Measures of Central Tendency

Measures of central tendency describe the center or typical value of a dataset, summarizing the data with a single representative value. We will explore the **Mean (Arithmetic Average)**, **Median**, and **Mode**'

1 Mean (Arithmetic Average)

The mean is a measure of central tendency that represents the average of a dataset. It is calculated by summing all the data values and dividing by the number of values. The mean is sensitive to outliers but provides a comprehensive summary of the data by considering all values.

Formula for the Mean

- **Ungrouped Data**:

$$\text{Mean}\,(\mu \text{ or } \bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

  where x_i are the individual data values, and n is the number of data points.

- **Grouped Data**:

$$\text{Mean}\,(\bar{x}) = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

  where:

  - x_i is the midpoint of the i-th class interval,
  - f_i is the frequency of the i-th class interval,
  - k is the number of class intervals,
  - $\sum f_i$ is the total frequency (total number of observations).

1.1 Mean for Ungrouped Data

Ungrouped data consists of individual data points that are not organized into intervals or classes. Each value is listed separately, and we calculate the mean directly from these values.

Formula (Repeated for Clarity):

$$\text{Mean}\,(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example 1: Simple Ungrouped Data

Dataset: Test scores of 5 students: 85, 90, 75, 95, 80

Step 1: List the data values.

- x_i: 85, 90, 75, 95, 80
- Number of values $(n) = 5$

Step 2: Sum the data values $(\sum x_i)$.

$$\sum x_i = 85 + 90 + 75 + 95 + 80 = 425$$

Step 3: Divide the sum by the number of values.

$$\text{Mean}\,(\bar{x}) = \frac{\sum x_i}{n} = \frac{425}{5} = 85$$

Result: The mean test score is 85.

Example 2: Ungrouped Data with Repeated Values

Dataset: Number of books read by 8 students in a month: 2, 3, 2, 5, 4, 3, 2, 6

Step 1: List the data values.

- x_i: 2, 3, 2, 5, 4, 3, 2, 6
- Number of values $(n) = 8$

Step 2: Sum the data values $(\sum x_i)$.

$$\sum x_i = 2 + 3 + 2 + 5 + 4 + 3 + 2 + 6 = 27$$

Step 3: Divide the sum by the number of values.

$$\text{Mean}\,(\bar{x}) = \frac{\sum x_i}{n} = \frac{27}{8} = 3.375$$

Result: The mean number of books read is 3.375 (or approximately 3.38).

1.2 Mean for Grouped Data

Grouped data is organized into class intervals, often with corresponding frequencies. This is common when dealing with large datasets or continuous data (e.g., heights, weights, or test scores). To calculate the mean, we use the midpoints of the class intervals and their frequencies.

Formula (Repeated for Clarity):

$$\text{Mean}\,(\bar{x}) = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

Steps for Calculation:

1. Identify the class intervals and their frequencies (f_i).
2. Calculate the midpoint (x_i) of each class interval.

$$\text{Midpoint} = \frac{\text{Lower Limit} + \text{Upper Limit}}{2}$$

3. Multiply each midpoint by its frequency $(f_i x_i)$.
4. Sum the products $(\sum f_i x_i)$.
5. Sum the frequencies $(\sum f_i)$.
6. Divide the sum of the products by the total frequency.

Example 1: Grouped Data (Test Scores)

Dataset: The following table shows the test scores of 50 students, grouped into intervals.

Class Interval	**Frequency** (f_i)
50–60	5
60–70	10
70–80	15
80–90	12
90–100	8

Step 1: Calculate the midpoint (x_i) for each class interval.

- 50–60: $x_1 = \frac{50+60}{2} = 55$
- 60–70: $x_2 = \frac{60+70}{2} = 65$
- 70–80: $x_3 = \frac{70+80}{2} = 75$
- 80–90: $x_4 = \frac{80+90}{2} = 85$
- 90–100: $x_5 = \frac{90+100}{2} = 95$

Step 2: Multiply each midpoint by its frequency ($f_i x_i$) and sum the frequencies.

Class Interval	**Frequency** (f_i)	**Midpoint** (x_i)	$f_i x_i$
50–60	5	55	$5 \times 55 = 275$
60–70	10	65	$10 \times 65 = 650$
70–80	15	75	$15 \times 75 = 1125$
80–90	12	85	$12 \times 85 = 1020$
90–100	8	95	$8 \times 95 = 760$

- Total frequency ($\sum f_i$) $= 5 + 10 + 15 + 12 + 8 = 50$
- Sum of products ($\sum f_i x_i$) $= 275 + 650 + 1125 + 1020 + 760 = 3830$

Step 3: Calculate the mean.

$$\text{Mean}\,(\bar{x}) = \frac{\sum f_i x_i}{\sum f_i} = \frac{3830}{50} = 76.6$$

Result: The mean test score is 76.6.

Example 2: Grouped Data (Heights of Plants)

Dataset: The heights (in cm) of 40 plants are grouped as follows.

Height (cm)	**Frequency** (f_i)
10–20	6
20–30	12
30–40	10
40–50	8
50–60	4

Step 1: Calculate the midpoint (x_i) for each class interval.

- 10–20: $x_1 = \frac{10+20}{2} = 15$
- 20–30: $x_2 = \frac{20+30}{2} = 25$
- 30–40: $x_3 = \frac{30+40}{2} = 35$
- 40–50: $x_4 = \frac{40+50}{2} = 45$
- 50–60: $x_5 = \frac{50+60}{2} = 55$

Step 2: Multiply each midpoint by its frequency ($f_i x_i$) and sum the frequencies.

Height (cm)	**Frequency** (f_i)	**Midpoint** (x_i)	$f_i x_i$
10–20	6	15	$6 \times 15 = 90$
20–30	12	25	$12 \times 25 = 300$
30–40	10	35	$10 \times 35 = 350$
40–50	8	45	$8 \times 45 = 360$
50–60	4	55	$4 \times 55 = 220$

- Total frequency $(\sum f_i) = 6 + 12 + 10 + 8 + 4 = 40$
- Sum of products $(\sum f_i x_i) = 90 + 300 + 350 + 360 + 220 = 1320$

Step 3: Calculate the mean.

$$\text{Mean}\,(\bar{x}) = \frac{\sum f_i x_i}{\sum f_i} = \frac{1320}{40} = 33$$

Result: The mean height of the plants is 33 cm.

Summary of Mean Calculations

- **Ungrouped Data**:

$$\text{Mean}\,(\bar{x}) = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example: Scores: 85, 90, 75, 95, 80

$$\text{Mean} = \frac{85 + 90 + 75 + 95 + 80}{5} = \frac{425}{5} = 85$$

- **Grouped Data**:

$$\text{Mean}\,(\bar{x}) = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i}$$

Example: Test scores of 50 students:

Class Interval	**Frequency** (f_i)	**Midpoint** (x_i)	$f_i x_i$
50–60	5	55	275
60–70	10	65	650
70–80	15	75	1125
80–90	12	85	1020
90–100	8	95	760

$$\text{Mean} = \frac{275 + 650 + 1125 + 1020 + 760}{5 + 10 + 15 + 12 + 8} = \frac{3830}{50} = 76.6$$

Strengths of the Mean

1. **Uses All Data Values**:
   - The mean incorporates every value in the dataset, making it a comprehensive summary of the data.
   - **Example (Ungrouped)**: In the scores 85, 90, 75, 95, 80, all values contribute to the mean (85), reflecting the overall performance.

- **Example (Grouped)**: For the test scores, all class intervals and frequencies contribute to the mean (76.6), capturing the distribution across the entire range.

2. **Mathematical Simplicity and Utility**:
   - The mean is easy to calculate and serves as a foundation for other statistical measures (e.g., variance, standard deviation).
   - It's widely used in statistical analysis because of its algebraic properties (e.g., the sum of deviations from the mean is zero).
   - **Example**: The mean test score of 76.6 can be used to calculate the variance by finding deviations from this central value.

3. **Good for Symmetric Data**:
   - The mean is an excellent measure of central tendency when the data is symmetric (e.g., follows a normal distribution) and has no extreme outliers.
   - **Example**: If test scores are symmetrically distributed around 76.6 with no extreme outliers, the mean accurately represents the typical score.

4. **Applicable to Both Ungrouped and Grouped Data**:
   - The mean can be calculated for both raw (ungrouped) and summarized (grouped) data, making it versatile.
   - **Example**: We calculated the mean for ungrouped data (85) and grouped data (76.6), showing its adaptability.

Weaknesses of the Mean

1. **Sensitive to Outliers**:
   - The mean is heavily influenced by extreme values, which can distort its representation of the "typical" value.
   - **Example (Ungrouped)**: Add an outlier to the scores: 85, 90, 75, 95, 80, **150**.

$$\text{Mean} = \frac{85 + 90 + 75 + 95 + 80 + 150}{6} = \frac{575}{6} \approx 95.83$$

     The mean jumps from 85 to 95.83 due to the outlier (150), which may not reflect the typical student's performance.
   - **Example (Grouped)**: If the grouped data had an additional interval like 150–160 with a frequency of 1, the mean would increase significantly, skewing the result.

2. **Not Suitable for Skewed Data**:
   - In skewed distributions (e.g., income data with a few very high values), the mean may not represent the central tendency well.
   - **Example**: Consider incomes: \$30,000, \$40,000, \$50,000, \$1,000,000 (ungrouped).

$$\text{Mean} = \frac{30,000 + 40,000 + 50,000 + 1,000,000}{4} = \frac{1,120,000}{4} = 280,000$$

     The mean (\$280,000) is far higher than most values due to the skewed distribution, making it misleading.

3. **Requires Numerical Data**:
    - The mean cannot be calculated for categorical data (e.g., colors, names), limiting its applicability.
    - **Example**: You cannot calculate the mean of favorite colors: "red, blue, red, green." The mean is only meaningful for numerical data like test scores or heights.
4. **Approximation Error in Grouped Data**:
    - For grouped data, the mean relies on midpoints, assuming data is evenly distributed within each interval, which may not be true.
    - **Example (Grouped)**: In the test score example, the mean of 76.6 assumes scores in the 70–80 interval are centered at 75. If most scores in that interval are closer to 70, the mean may overestimate the average.

1.3 Weighted Mean

The **weighted mean** accounts for the importance (weight) of each value in a dataset. It's useful when some values contribute more to the average than others.

Formula:

$$\text{Weighted Mean } (\bar{x}_w) = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where x_i are the data values, w_i are the weights, and n is the number of values.

Example: Scores: 80, 90, 100 with weights 1, 2, 3.

$$\bar{x}_w = \frac{(1 \times 80) + (2 \times 90) + (3 \times 100)}{1 + 2 + 3} = \frac{80 + 180 + 300}{6} = \frac{560}{6} \approx 93.33$$

The weighted mean is approximately 93.33.

1.4 Trimmed Mean

The **trimmed mean** reduces the impact of outliers by removing a percentage of the smallest and largest values before calculating the mean.

