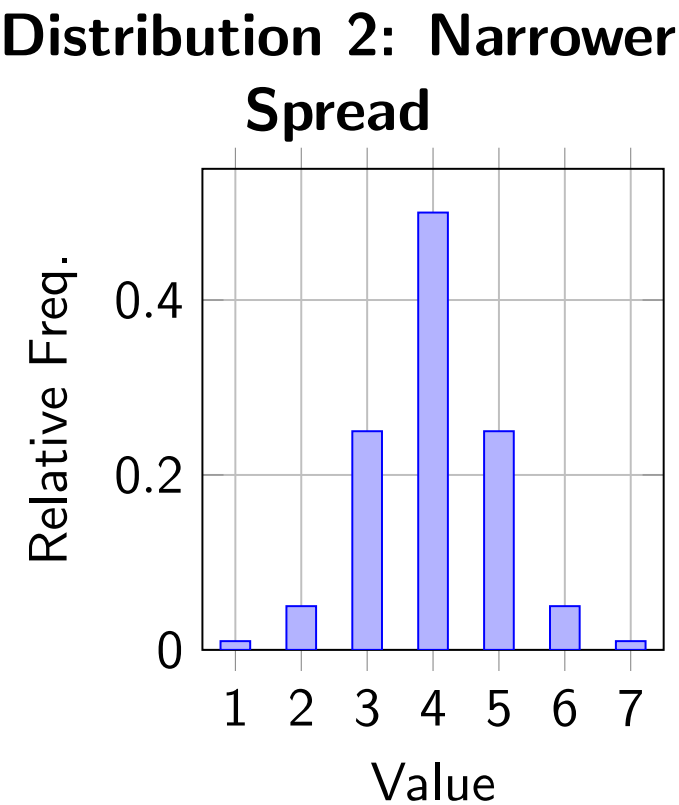
Measures of Dispersion

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





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.