Steps:

1. Sort the data in ascending order.
2. Trim the specified percentage (e.g., 10%) from both ends.
3. Calculate the mean of the remaining values.

Example: Dataset: 10, 20, 30, 90, 100 (5 values). Trim 20% (1 value from each end).

- Sorted: 10, 20, 30, 90, 100
- Trimmed: 20, 30, 90
- Mean: $\frac{20+30+90}{3} = \frac{140}{3} \approx 46.67$

The trimmed mean is approximately 46.67.

2 Median

The **median** is the middle value of a dataset when the data points are arranged in ascending order. It divides the dataset into two equal halves, with 50% of the values below it and 50% above it. The median is particularly useful because it is not affected by extreme values, making it a robust measure of central tendency.

Formula for the Median

- **Ungrouped Data**:
  - If n (number of observations) is **odd**: The median is the value at position $\frac{n+1}{2}$.
  - If n is **even**: The median is the average of the values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$.
- **Grouped Data**: The median for grouped data is found using the cumulative frequency and the median class (the class where the cumulative frequency reaches or exceeds $\frac{n}{2}$). The formula is:

$$\text{Median} = L + \left(\frac{\frac{n}{2} - CF}{f} \right) \times h$$

  where:
  - L = lower boundary of the median class,
  - n = total number of observations ($\sum f_i$),
  - CF = cumulative frequency of the class before the median class,
  - f = frequency of the median class,
  - h = width of the median class interval.

2.1 Median for Ungrouped Data

Ungrouped data consists of individual data points listed separately. To find the median, we first arrange the data in ascending order and then apply the appropriate formula based on whether the number of observations is odd or even.

Example 1: Ungrouped Data (Odd Number of Observations)

Dataset: Test scores of 7 students: 85, 90, 75, 95, 80, 88, 92

Step 1: Arrange the data in ascending order.

- Ordered dataset: 75, 80, 85, 88, 90, 92, 95
- Number of observations (n) = 7 (odd)

Step 2: Find the median position.

- Since n is odd, the median is at position $\frac{n+1}{2}$.

$$\frac{7+1}{2} = \frac{8}{2} = 4$$

Step 3: Identify the value at the 4th position.

- Ordered data: 75, 80, 85, **88**, 90, 92, 95
- The 4th value is 88.

Result: The median test score is 88.

Example 2: Ungrouped Data (Even Number of Observations)

Dataset: Number of books read by 6 students in a month: 2, 3, 5, 4, 3, 6

Step 1: Arrange the data in ascending order.

- Ordered dataset: 2, 3, 3, 4, 5, 6
- Number of observations (n) = 6 (even)

Step 2: Find the median positions.

- Since n is even, the median is the average of the values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

$$\frac{n}{2} = \frac{6}{2} = 3 \quad \text{and} \quad \frac{n}{2}+1 = 3+1 = 4$$

Step 3: Identify the values at the 3rd and 4th positions and compute their average.

- Ordered data: 2, 3, **3**, **4**, 5, 6
- 3rd value = 3, 4th value = 4
- Median = $\frac{3+4}{2} = \frac{7}{2} = 3.5$

Result: The median number of books read is 3.5.

2.2 Median for Grouped Data

Grouped data is organized into class intervals with corresponding frequencies, often used for large datasets or continuous data. To find the median, we identify the median class using cumulative frequency and then apply the interpolation formula.

Formula (Repeated for Clarity):

$$\text{Median} = L + \left(\frac{\frac{n}{2} - CF}{f} \right) \times h$$

Example 1: Grouped Data (Test Scores)

Dataset: Test scores of 50 students, grouped into intervals.

Class Interval	**Frequency** (f_i)
50–60	5
60–70	10
70–80	15
80–90	12
90–100	8

Step 1: Calculate the total frequency (n) and find $\frac{n}{2}$.

- Total frequency (n) = $5+10+15+12+8 = 50$
- $\frac{n}{2} = \frac{50}{2} = 25$

Step 2: Compute the cumulative frequency to identify the median class.

Class Interval	Frequency (f_i)	Cumulative Frequency
50–60	5	5
60–70	10	15
70–80	15	30
80–90	12	42
90–100	8	50

- The median class is the interval where the cumulative frequency first reaches or exceeds $\frac{n}{2} = 25$. Here, the cumulative frequency reaches 30 at the 70–80 interval, so the median class is 70–80.

Step 3: Apply the median formula.

- L (lower boundary of the median class) $= 70$
- CF (cumulative frequency before the median class) $= 15$
- f (frequency of the median class) $= 15$
- h (class width) $= 80 - 70 = 10$
- $\frac{n}{2} = 25$

$$\text{Median} = L + \left(\frac{\frac{n}{2} - CF}{f}\right) \times h$$

$$\text{Median} = 70 + \left(\frac{25 - 15}{15}\right) \times 10 = 70 + \left(\frac{10}{15}\right) \times 10 = 70 + \frac{10}{1.5} = 70 + 6.67 = 76.67$$

Result: The median test score is approximately 76.67.

Example 2: Grouped Data (Heights of Plants)

Dataset: Heights (in cm) of 40 plants, grouped as follows.

Height (cm)	Frequency (f_i)
10–20	6
20–30	12
30–40	10
40–50	8
50–60	4

Step 1: Calculate the total frequency (n) and find $\frac{n}{2}$.

- Total frequency (n) $= 6 + 12 + 10 + 8 + 4 = 40$
- $\frac{n}{2} = \frac{40}{2} = 20$

Step 2: Compute the cumulative frequency to identify the median class.

Height (cm)	Frequency (f_i)	Cumulative Frequency
10–20	6	6
20–30	12	18
30–40	10	28
40–50	8	36
50–60	4	40

- The cumulative frequency reaches 28 at the 30–40 interval, which exceeds $\frac{n}{2} = 20$. So, the median class is 30–40.

Step 3: Apply the median formula.

- L (lower boundary of the median class) = 30
- CF (cumulative frequency before the median class) = 18
- f (frequency of the median class) = 10
- h (class width) = $40 - 30 = 10$
- $\frac{n}{2} = 20$

$$\text{Median} = 30 + \left(\frac{20 - 18}{10}\right) \times 10 = 30 + \left(\frac{2}{10}\right) \times 10 = 30 + 0.2 \times 10 = 30 + 2 = 32$$

Result: The median height of the plants is 32 cm.

Strengths of the Median

1. **Robust to Outliers**:
   - The median is not affected by extreme values, making it a better measure of central tendency for skewed datasets.
   - **Example (Ungrouped)**: Add an outlier to the scores: 75, 80, 85, 88, 90, 92, 95, **150**.
     - Ordered data: 75, 80, 85, 88, 90, 92, 95, 150
     - $n = 8$, median = average of 4th and 5th values: $\frac{88+90}{2} = 89$
     - The median (89) is barely affected by the outlier (150), unlike the mean, which would increase significantly.
   - **Example (Grouped)**: If the test scores had an additional class like 150–160 with a small frequency, the median class would likely remain 70–80, keeping the median stable.
2. **Good for Skewed Data**:
   - The median is ideal for datasets with skewed distributions (e.g., income, time to failure), where the mean might be misleading.
   - **Example**: Incomes: \$30,000, \$40,000, \$50,000, \$1,000,000.
     - Ordered: \$30,000, \$40,000, \$50,000, \$1,000,000
     - Median = $\frac{40{,}000+50{,}000}{2} = 45,000$, which better represents the typical income than the mean (\$280,000).
3. **Applicable to Ordinal Data**:
   - The median can be used with ordinal data (e.g., rankings, Likert scales) as long as the data can be ordered.
   - **Example**: Survey responses (1 = Poor, 2 = Fair, 3 = Good, 4 = Excellent): 1, 2, 3, 3, 4
     - Ordered: 1, 2, 3, 3, 4
     - Median = 3 (Good), which is meaningful for ordinal data.
4. **Simple to Understand and Calculate**:

- The concept of the "middle value" is intuitive, and the calculation is straightforward for ungrouped data.
- **Example**: For the books dataset (2, 3, 3, 4, 5, 6), the median (3.5) is easily found by ordering and averaging the middle two values.

Weaknesses of the Median

1. **Ignores the Magnitude of Other Values**:
   - The median only considers the middle value(s) and does not account for the magnitude of other data points, potentially losing information.
   - **Example (Ungrouped)**: In the scores 75, 80, 85, 88, 90, 92, 95, the median is 88, but it doesn't reflect the spread or the actual values of the other scores.
   - **Example (Grouped)**: The median test score (76.67) doesn't indicate how many students scored in the higher intervals like 90–100.
2. **Less Useful for Further Statistical Analysis**:
   - Unlike the mean, the median lacks algebraic properties that make it useful for advanced statistical calculations (e.g., variance, standard deviation).
   - **Example**: You cannot directly use the median to calculate the variance of the test scores dataset; the mean is required for such computations.
3. **Approximation in Grouped Data**:
   - For grouped data, the median relies on interpolation within the median class, assuming a uniform distribution within the interval, which may not be accurate.
   - **Example (Grouped)**: In the heights example, the median (32) assumes a linear distribution within the 30–40 interval, but if most heights are closer to 30, the true median might be lower.
4. **May Not Be Unique in Discrete Data**:
   - In datasets with repeated values or discrete data, the median might not be a unique value, especially in even-sized datasets.
   - **Example**: Dataset: 1, 1, 2, 2
     - Median $= \frac{1+2}{2} = 1.5$, which isn't an actual data value, potentially making it less intuitive.

Practical Notes

- Use the median when dealing with skewed data or datasets with outliers, as it provides a better representation of the "typical" value than the mean in such cases.
- For grouped data, the median is an estimate due to the interpolation assumption, so interpret it with caution if the distribution within intervals is uneven.
- The median is especially useful in fields like economics (e.g., median income) or real estate (e.g., median house price) where extreme values are common.

3 Mode

The **mode** is the value or values that appear most frequently in a dataset. It represents the most common observation and is unique among measures of central tendency because a dataset can have no mode, one mode (unimodal), or multiple modes (bimodal or multimodal). The mode is particularly useful for identifying the most typical value in categorical or discrete data.

Formula for the Mode

- **Ungrouped Data**:
  - There is no explicit formula; the mode is simply the value(s) with the highest frequency.
  - Identify the value that appears most often in the dataset.
- **Grouped Data**: For grouped data, the mode is estimated using the modal class (the class interval with the highest frequency). The formula for the mode is:

$$\text{Mode} = L + \left(\frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \right) \times h$$

  where:
  - L = lower boundary of the modal class,
  - f_m = frequency of the modal class,
  - f_1 = frequency of the class before the modal class,
  - f_2 = frequency of the class after the modal class,
  - h = width of the modal class interval.

3.1 Mode for Ungrouped Data

Ungrouped data consists of individual data points listed separately. To find the mode, we identify the value(s) that occur most frequently.

Example 1: Ungrouped Data (Unimodal)

Dataset: Number of books read by 8 students in a month: 2, 3, 2, 5, 4, 3, 2, 6

Step 1: List the data and count the frequency of each value.

- 2 appears 3 times
- 3 appears 2 times
- 4 appears 1 time
- 5 appears 1 time
- 6 appears 1 time

Step 2: Identify the value with the highest frequency.

- The value 2 appears 3 times, which is the highest frequency.

Result: The mode is 2.

Example 2: Ungrouped Data (Bimodal)

Dataset: Test scores of 6 students: 85, 90, 85, 90, 88, 92

Step 1: Count the frequency of each value.

- 85 appears 2 times
- 90 appears 2 times
- 88 appears 1 time
- 92 appears 1 time

Step 2: Identify the value(s) with the highest frequency.

- Both 85 and 90 appear 2 times, which is the highest frequency.

Result: The dataset is bimodal, with modes 85 and 90.

Example 3: Ungrouped Data (No Mode)

Dataset: Heights of 5 people (in cm): 160, 165, 170, 175, 180

Step 1: Count the frequency of each value.

- 160 appears 1 time
- 165 appears 1 time
- 170 appears 1 time
- 175 appears 1 time
- 180 appears 1 time

Step 2: Identify the value(s) with the highest frequency.

- All values appear exactly once, so there is no value that stands out as the most frequent.

Result: There is no mode for this dataset.

3.2 Mode for Grouped Data

Grouped data is organized into class intervals with corresponding frequencies, often used for large or continuous datasets. To find the mode, we identify the modal class (the class with the highest frequency) and use the formula to estimate the mode within that class.

Formula (Repeated for Clarity):

$$\text{Mode} = L + \left(\frac{f_m - f_1}{(f_m - f_1) + (f_m - f_2)} \right) \times h$$

Example 1: Grouped Data (Test Scores)

Dataset: Test scores of 50 students, grouped into intervals.

Class Interval	**Frequency** (f_i)
50–60	5
60–70	10
70–80	15
80–90	12
90–100	8

Step 1: Identify the modal class.

- The highest frequency is 15, which corresponds to the 70–80 interval.
- Modal class = 70–80

Step 2: Apply the mode formula.

- L (lower boundary of the modal class) = 70
- f_m (frequency of the modal class) = 15
- f_1 (frequency of the class before) = 10 (60–70)
- f_2 (frequency of the class after) = 12 (80–90)
- h (class width) = $80 - 70 = 10$

$$\text{Mode} = 70 + \left(\frac{15 - 10}{(15 - 10) + (15 - 12)} \right) \times 10$$

$$= 70 + \left(\frac{5}{5 + 3} \right) \times 10 = 70 + \left(\frac{5}{8} \right) \times 10 = 70 + 0.625 \times 10 = 70 + 6.25 = 76.25$$

Result: The mode of the test scores is approximately 76.25.

Example 2: Grouped Data (Heights of Plants)

Dataset: Heights (in cm) of 40 plants, grouped as follows.

Height (cm)	**Frequency** (f_i)
10–20	6
20–30	12
30–40	10
40–50	8
50–60	4

Step 1: Identify the modal class.

- The highest frequency is 12, which corresponds to the 20–30 interval.
- Modal class = 20–30

Step 2: Apply the mode formula.

- L (lower boundary of the modal class) = 20

- f_m (frequency of the modal class) = 12
- f_1 (frequency of the class before) = 6 (10–20)
- f_2 (frequency of the class after) = 10 (30–40)
- h (class width) = $30 - 20 = 10$

$$\text{Mode} = 20 + \left(\frac{12-6}{(12-6)+(12-10)}\right) \times 10$$

$$= 20 + \left(\frac{6}{6+2}\right) \times 10 = 20 + \left(\frac{6}{8}\right) \times 10 = 20 + 0.75 \times 10 = 20 + 7.5 = 27.5$$

Result: The mode of the plant heights is 27.5 cm.

Strengths of the Mode

1. **Useful for Categorical Data**:
   - The mode is the only measure of central tendency that can be used with categorical (non-numerical) data, such as colors, names, or preferences.
   - **Example**: Favorite colors of 10 people: red, blue, red, green, red.
     - Mode = red (appears 3 times), which is meaningful for categorical data.
2. **Not Affected by Extreme Values**:
   - The mode is robust to outliers since it only depends on frequency, not the magnitude of the values.
   - **Example (Ungrouped)**: Add an outlier to the books dataset: 2, 3, 2, 5, 4, 3, 2, **50**.
     - The mode remains 2 (appears 3 times), unaffected by the outlier 50.
   - **Example (Grouped)**: If the test scores had an additional class like 150–160 with a frequency of 1, the modal class (70–80) and the mode (76.25) would remain unchanged.
3. **Reflects the Most Common Value**:
   - The mode directly identifies the most frequent or typical value, which is useful in fields like marketing or sociology.
   - **Example**: In the books dataset, the mode (2) indicates that most students read 2 books, providing insight into common behavior.
4. **Simple to Identify in Ungrouped Data**:
   - For ungrouped data, the mode is easy to determine by counting frequencies, requiring no complex calculations.
   - **Example**: In the test scores 85, 90, 85, 90, 88, 92, the modes (85 and 90) are quickly identified by observing frequencies.

Weaknesses of the Mode

1. **May Not Exist or Be Unique**:
   - A dataset may have no mode (if all values occur equally often) or multiple modes (bimodal or multimodal), which can make it less definitive.
   - **Example (Ungrouped)**: Heights 160, 165, 170, 175, 180 have no mode (all frequencies are 1).
   - **Example (Ungrouped)**: Test scores 85, 90, 85, 90, 88, 92 are bimodal (modes 85 and 90), which may complicate interpretation.
2. **Ignores the Distribution of Other Values**:
   - The mode only focuses on the most frequent value(s) and does not consider the overall distribution or magnitude of other data points.
   - **Example (Ungrouped)**: In the books dataset (2, 3, 2, 5, 4, 3, 2), the mode is 2, but it doesn't reflect the presence of higher values like 5 or 6.
   - **Example (Grouped)**: The mode of the test scores (76.25) doesn't indicate the spread of scores in other intervals like 90–100.
3. **Less Useful for Continuous Data**:
   - In continuous datasets, exact repeated values are rare, so the mode is often estimated (as in grouped data), which may not be precise.
   - **Example (Grouped)**: The mode of the plant heights (27.5) is an estimate based on the modal class, but the actual most common height may differ slightly due to the assumption of uniform distribution within the interval.
4. **Not Suitable for Further Statistical Analysis**:
   - The mode lacks algebraic properties, making it less useful for advanced statistical calculations like variance or regression.
   - **Example**: You cannot use the mode (76.25) of the test scores to directly calculate the variance; the mean is required for such computations.

Practical Notes

- Use the mode when you need to identify the most frequent or typical value, especially in categorical data (e.g., most popular product, most common response).
- For grouped data, the mode is an estimate based on the modal class, so interpret it with caution if the distribution within intervals is uneven.
- The mode is particularly valuable in fields like market research (e.g., most common shoe size) or sociology (e.g., most frequent age group), but it's less informative for continuous or highly variable data.

Measures of Dispersion

Measures of dispersion describe the spread or variability of a dataset, indicating how much the data points differ from each other or from the central value. They complement measures of central tendency by providing a fuller picture of the data's distribution. We will explore the **Range**, **Variance**, **Standard Deviation**, and **Coefficient of Variation**, with detailed examples for ungrouped and grouped data, along with their strengths, weaknesses, and practical notes.

1 Range

The **range** is the simplest measure of dispersion, representing the difference between the maximum and minimum values in a dataset. It provides a quick estimate of the spread but only considers the extreme values, ignoring the distribution of other data points.

Formula for the Range

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

The range can be calculated for both ungrouped and grouped data. For grouped data, the maximum and minimum values are typically taken as the upper boundary of the highest class and the lower boundary of the lowest class, respectively.

1.1 Range for Ungrouped Data

Ungrouped data consists of individual data points. To calculate the range, we identify the largest and smallest values and compute their difference.

Example 1: Simple Ungrouped Data

Dataset: Test scores of 5 students: 85, 90, 75, 95, 80

Step 1: Identify the maximum and minimum values.

- Maximum value = 95
- Minimum value = 75

Step 2: Calculate the range.

$$\text{Range} = 95 - 75 = 20$$

Result: The range of the test scores is 20.

Example 2: Ungrouped Data with Repeated Values

Dataset: Number of books read by 8 students in a month: 2, 3, 2, 5, 4, 3, 2, 6

Step 1: Identify the maximum and minimum values.

- Maximum value = 6
- Minimum value = 2

Step 2: Calculate the range.

$$\text{Range} = 6 - 2 = 4$$

Result: The range of the number of books read is 4.

1.2 Range for Grouped Data

Grouped data is organized into class intervals with corresponding frequencies. To calculate the range, we use the boundaries of the class intervals (the lower boundary of the lowest class and the upper boundary of the highest class).

Example 1: Grouped Data (Test Scores)

Dataset: Test scores of 50 students, grouped into intervals.

Class Interval	**Frequency** (f_i)
50–60	5
60–70	10
70–80	15
80–90	12
90–100	8

Step 1: Identify the boundaries of the lowest and highest classes.

- Lowest class: 50–60, so the lower boundary is 50.
- Highest class: 90–100, so the upper boundary is 100.

Step 2: Calculate the range.

$$\text{Range} = 100 - 50 = 50$$

Result: The range of the test scores is 50.

Example 2: Grouped Data (Heights of Plants)

Dataset: Heights (in cm) of 40 plants, grouped as follows.

Height (cm)	**Frequency** (f_i)
10–20	6
20–30	12
30–40	10
40–50	8
50–60	4

Step 1: Identify the boundaries of the lowest and highest classes.

- Lowest class: 10–20, so the lower boundary is 10.
- Highest class: 50–60, so the upper boundary is 60.

Step 2: Calculate the range.

$$\text{Range} = 60 - 10 = 50$$

Result: The range of the plant heights is 50 cm.

Strengths of the Range

1. **Simple to Calculate**:
   - The range requires only the maximum and minimum values, making it quick and easy to compute.
   - **Example (Ungrouped)**: For the test scores 85, 90, 75, 95, 80, the range (20) is found with minimal effort.
   - **Example (Grouped)**: For the test scores dataset, the range (50) is straightforward to calculate using class boundaries.
2. **Useful for Quick Estimates**:
   - The range provides a rapid estimate of the spread, which can be useful for initial data analysis.
   - **Example**: A range of 50 in the test scores dataset immediately indicates a wide spread of scores.
3. **Applicable to Both Ungrouped and Grouped Data**:
   - The range can be calculated for both raw (ungrouped) and summarized (grouped) data.
   - **Example**: We computed the range for ungrouped data (books: 4) and grouped data (heights: 50), showing its versatility.

Weaknesses of the Range

1. **Sensitive to Outliers**:
   - The range is heavily influenced by extreme values, which may not represent the overall variability.
   - **Example (Ungrouped)**: Add an outlier to the test scores: 85, 90, 75, 95, 80, **150**.

     $$\text{Range} = 150 - 75 = 75$$

     The range increases from 20 to 75 due to the outlier, which may exaggerate the perceived spread.
   - **Example (Grouped)**: If the test scores dataset had an additional class like 150–160, the range would jump to $160 - 50 = 110$, even if only one student scored in that range.
2. **Ignores Intermediate Values**:
   - The range only considers the extremes and does not account for the distribution of other data points.
   - **Example (Ungrouped)**: In the books dataset (2, 3, 2, 5, 4, 3, 2, 6), the range is 4, but it doesn't reflect how most values cluster around 2 and 3.
   - **Example (Grouped)**: The range of the test scores (50) doesn't indicate the concentration of scores in the 70–80 interval.
3. **Less Informative for Large Datasets**:
   - For large or complex datasets, the range provides limited insight into the overall variability.
   - **Example**: A range of 50 in the plant heights dataset doesn't reveal whether the data is evenly spread or clustered within certain intervals.

Practical Notes

- Use the range for a quick, preliminary assessment of variability, but complement it with other measures like variance or standard deviation for a more comprehensive analysis.
- Be cautious when interpreting the range in datasets with outliers, as it may overstate the spread.
- The range is often used in quality control (e.g., range of temperatures in a manufacturing process) or weather forecasting (e.g., daily temperature range).

2 Variance

The **variance** measures the average squared deviation of each data point from the mean, providing a more detailed picture of the spread. It accounts for all data points, making it a more comprehensive measure than the range, but it's in squared units, which can make interpretation less intuitive.

Formula for the Variance

- **Ungrouped Data (Population Variance)**:

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

  where x_i are the data values, μ is the population mean, and n is the number of data points.

- **Ungrouped Data (Sample Variance)**:

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

  where $\bar{x}$ is the sample mean, and $n-1$ is used to correct for bias in a sample.

- **Grouped Data (Population Variance)**:

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

- **Grouped Data (Sample Variance)**:

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

  where x_i is the midpoint of the i-th class interval, f_i is the frequency, and $\sum f_i$ is the total frequency.

2.1 Variance for Ungrouped Data

We'll calculate the sample variance (s^2) for ungrouped data, as it's more common in practice when working with samples.

Example 1: Simple Ungrouped Data

Dataset: Test scores of 5 students: 85, 90, 75, 95, 80

Step 1: Calculate the sample mean ($\bar{x}$).

$$\bar{x} = \frac{85 + 90 + 75 + 95 + 80}{5} = \frac{425}{5} = 85$$

Step 2: Compute the squared deviations from the mean and sum them.

Score (x_i)	**Deviation** ($x_i - \bar{x}$)	**Squared Deviation** ($(x_i - \bar{x})^2$)
85	$85 - 85 = 0$	$0^2 = 0$
90	$90 - 85 = 5$	$5^2 = 25$
75	$75 - 85 = -10$	$(-10)^2 = 100$
95	$95 - 85 = 10$	$10^2 = 100$
80	$80 - 85 = -5$	$(-5)^2 = 25$

$$\sum (x_i - \bar{x})^2 = 0 + 25 + 100 + 100 + 25 = 250$$

Step 3: Calculate the sample variance.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{250}{5 - 1} = \frac{250}{4} = 62.5$$

Result: The sample variance of the test scores is 62.5.

Example 2: Ungrouped Data with Repeated Values

Dataset: Number of books read by 8 students: 2, 3, 2, 5, 4, 3, 2, 6

Step 1: Calculate the sample mean ($\bar{x}$).

$$\bar{x} = \frac{2 + 3 + 2 + 5 + 4 + 3 + 2 + 6}{8} = \frac{27}{8} = 3.375$$

Step 2: Compute the squared deviations from the mean and sum them.

Books (x_i)	**Deviation** ($x_i - \bar{x}$)	**Squared Deviation** ($(x_i - \bar{x})^2$)
2	$2 - 3.375 = -1.375$	$(-1.375)^2 = 1.890625$
3	$3 - 3.375 = -0.375$	$(-0.375)^2 = 0.140625$
2	$2 - 3.375 = -1.375$	$(-1.375)^2 = 1.890625$
5	$5 - 3.375 = 1.625$	$1.625^2 = 2.640625$
4	$4 - 3.375 = 0.625$	$0.625^2 = 0.390625$
3	$3 - 3.375 = -0.375$	$(-0.375)^2 = 0.140625$
2	$2 - 3.375 = -1.375$	$(-1.375)^2 = 1.890625$
6	$6 - 3.375 = 2.625$	$2.625^2 = 6.890625$

$$\sum (x_i - \bar{x})^2 = 1.890625 \times 3 + 0.140625 \times 2 + 2.640625 + 0.390625 + 6.890625 = 15.75$$

Step 3: Calculate the sample variance.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1} = \frac{15.75}{8 - 1} = \frac{15.75}{7} = 2.25$$

Result: The sample variance of the number of books read is 2.25.

2.2 Variance for Grouped Data

For grouped data, we calculate the sample variance using the midpoints of the class intervals and their frequencies.

Example 1: Grouped Data (Test Scores)

Dataset: Test scores of 50 students, grouped into intervals.

Class Interval	**Frequency** (f_i)
50–60	5
60–70	10
70–80	15
80–90	12
90–100	8

Step 1: Calculate the mean ($\bar{x}$).

- Midpoints (x_i): 55, 65, 75, 85, 95
- $\sum f_i x_i = (5 \times 55) + (10 \times 65) + (15 \times 75) + (12 \times 85) + (8 \times 95) = 275 + 650 + 1125 + 1020 + 760 = 3830$
- Total frequency ($\sum f_i$) = 50
- $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{3830}{50} = 76.6$

Step 2: Compute the squared deviations from the mean, weighted by frequency.

Class Interval	**Frequency** (f_i)	**Midpoint** (x_i)	**Deviation** ($x_i - \bar{x}$)	$f_i(x_i - \bar{x})^2$
50–60	5	55	$55 - 76.6 = -21.6$	$5 \times (-21.6)^2 = 2332.8$
60–70	10	65	$65 - 76.6 = -11.6$	$10 \times (-11.6)^2 = 1345.6$
70–80	15	75	$75 - 76.6 = -1.6$	$15 \times (-1.6)^2 = 38.4$
80–90	12	85	$85 - 76.6 = 8.4$	$12 \times 8.4^2 = 846.72$
90–100	8	95	$95 - 76.6 = 18.4$	$8 \times 18.4^2 = 2708.48$

$$\sum f_i(x_i - \bar{x})^2 = 2332.8 + 1345.6 + 38.4 + 846.72 + 2708.48 = 7272$$

Step 3: Calculate the sample variance.

$$s^2 = \frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i - 1} = \frac{7272}{50 - 1} = \frac{7272}{49} \approx 148.41$$

Result: The sample variance of the test scores is approximately 148.41.

Example 2: Grouped Data (Heights of Plants)

Dataset: Heights (in cm) of 40 plants, grouped as follows.

Height (cm)	**Frequency** (f_i)
10–20	6
20–30	12
30–40	10
40–50	8
50–60	4

Step 1: Calculate the mean ($\bar{x}$).

- Midpoints (x_i): 15, 25, 35, 45, 55
- $\sum f_i x_i = (6 \times 15) + (12 \times 25) + (10 \times 35) + (8 \times 45) + (4 \times 55) = 90 + 300 + 350 + 360 + 220 = 1320$
- Total frequency ($\sum f_i$) = 40
- $\bar{x} = \frac{\sum f_i x_i}{\sum f_i} = \frac{1320}{40} = 33$

Step 2: Compute the squared deviations from the mean, weighted by frequency.

Height (cm)	**Frequency** (f_i)	**Midpoint** (x_i)	**Deviation** ($x_i - \bar{x}$)	$f_i(x_i - \bar{x})^2$
10–20	6	15	$15 - 33 = -18$	$6 \times (-18)^2 = 1944$
20–30	12	25	$25 - 33 = -8$	$12 \times (-8)^2 = 768$
30–40	10	35	$35 - 33 = 2$	$10 \times 2^2 = 40$
40–50	8	45	$45 - 33 = 12$	$8 \times 12^2 = 1152$
50–60	4	55	$55 - 33 = 22$	$4 \times 22^2 = 1936$

$$\sum f_i(x_i - \bar{x})^2 = 1944 + 768 + 40 + 1152 + 1936 = 5840$$

Step 3: Calculate the sample variance.

$$s^2 = \frac{\sum f_i(x_i - \bar{x})^2}{\sum f_i - 1} = \frac{5840}{40 - 1} = \frac{5840}{39} \approx 149.74$$

Result: The sample variance of the plant heights is approximately 149.74.

Strengths of the Variance

1. **Uses All Data Values**:
   - The variance accounts for every data point's deviation from the mean, providing a comprehensive measure of spread.
   - **Example (Ungrouped)**: In the test scores dataset, the variance (62.5) reflects the spread of all scores relative to the mean (85).
   - **Example (Grouped)**: The variance of the test scores (148.41) considers all class intervals and their frequencies.
2. **Mathematically Robust**:
   - The variance is widely used in statistical analysis (e.g., hypothesis testing, ANOVA) due to its algebraic properties.
   - **Example**: The variance of the test scores can be used to compute the standard deviation or perform further statistical tests.
3. **Applicable to Both Ungrouped and Grouped Data**:
   - The variance can be calculated for both raw and grouped data, making it versatile.
   - **Example**: We computed the variance for ungrouped data (books: 2.25) and grouped data (heights: 149.74).

Weaknesses of the Variance

1. **Squared Units**:
   - The variance is in squared units, which can make interpretation less intuitive.
   - **Example**: The variance of the test scores (148.41) is in "squared score units," which isn't directly comparable to the original scores.
2. **Sensitive to Outliers**:
   - Since deviations are squared, outliers have a disproportionately large impact on the variance.
   - **Example (Ungrouped)**: Add an outlier to the test scores: 85, 90, 75, 95, 80, **150**.

     $$\bar{x} = \frac{575}{6} \approx 95.83, \quad \sum (x_i - \bar{x})^2 \approx 4929.17, \quad s^2 = \frac{4929.17}{5} \approx 985.83$$

     The variance jumps from 62.5 to 985.83 due to the outlier.
3. **Complex to Calculate Manually**:
   - The variance requires multiple steps (mean, deviations, squaring, summing), which can be tedious for large datasets.
   - **Example**: Calculating the variance for the grouped test scores dataset required several steps and careful computation.

3 Standard Deviation

The **standard deviation** is the square root of the variance, providing a measure of dispersion in the same units as the data. It's widely used because it's more interpretable than the variance and indicates the typical deviation from the mean.

Formula for the Standard Deviation

- **Ungrouped Data (Sample Standard Deviation)**:

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

- **Grouped Data (Sample Standard Deviation)**:

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

3.1 Standard Deviation for Ungrouped Data

The standard deviation is simply the square root of the variance calculated earlier.

Example 1: Simple Ungrouped Data

Using the variance from the test scores dataset (85, 90, 75, 95, 80):

$$s^2 = 62.5$$

$$s = \sqrt{62.5} \approx 7.91$$

Result: The standard deviation of the test scores is approximately 7.91.

Example 2: Ungrouped Data with Repeated Values

Using the variance from the books dataset (2, 3, 2, 5, 4, 3, 2, 6):

$$s^2 = 2.25$$

$$s = \sqrt{2.25} = 1.5$$

Result: The standard deviation of the number of books read is 1.5.

3.2 Standard Deviation for Grouped Data

Example 1: Grouped Data (Test Scores)

Using the variance from the test scores dataset:

$$s^2 \approx 148.41$$

$$s = \sqrt{148.41} \approx 12.18$$

Result: The standard deviation of the test scores is approximately 12.18.

Example 2: Grouped Data (Heights of Plants)

Using the variance from the plant heights dataset:

$$s^2 \approx 149.74$$

$$s = \sqrt{149.74} \approx 12.24$$

Result: The standard deviation of the plant heights is approximately 12.24.

Strengths of the Standard Deviation

1. **Same Units as the Data**:
   - The standard deviation is in the same units as the original data, making it easier to interpret than the variance.
   - **Example**: The standard deviation of the test scores (12.18) is in score units, directly comparable to the original scores.
2. **Widely Used in Statistics**:
   - The standard deviation is a key measure in statistical analysis, used in confidence intervals, hypothesis testing, and normal distribution analysis.
   - **Example**: In a normal distribution, about 68% of the data lies within one standard deviation of the mean (e.g., 76.6 ± 12.18 for the test scores).
3. **Reflects Typical Deviation**:
   - It provides a measure of the typical deviation from the mean, which is intuitive for understanding variability.
   - **Example**: A standard deviation of 1.5 for the books dataset indicates that most values are within 1.5 books of the mean (3.375).

Weaknesses of the Standard Deviation

1. **Sensitive to Outliers**:
   - Like the variance, the standard deviation is affected by extreme values due to the squaring of deviations.
   - **Example**: The outlier in the test scores (150) increased the standard deviation to $\sqrt{985.83} \approx 31.4$, which may not reflect the typical spread.
2. **Complex to Calculate Manually**:
   - It requires calculating the variance first, which can be time-consuming for large datasets.
   - **Example**: The standard deviation for the grouped test scores dataset required multiple steps to compute.
3. **Less Intuitive for Non-Normal Data**:
   - The standard deviation is most meaningful for symmetric, normal-like distributions; it may be less informative for skewed data.
   - **Example**: If the test scores were heavily skewed, the standard deviation (12.18) might not accurately represent the spread.

4 Coefficient of Variation

The **coefficient of variation (CV)** is a relative measure of dispersion, expressed as a percentage, that allows comparison of variability across datasets with different units or means. It's particularly useful for comparing the relative spread of different datasets.

Formula for the Coefficient of Variation

$$\text{CV} = \left(\frac{\text{Standard Deviation}}{\text{Mean}} \right) \times 100$$

4.1 Coefficient of Variation for Ungrouped Data

Example 1: Test Scores

Using the test scores dataset (mean = 85, standard deviation ≈ 7.91):

$$\text{CV} = \left(\frac{7.91}{85} \right) \times 100 \approx 9.31\%$$

Result: The coefficient of variation for the test scores is approximately 9.31%.

Example 2: Books

Using the books dataset (mean = 3.375, standard deviation = 1.5):

$$\text{CV} = \left(\frac{1.5}{3.375} \right) \times 100 \approx 44.44\%$$

Result: The coefficient of variation for the number of books read is approximately 44.44%.

4.2 Coefficient of Variation for Grouped Data

Example 1: Test Scores

Using the grouped test scores dataset (mean = 76.6, standard deviation ≈ 12.18):

$$\text{CV} = \left(\frac{12.18}{76.6}\right) \times 100 \approx 15.90\%$$

Result: The coefficient of variation for the test scores is approximately 15.90%.

Example 2: Heights of Plants

Using the plant heights dataset (mean = 33, standard deviation ≈ 12.24):

$$\text{CV} = \left(\frac{12.24}{33}\right) \times 100 \approx 37.09\%$$

Result: The coefficient of variation for the plant heights is approximately 37.09%.

Strengths of the Coefficient of Variation

1. **Relative Measure**:
   - The CV allows comparison of variability across datasets with different units or scales.
   - **Example**: The CV of the test scores (15.90%) and plant heights (37.09%) shows that the heights dataset has relatively more variability, despite different units.
2. **Unitless**:
   - Being a percentage, the CV is unitless, making it ideal for comparing datasets.
   - **Example**: You can compare the CV of test scores (in points) and plant heights (in cm) directly.
3. **Useful in Decision-Making**:
   - The CV is often used in fields like finance (e.g., comparing the risk of investments) or quality control (e.g., consistency of manufacturing processes).
   - **Example**: A lower CV for test scores (15.90%) compared to plant heights (37.09%) indicates more consistency in the scores.

Weaknesses of the Coefficient of Variation

1. **Dependent on the Mean**:
   - The CV is less meaningful when the mean is close to zero, as it can inflate the result.
   - **Example**: If a dataset had a mean of 0.1 and a standard deviation of 0.5, the CV would be 500%, which may be misleading.
2. **Sensitive to Outliers**:
   - Since it relies on the standard deviation, the CV is also affected by extreme values.
   - **Example**: The outlier in the test scores (150) increased the standard deviation, thus inflating the CV.

3. **Not Always Intuitive**:
    - For datasets with very small variability, the CV might be small but still indicate important relative differences.
    - **Example**: A CV of 9.31% for the test scores might seem low, but in some contexts (e.g., standardized testing), this variability could be significant.

Practical Notes

- Use the CV when comparing the relative variability of datasets with different units or scales (e.g., heights in cm vs. weights in kg).
- Be cautious when the mean is close to zero, as the CV can become disproportionately large.
- The CV is valuable in fields like finance (e.g., comparing the risk of different investments) or biology (e.g., comparing variability in growth rates across species).

Categorical Frequency Distribution and Cumulative Frequency

A frequency distribution table is a table that summarizes the number of times (or frequency) that each value occurs in a dataset. For categorical data, it displays the frequency of each category, providing a clear overview of how the data is distributed across the categories.

Example: Survey of Favorite Vacation Types

Let's say we have a survey of 200 people, and we ask them about their favorite type of vacation, which could be one of six categories: Beach, City, Adventure, Nature, Cruise, or Other. The frequency distribution table for this dataset is as follows:

Type of Vacation	**Frequency**
Beach	60
City	40
Adventure	30
Nature	35
Cruise	20
Other	15

Total Frequency: The sum of the frequencies should equal the total number of observations.

$$60 + 40 + 30 + 35 + 20 + 15 = 200$$

The total number of people surveyed is 200, which matches the sum of the frequencies, confirming the table is correct.

Relative Frequency

Relative frequency is the proportion or percentage of a category in a dataset or sample. It is calculated by dividing the frequency of a category by the total number of observations in the dataset or sample.

$$\text{Relative Frequency} = \frac{\text{Frequency of Category}}{\text{Total Number of Observations}}$$

Using the survey data, we calculate the relative frequency for each category (total observations = 200):

- Beach: $\frac{60}{200} = 0.3$
- City: $\frac{40}{200} = 0.2$
- Adventure: $\frac{30}{200} = 0.15$
- Nature: $\frac{35}{200} = 0.175$
- Cruise: $\frac{20}{200} = 0.1$
- Other: $\frac{15}{200} = 0.075$

The updated table with relative frequencies is:

Type of Vacation	Frequency	Relative Frequency
Beach	60	0.3
City	40	0.2
Adventure	30	0.15
Nature	35	0.175
Cruise	20	0.1
Other	15	0.075

Verification: The sum of the relative frequencies should equal 1.

$$0.3 + 0.2 + 0.15 + 0.175 + 0.1 + 0.075 = 1$$

This confirms the relative frequencies are correctly calculated.

Cumulative Frequency

Cumulative frequency is the running total of frequencies of a variable or category in a dataset or sample. It is calculated by adding up the frequencies of the current category and all previous categories in the dataset or sample.

Using the survey data, we calculate the cumulative frequency for each category:

- Beach: 60
- City: $60 + 40 = 100$
- Adventure: $100 + 30 = 130$
- Nature: $130 + 35 = 165$
- Cruise: $165 + 20 = 185$
- Other: $185 + 15 = 200$

The final table including cumulative frequency is:

Type of Vacation	Frequency	Relative Frequency	Cumulative Frequency
Beach	60	0.3	60
City	40	0.2	100
Adventure	30	0.15	130
Nature	35	0.175	165
Cruise	20	0.1	185
Other	15	0.075	200

Verification: The final cumulative frequency should equal the total number of observations, which is 200, confirming the calculation is correct.

Practical Notes

- Frequency distribution tables are useful for summarizing categorical data, providing a quick overview of how often each category occurs.
- Relative frequency helps in understanding the proportional contribution of each category, making it easier to compare datasets of different sizes.

- Cumulative frequency is often used to determine the number of observations below a certain category, which can be helpful in percentile calculations or understanding the distribution's progression.
- These tables are widely used in survey analysis, market research (e.g., customer preferences), and social sciences (e.g., demographic studies).

Graphs for Bivariate Analysis

Bivariate analysis involves examining the relationship between two variables to understand how one influences or relates to the other. It is a fundamental tool in statistics for uncovering patterns, trends, and dependencies in data. In this section, we will explore appropriate graphs for different combinations of variable types, provide detailed examples, and discuss their applications in real-world scenarios.

What is Bivariate Analysis?

- **Bivariate Analysis** refers to the analysis of two variables to determine the **empirical relationship** between them.
- It helps us understand **how one variable influences or relates to another**.

Types of Variables

Variables in bivariate analysis can be classified into two main types:

Type	Description	Example
Categorical	Variables with categories or labels (non-numeric)	Gender, Region, Product Type
Numerical	Variables that are measured in numbers	Age, Salary, Temperature

Combinations of Variables and Suitable Graphs

Depending on the types of the two variables, we choose an appropriate graph to visualize their relationship:

Variable 1	Variable 2	Suitable Graph
Categorical	Categorical	Contingency Table
Numerical	Numerical	Scatter Plot
Categorical	Numerical	Box Plot / Bar Plot

1 Categorical - Categorical

Graph: Contingency Table (Crosstab)

Theory

- Used when both variables are **categorical**.
- Shows how frequently each combination of categories occurs.
- Helps identify any **association or dependency** between the variables.

Example: Gender vs Purchase Decision

Consider a dataset of 200 customers, where we analyze the relationship between *Gender* (Male, Female) and *Purchase Decision* (Yes, No):

Gender	**Purchase: Yes**	**Purchase: No**	**Total**
Male	80	20	100
Female	90	10	100
Total	170	30	200

Statistical Test Used

To determine if there is a significant association between Gender and Purchase Decision, we use the **Chi-Square Test of Independence**. This test checks if the row and column variables are independent.

Sample Calculation Expected count for (Male, Yes):

$$\frac{(\text{Row Total} \times \text{Column Total})}{\text{Grand Total}} = \frac{100 \times 170}{200} = 85$$

We repeat this for all cells to compute the expected frequencies, then apply the Chi-Square formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

where O is the observed frequency, and E is the expected frequency.

Real-World Case Study

- **Retail Marketing**: A company wants to understand if gender affects buying decisions for a product.
- **Insight**: If more females purchase the product (90 vs 80 for males), marketing strategies can target female audiences more effectively.

2 Numerical - Numerical

Graph: Scatter Plot

Theory

- Used when both variables are **numerical**.
- Each point on the graph represents one observation (X, Y).
- Helps detect:
  - **Trends** (positive or negative),
  - **Correlation**,
  - **Outliers**.

Example: Study Hours vs Exam Score

Consider a dataset of students, where we analyze the relationship between *Hours Studied* and *Exam Score*:

Hours Studied	Exam Score
1	50
2	55
3	60
4	65
5	70

A scatter plot would plot each pair (Hours Studied, Exam Score) as a point, revealing a potential trend (e.g., increasing scores with more study hours).

Statistical Calculation: Correlation Coefficient (r)

To quantify the relationship, we calculate the **Pearson Correlation Coefficient (r)**:

- If $r \approx 1$: Strong positive correlation.
- If $r \approx -1$: Strong negative correlation.
- If $r \approx 0$: No correlation.

While we won't compute r here, the scatter plot visually indicates a positive trend (as hours increase, scores increase).

Real-World Case Study

- **Education Sector**: Analyze whether more study time results in better exam scores.
- **Insight**: If a strong positive correlation is observed, educators can create data-driven study plans encouraging more study hours.

3 Categorical - Numerical

Graph: Box Plot / Bar Plot

Theory

- Used when one variable is **categorical** and the other is **numerical**.
- **Box Plots** show the median, quartiles, and outliers for the numerical variable across categories.
- **Bar Plots** compare the **average values** of the numerical variable across categories.

Example: Department vs Salary

Consider a dataset of employees, where we analyze the relationship between *Department* (HR, IT, Finance) and *Salary*:

Department	Salaries ($)
HR	38,000; 40,000; 42,000
IT	60,000; 62,000; 65,000
Finance	53,000; 55,000; 57,000

Mean Calculation

- **HR Mean**:
$$\frac{38000 + 40000 + 42000}{3} = 40000$$
- **IT Mean**:
$$\frac{60000 + 62000 + 65000}{3} = 62333$$
- **Finance Mean**:
$$\frac{53000 + 55000 + 57000}{3} = 55000$$

A bar plot would display these means as bars for each department, while a box plot would show the distribution (median, quartiles, and range) of salaries within each department.

Graph Interpretation

- **Bar Plot**: Compare the average salaries across departments (e.g., IT has the highest mean salary at $62,333).
- **Box Plot**: Spot departments with greater salary variation (e.g., IT salaries range from $60,000 to $65,000, a wider spread than HR).

Real-World Case Study

- **HR Analytics**: Evaluate salary fairness across departments.
- **Insight**: If IT salaries are significantly higher, HR can investigate potential inequalities or plan budget allocations accordingly.

Summary Table

The following table summarizes the graphs for bivariate analysis based on variable types:

Variable Type Combination	**Graph Type**	**Used For**	**Industry Example**
Categorical - Categorical	Contingency Table	Comparing two groups of categories	Marketing, Healthcare
Numerical - Numerical	Scatter Plot	Finding trends or correlations between values	Education, Finance
Categorical - Numerical	Box/Bar Plot	Comparing numerical values by groups	HR, Customer Analysis

Five-Number Summary and Boxplots

The **five-number summary** is a descriptive statistic that provides a concise summary of a dataset by dividing it into four equal parts, also known as quartiles. It is a powerful tool for understanding the central tendency, variability, and distribution of a dataset. Additionally, the five-number summary is often visualized using a **boxplot** (or box-and-whisker plot), which provides a graphical representation of the data's distribution.

1 The Five-Number Summary

The five-number summary consists of the following five values:

1. **Minimum Value**: The smallest value in the dataset.
2. **First Quartile (Q_1)**: The value that separates the lowest 25% of the data from the rest (25th percentile).
3. **Median (Q_2)**: The value that separates the lowest 50% from the highest 50% of the data (50th percentile).
4. **Third Quartile (Q_3)**: The value that separates the lowest 75% of the data from the highest 25% (75th percentile).
5. **Maximum Value**: The largest value in the dataset.

These values divide the dataset into four equal parts, each containing 25% of the observations:

Minimum	Q_1	Median (Q_2)	Q_3	Maximum
25%	25%	25%	25%	

The five-number summary is often used to construct a boxplot, which visually represents the range, median, and quartiles of the dataset.

2 What is a Boxplot?

A **boxplot**, also known as a box-and-whisker plot, is a graphical representation of a dataset that displays its distribution. The boxplot provides a visual summary of the data, including the minimum and maximum values (adjusted for outliers), the first quartile (Q_1), the median (Q_2), and the third quartile (Q_3).

Components of a Boxplot

A boxplot typically includes the following components:

- **Box**: The central box represents the interquartile range (IQR), which is the range between Q_1 (25th percentile) and Q_3 (75th percentile). It contains the middle 50% of the data.
- **Median Line**: A line inside the box marks the median (Q_2, 50th percentile), dividing the dataset into two halves.

- **Whiskers**: Lines extending from the edges of the box to the "minimum" and "maximum" values, adjusted for outliers:
  - **"Minimum"**: Typically defined as $Q_1 - 1.5 \times \text{IQR}$, but not less than the smallest data point.
  - **"Maximum"**: Typically defined as $Q_3 + 1.5 \times \text{IQR}$, but not more than the largest data point.
- **Outliers**: Data points that fall outside the whiskers (i.e., below $Q_1 - 1.5 \times \text{IQR}$ or above $Q_3 + 1.5 \times \text{IQR}$) are plotted as individual points.
- **Interquartile Range (IQR)**: The difference between Q_3 and Q_1, i.e., $\text{IQR} = Q_3 - Q_1$.

Note: A visual representation of a boxplot would typically be included here, with labels for each component (e.g., minimum as the starting point of the lower whisker, Q_1 as the lower edge of the box, median as a line inside the box, Q_3 as the upper edge of the box, maximum as the endpoint of the upper whisker, and outliers as individual points). However, due to limitations, the image is not included in this document.

3 How to Create a Boxplot with Examples

To create a boxplot, we need to compute the five-number summary and identify any outliers using the IQR. Below, we provide detailed examples for both ungrouped and grouped data.

Example 1: Boxplot for Ungrouped Data

Dataset: Test scores of 12 students: 65, 70, 72, 75, 78, 80, 82, 85, 88, 90, 95, 100.

Step 1: Sort the data in ascending order (already sorted):

$$65, 70, 72, 75, 78, 80, 82, 85, 88, 90, 95, 100$$

Positions: 1, 2, 3, ..., 12 ($N = 12$).

Step 2: Compute the five-number summary.

- **Minimum**: The smallest value is 65.
- **Maximum**: The largest value is 100.
- **Median (Q_2)**: Since $N = 12$ (even), the median is the average of the 6th and 7th values:

$$\text{6th value} = 80, \quad \text{7th value} = 82$$

$$Q_2 = \frac{80 + 82}{2} = 81$$

- **First Quartile (Q_1)**: The 25th percentile position:

$$PL = \frac{25}{100}(12 + 1) = 0.25 \times 13 = 3.25$$

Between the 3rd and 4th values:

$$\text{3rd value} = 72, \quad \text{4th value} = 75$$

$$Q_1 = 72 + 0.25 \times (75 - 72) = 72 + 0.25 \times 3 = 72 + 0.75 = 72.75$$

- **Third Quartile (Q_3)**: The 75th percentile position:

$$PL = \frac{75}{100}(12+1) = 0.75 \times 13 = 9.75$$

Between the 9th and 10th values:

$$\text{9th value} = 88, \quad \text{10th value} = 90$$

$$Q_3 = 88 + 0.75 \times (90 - 88) = 88 + 0.75 \times 2 = 88 + 1.5 = 89.5$$

Five-Number Summary:

$$\text{Minimum} = 65, \quad Q_1 = 72.75, \quad Q_2 = 81, \quad Q_3 = 89.5, \quad \text{Maximum} = 100$$

Step 3: Calculate the IQR and determine the whiskers and outliers.

$$\text{IQR} = Q_3 - Q_1 = 89.5 - 72.75 = 16.75$$

$$1.5 \times \text{IQR} = 1.5 \times 16.75 = 25.125$$

- Lower whisker: $Q_1 - 1.5 \times \text{IQR} = 72.75 - 25.125 = 47.625$. Since 65 (the minimum) is greater than 47.625, the lower whisker extends to 65.
- Upper whisker: $Q_3 + 1.5 \times \text{IQR} = 89.5 + 25.125 = 114.625$. Since 100 (the maximum) is less than 114.625, the upper whisker extends to 100.
- Outliers: There are no values below 47.625 or above 114.625, so there are no outliers.

Step 4: Describe the boxplot:

- The box extends from $Q_1 = 72.75$ to $Q_3 = 89.5$, with a line at the median $Q_2 = 81$.
- The lower whisker extends from 65 to 72.75.
- The upper whisker extends from 89.5 to 100.
- There are no outliers to plot.

Example 2: Boxplot for Ungrouped Data with Outliers

Dataset: Heights (in cm) of 10 individuals: 150, 155, 160, 162, 165, 168, 170, 175, 180, 200.

Step 1: Sort the data (already sorted):

$$150, 155, 160, 162, 165, 168, 170, 175, 180, 200$$

Step 2: Compute the five-number summary.

- **Minimum**: 150.
- **Maximum**: 200.

- **Median (Q_2)**: $N = 10$, so the median is the average of the 5th and 6th values:

$$\text{5th value} = 165, \quad \text{6th value} = 168$$

$$Q_2 = \frac{165 + 168}{2} = 166.5$$

- Q_1: Position:

$$PL = \frac{25}{100}(10 + 1) = 0.25 \times 11 = 2.75$$

Between the 2nd and 3rd values:

$$\text{2nd value} = 155, \quad \text{3rd value} = 160$$

$$Q_1 = 155 + 0.75 \times (160 - 155) = 155 + 0.75 \times 5 = 155 + 3.75 = 158.75$$

- Q_3: Position:

$$PL = \frac{75}{100}(10 + 1) = 0.75 \times 11 = 8.25$$

Between the 8th and 9th values:

$$\text{8th value} = 175, \quad \text{9th value} = 180$$

$$Q_3 = 175 + 0.25 \times (180 - 175) = 175 + 0.25 \times 5 = 175 + 1.25 = 176.25$$

Five-Number Summary:

$$\text{Minimum} = 150, \quad Q_1 = 158.75, \quad Q_2 = 166.5, \quad Q_3 = 176.25, \quad \text{Maximum} = 200$$

Step 3: Calculate the IQR and determine the whiskers and outliers.

$$\text{IQR} = 176.25 - 158.75 = 17.5$$

$$1.5 \times \text{IQR} = 1.5 \times 17.5 = 26.25$$

- Lower whisker: $158.75 - 26.25 = 132.5$. The smallest value (150) is above 132.5, so the lower whisker extends to 150.
- Upper whisker: $176.25 + 26.25 = 202.5$. The largest value (200) is below 202.5, but we check for outliers.
- Outliers: Values below 132.5 or above 202.5. There are none below 132.5, but 200 is below 202.5. We need to check if any values are outside the whiskers relative to the data: all values between 150 and 180 are within bounds, but 200 might be an outlier based on typical distribution. Since 200 is very close to 202.5, we note it but often include it in the whisker for simplicity unless further context suggests otherwise.

For simplicity, we assume 200 is not treated as an outlier in this context (as it's just below the threshold).

Step 4: Describe the boxplot:

- The box extends from $Q_1 = 158.75$ to $Q_3 = 176.25$, with a line at $Q_2 = 166.5$.
- The lower whisker extends from 150 to 158.75.
- The upper whisker extends from 176.25 to 200.
- No outliers are plotted (assuming 200 is included in the whisker).

Example 3: Boxplot for Grouped Data

Dataset: Test scores of 50 students, grouped into intervals.

Class Interval	**Frequency** (f_i)
50–60	5
60–70	10
70–80	15
80–90	12
90–100	8

Step 1: Compute the cumulative frequency to find the five-number summary.

Class Interval	**Frequency** (f_i)	**Cumulative Frequency**
50–60	5	5
60–70	10	15
70–80	15	30
80–90	12	42
90–100	8	50

Total frequency $N = 50$.
Step 2: Compute the five-number summary.

- **Minimum**: The lower boundary of the first class, 50.
- **Maximum**: The upper boundary of the last class, 100.
- **Median (Q_2)**: The 50th percentile, position $\frac{50}{100} \times 50 = 25$. Cumulative frequency exceeds 25 at 70–80:

$$Q_2 = 70 + \left(\frac{25 - 15}{15}\right) \times 10 = 70 + \left(\frac{10}{15}\right) \times 10 = 70 + 6.67 = 76.67$$

- Q_1: The 25th percentile, position $\frac{25}{100} \times 50 = 12.5$. Cumulative frequency exceeds 12.5 at 60–70:

$$Q_1 = 60 + \left(\frac{12.5 - 5}{10}\right) \times 10 = 60 + \left(\frac{7.5}{10}\right) \times 10 = 60 + 7.5 = 67.5$$

- Q_3: The 75th percentile, position $\frac{75}{100} \times 50 = 37.5$. Cumulative frequency exceeds 37.5 at 80–90:

$$Q_3 = 80 + \left(\frac{37.5 - 30}{12}\right) \times 10 = 80 + \left(\frac{7.5}{12}\right) \times 10 = 80 + 6.25 = 86.25$$

Five-Number Summary:

$$\text{Minimum} = 50, \quad Q_1 = 67.5, \quad Q_2 = 76.67, \quad Q_3 = 86.25, \quad \text{Maximum} = 100$$

Step 3: Calculate the IQR and determine the whiskers and outliers.

$$\text{IQR} = 86.25 - 67.5 = 18.75$$

$$1.5 \times \text{IQR} = 1.5 \times 18.75 = 28.125$$

- Lower whisker: $67.5 - 28.125 = 39.375$. The minimum (50) is above 39.375, so the lower whisker extends to 50.
- Upper whisker: $86.25 + 28.125 = 114.375$. The maximum (100) is below 114.375, so the upper whisker extends to 100.
- Outliers: We cannot identify specific outliers in grouped data, but the whiskers are set based on the boundaries.

Step 4: Describe the boxplot:

- The box extends from $Q_1 = 67.5$ to $Q_3 = 86.25$, with a line at $Q_2 = 76.67$.
- The lower whisker extends from 50 to 67.5.
- The upper whisker extends from 86.25 to 100.
- Outliers cannot be plotted for grouped data in this format.

Practical Notes

- **Understanding Distribution**: The five-number summary and boxplot provide a quick overview of the dataset's central tendency, spread, and skewness. For example, if $Q_2 - Q_1 > Q_3 - Q_2$, the data is skewed to the left.
- **Outlier Detection**: Boxplots are particularly useful for identifying outliers, which can be critical in fields like finance (e.g., detecting fraudulent transactions) or quality control (e.g., identifying defective products).
- **Comparing Datasets**: Boxplots allow for easy comparison of multiple datasets. For instance, comparing test scores of two classes using side-by-side boxplots can reveal differences in performance and variability.
- **Limitations**: For grouped data, the five-number summary is an approximation, and individual outliers cannot be identified. Additionally, boxplots do not show the full distribution shape (e.g., bimodality).

Quantiles and Percentiles

Quantiles are statistical measures used to divide a dataset into equal-sized groups, where each group contains the same number of observations. They are crucial for understanding the distribution of data, summarizing and comparing datasets, and identifying outliers. Quantiles are measures of variability and position, providing insights into the spread and shape of the data.

Importance of Quantiles

Quantiles are used to:

- Understand the **distribution of data** by showing how data is spread across different segments.
- **Summarize and compare datasets** by providing key positional values (e.g., median, quartiles).
- **Identify outliers** by examining extreme quantiles (e.g., values below the 1st percentile or above the 99th percentile).

Types of Quantiles

There are several types of quantiles commonly used in statistical analysis:

- **Quartiles**: Divide the data into four equal parts:
  - Q_1 (25th percentile): 25% of the data lies below this value.
  - Q_2 (50th percentile or median): 50% of the data lies below this value.
  - Q_3 (75th percentile): 75% of the data lies below this value.
- **Deciles**: Divide the data into ten equal parts:
  - D_1 (10th percentile), D_2 (20th percentile), ..., D_9 (90th percentile).
- **Percentiles**: Divide the data into 100 equal parts:
  - P_1 (1st percentile), P_2 (2nd percentile), ..., P_{99} (99th percentile).
- **Quintiles**: Divide the data into five equal parts:
  - 20th percentile, 40th percentile, 60th percentile, 80th percentile.

Key Points to Remember

When calculating quantiles:

1. The data must be **sorted in ascending order** (from low to high).
2. Quantiles represent the **position of an observation**, not necessarily an actual data value.
3. They often require **interpolation** if the position is not an integer.
4. All other quantiles (quartiles, deciles, quintiles) can be derived from percentiles.

1 Percentiles

A **percentile** is a statistical measure that indicates the percentage of observations in a dataset that fall below a particular value. For example, the 75th percentile (P_{75}) is the value below which 75% of the observations lie.

Formula to Calculate Percentile Position

The position of the p-th percentile in a dataset can be calculated using:

$$PL = \frac{p}{100}(N+1)$$

where:

- PL = the desired percentile position,
- p = the percentile rank (expressed as a percentage, e.g., 75 for the 75th percentile),
- N = the total number of observations in the dataset.

If PL is not an integer, interpolation is required between the two nearest data points.

Percentile for Ungrouped Data

Example 1: Finding the 75th Percentile

Dataset: Test scores of 10 students: 78, 82, 84, 88, 91, 93, 94, 96, 98, 99.

Step 1: Sort the data in ascending order (already sorted):

$$78, 82, 84, 88, 91, 93, 94, 96, 98, 99$$

Positions: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

Step 2: Calculate the position of the 75th percentile ($p = 75$, $N = 10$):

$$PL = \frac{75}{100}(10+1) = \frac{75}{100} \times 11 = \frac{3}{4} \times 11 = 8.25$$

The position 8.25 lies between the 8th and 9th values in the dataset.

Step 3: Identify the 8th and 9th values:

$$\text{8th value} = 96, \quad \text{9th value} = 98$$

Step 4: Interpolate to find the 75th percentile:

$$P_{75} = 96 + 0.25 \times (98 - 96) = 96 + 0.25 \times 2 = 96 + 0.5 = 96.5$$

Result: The 75th percentile of the test scores is 96.5. This means 75% of the students scored below 96.5.

Example 2: Finding the 90th Percentile

Using the same dataset: 78, 82, 84, 88, 91, 93, 94, 96, 98, 99.

Step 1: The data is already sorted.

Step 2: Calculate the position of the 90th percentile ($p = 90$, $N = 10$):

$$PL = \frac{90}{100}(10 + 1) = 0.9 \times 11 = 9.9$$

The position 9.9 lies between the 9th and 10th values.

Step 3: Identify the 9th and 10th values:

$$\text{9th value} = 98, \quad \text{10th value} = 99$$

Step 4: Interpolate:

$$P_{90} = 98 + 0.9 \times (99 - 98) = 98 + 0.9 \times 1 = 98 + 0.9 = 98.9$$

Result: The 90th percentile is 98.9, meaning 90% of the scores are below 98.9.

Percentile for Grouped Data

For grouped data, percentiles are estimated using cumulative frequencies and interpolation within the appropriate class interval.

Formula for Percentile in Grouped Data:

$$P_p = L + \left(\frac{\frac{p}{100}N - CF}{f} \right) \times h$$

where:

- P_p = the p-th percentile,
- L = lower boundary of the percentile class (the class where the cumulative frequency first exceeds $\frac{p}{100}N$),
- N = total number of observations,
- CF = cumulative frequency of the class before the percentile class,
- f = frequency of the percentile class,
- h = width of the class interval.

Example: 60th Percentile for Grouped Data

Dataset: Test scores of 50 students, grouped into intervals.

Class Interval	**Frequency** (f_i)
50–60	5
60–70	10
70–80	15
80–90	12
90–100	8

Step 1: Calculate the total frequency (N) and find $\frac{p}{100}N$ for $p = 60$:

$$N = 5 + 10 + 15 + 12 + 8 = 50, \quad \frac{60}{100} \times 50 = 30$$

Step 2: Compute the cumulative frequency to identify the percentile class:

Class Interval	**Frequency** (f_i)	**Cumulative Frequency**
50–60	5	5
60–70	10	15
70–80	15	30
80–90	12	42
90–100	8	50

The cumulative frequency first exceeds 30 at the 70–80 interval, so the percentile class is 70–80.
Step 3: Apply the percentile formula:

- L (lower boundary of the percentile class) = 70,
- CF (cumulative frequency before the percentile class) = 15,
- f (frequency of the percentile class) = 15,
- h (class width) = $80 - 70 = 10$,
- $\frac{p}{100}N = 30$.

$$P_{60} = 70 + \left(\frac{30 - 15}{15}\right) \times 10 = 70 + \left(\frac{15}{15}\right) \times 10 = 70 + 1 \times 10 = 80$$

Result: The 60th percentile of the test scores is 80, meaning 60% of the students scored below 80.

2 Quartiles

Quartiles divide the data into four equal parts: Q_1 (25th percentile), Q_2 (50th percentile or median), and Q_3 (75th percentile).

Quartiles for Ungrouped Data

Using the same dataset: 78, 82, 84, 88, 91, 93, 94, 96, 98, 99 ($N = 10$).

Q_1 **(25th Percentile)**:

$$PL = \frac{25}{100}(10 + 1) = 0.25 \times 11 = 2.75$$

Between the 2nd and 3rd values:

$$\text{2nd value} = 82, \quad \text{3rd value} = 84$$

$$Q_1 = 82 + 0.75 \times (84 - 82) = 82 + 0.75 \times 2 = 82 + 1.5 = 83.5$$

Q_2 **(50th Percentile or Median)**:

$$PL = \frac{50}{100}(10 + 1) = 0.5 \times 11 = 5.5$$

Between the 5th and 6th values:

$$\text{5th value} = 91, \quad \text{6th value} = 93$$

$$Q_2 = 91 + 0.5 \times (93 - 91) = 91 + 0.5 \times 2 = 91 + 1 = 92$$

Q_3 **(75th Percentile)**: Already calculated as 96.5 (see percentile example).
Result: $Q_1 = 83.5$, $Q_2 = 92$, $Q_3 = 96.5$.

Quartiles for Grouped Data

Using the grouped test scores dataset (from the percentile example):

Q_1 **(25th Percentile)**:

$$\frac{25}{100} \times 50 = 12.5$$

Cumulative frequency exceeds 12.5 at 60–70 ($CF = 15$):

$$Q_1 = 60 + \left(\frac{12.5 - 5}{10}\right) \times 10 = 60 + \left(\frac{7.5}{10}\right) \times 10 = 60 + 7.5 = 67.5$$

Q_2 **(50th Percentile)**:

$$\frac{50}{100} \times 50 = 25$$

Cumulative frequency exceeds 25 at 70–80 ($CF = 30$):

$$Q_2 = 70 + \left(\frac{25 - 15}{15}\right) \times 10 = 70 + \left(\frac{10}{15}\right) \times 10 = 70 + 6.67 = 76.67$$

Q_3 **(75th Percentile)**:

$$\frac{75}{100} \times 50 = 37.5$$

Cumulative frequency exceeds 37.5 at 80–90 ($CF = 42$):

$$Q_3 = 80 + \left(\frac{37.5 - 30}{12}\right) \times 10 = 80 + \left(\frac{7.5}{12}\right) \times 10 = 80 + 6.25 = 86.25$$

Result: $Q_1 = 67.5$, $Q_2 = 76.67$, $Q_3 = 86.25$.

3 Deciles

Deciles divide the data into ten equal parts: D_1 (10th percentile), D_2 (20th percentile), ..., D_9 (90th percentile).

Deciles for Ungrouped Data

Using the same dataset: 78, 82, 84, 88, 91, 93, 94, 96, 98, 99.

D_4 **(40th Percentile)**:

$$PL = \frac{40}{100}(10 + 1) = 0.4 \times 11 = 4.4$$

Between the 4th and 5th values:

$$\text{4th value} = 88, \quad \text{5th value} = 91$$

$$D_4 = 88 + 0.4 \times (91 - 88) = 88 + 0.4 \times 3 = 88 + 1.2 = 89.2$$

Result: The 4th decile (D_4) is 89.2.

Deciles for Grouped Data

Using the grouped test scores dataset:

D_7 **(70th Percentile)**:

$$\frac{70}{100} \times 50 = 35$$

Cumulative frequency exceeds 35 at 80–90 ($CF = 42$):

$$D_7 = 80 + \left(\frac{35 - 30}{12}\right) \times 10 = 80 + \left(\frac{5}{12}\right) \times 10 = 80 + 4.17 = 84.17$$

Result: The 7th decile (D_7) is 84.17.

4 Quintiles

Quintiles divide the data into five equal parts: 20th percentile, 40th percentile, 60th percentile, and 80th percentile.

Quintiles for Ungrouped Data

Using the same dataset:

2nd Quintile (40th Percentile): Already calculated as $D_4 = 89.2$.

3rd Quintile (60th Percentile):

$$PL = \frac{60}{100}(10 + 1) = 0.6 \times 11 = 6.6$$

Between the 6th and 7th values:

$$\text{6th value} = 93, \quad \text{7th value} = 94$$

$$\text{3rd Quintile} = 93 + 0.6 \times (94 - 93) = 93 + 0.6 \times 1 = 93.6$$

Result: The 3rd quintile is 93.6.

Quintiles for Grouped Data

Using the grouped test scores dataset:

4th Quintile (80th Percentile):

$$\frac{80}{100} \times 50 = 40$$

Cumulative frequency exceeds 40 at 80–90 ($CF = 42$):

$$\text{4th Quintile} = 80 + \left(\frac{40 - 30}{12}\right) \times 10 = 80 + \left(\frac{10}{12}\right) \times 10 = 80 + 8.33 = 88.33$$

Result: The 4th quintile is 88.33.

5 Percentile Rank of a Value

The percentile rank of a specific value in a dataset indicates the percentage of observations that fall below that value.

Formula for Percentile Rank

$$\text{Percentile Rank} = \frac{x + 0.5y}{N} \times 100$$

where:

- x = number of values strictly below the given value,
- y = number of values equal to the given value,
- N = total number of values in the dataset.

Example: Percentile Rank of 84

Using the dataset: 78, 82, 84, 88, 91, 93, 94, 96, 98, 99.

Step 1: Identify x, y, and N for the value 84:

- x (values below 84): 78, 82 $\rightarrow$ 2 values,
- y (values equal to 84): 1 value,
- N (total values): 10.

Step 2: Calculate the percentile rank:

$$\text{Percentile Rank} = \frac{2 + 0.5 \times 1}{10} \times 100 = \frac{2 + 0.5}{10} \times 100 = \frac{2.5}{10} \times 100 = 25\%$$

Result: The percentile rank of 84 is 25%, meaning 25% of the scores are below 84.

Example: Percentile Rank with Repeated Values

Dataset: 78, 82, 84, 84, 88, 91, 93, 94, 96, 98.

Step 1: For the value 84:

- x (values below 84): 78, 82 $\rightarrow$ 2 values,
- y (values equal to 84): 2 values,
- $N = 10$.

Step 2: Calculate the percentile rank:

$$\text{Percentile Rank} = \frac{2 + 0.5 \times 2}{10} \times 100 = \frac{2 + 1}{10} \times 100 = \frac{3}{10} \times 100 = 30\%$$

Result: The percentile rank of 84 is 30%.

Practical Notes

- **Comparing Distributions**: Quartiles and percentiles are useful for comparing the spread and central tendency of different datasets (e.g., test scores of two classes).
- **Outlier Detection**: Values below the 1st percentile or above the 99th percentile are often considered outliers.
- **Grouped Data Approximation**: For grouped data, quantiles are estimates based on interpolation, assuming a uniform distribution within each class interval.
- **Applications**: Quantiles are widely used in fields like education (e.g., standardized test scores), finance (e.g., income distributions), and healthcare (e.g., growth charts for children).

Comparing Datasets Using Five-Number Summary and Boxplots

The **five-number summary** and **boxplots** are powerful tools for comparing the distribution, central tendency, and variability of multiple datasets. By calculating the five-number summary (minimum, Q_1, median (Q_2), Q_3, maximum) for each dataset and visualizing them with side-by-side boxplots, we can quickly identify differences in spread, skewness, and potential outliers. This section builds on the concepts of quantiles (e.g., quartiles) and boxplots, providing practical examples for both ungrouped and grouped data.

1 Comparing Ungrouped Datasets

We'll compare the test scores of two classes, Class A and Class B, using their five-number summaries and describe how their boxplots would look side by side. Then, we'll add a new example comparing the weights of cats by sex to further illustrate the use of side-by-side boxplots.

Dataset 1: Class A Test Scores

Data: 65, 70, 72, 75,78, 80, 82, 85, 88, 90, 95, 100 (from previous example).

Five-Number Summary (computed previously):

$$\text{Minimum} = 65, \quad Q_1 = 72.75, \quad Q_2 = 81, \quad Q_3 = 89.5, \quad \text{Maximum} = 100$$

IQR and Whiskers:

$$\text{IQR} = 89.5 - 72.75 = 16.75, \quad 1.5 \times \text{IQR} = 25.125$$

$$\text{Lower Whisker} = 72.75 - 25.125 = 47.625 \quad \text{(extends to 65)},$$

$$\text{Upper Whisker} = 89.5 + 25.125 = 114.625 \quad \text{(extends to 100)}$$

No outliers.

Dataset 2: Class B Test Scores

Data: 55, 60, 62, 65, 68, 70, 72, 75, 80, 85, 90, 92.

Step 1: Sort the data (already sorted):

$$55, 60, 62, 65, 68, 70, 72, 75, 80, 85, 90, 92$$

Positions: 1, 2, ..., 12 ($N = 12$).

Step 2: Compute the five-number summary.

- **Minimum**: 55.
- **Maximum**: 92.

- **Median (Q_2)**: Average of the 6th and 7th values:

$$\text{6th value} = 70, \quad \text{7th value} = 72$$
$$Q_2 = \frac{70 + 72}{2} = 71$$

- Q_1: Position:

$$PL = \frac{25}{100}(12 + 1) = 0.25 \times 13 = 3.25$$

Between the 3rd and 4th values:

$$\text{3rd value} = 62, \quad \text{4th value} = 65$$
$$Q_1 = 62 + 0.25 \times (65 - 62) = 62 + 0.25 \times 3 = 62 + 0.75 = 62.75$$

- Q_3: Position:

$$PL = \frac{75}{100}(12 + 1) = 0.75 \times 13 = 9.75$$

Between the 9th and 10th values:

$$\text{9th value} = 80, \quad \text{10th value} = 85$$
$$Q_3 = 80 + 0.75 \times (85 - 80) = 80 + 0.75 \times 5 = 80 + 3.75 = 83.75$$

Five-Number Summary for Class B:

$$\text{Minimum} = 55, \quad Q_1 = 62.75, \quad Q_2 = 71, \quad Q_3 = 83.75, \quad \text{Maximum} = 92$$

Step 3: Calculate the IQR and whiskers.

$$\text{IQR} = 83.75 - 62.75 = 21$$
$$1.5 \times \text{IQR} = 1.5 \times 21 = 31.5$$
$$\text{Lower Whisker} = 62.75 - 31.5 = 31.25 \quad \text{(extends to 55)},$$
$$\text{Upper Whisker} = 83.75 + 31.5 = 115.25 \quad \text{(extends to 92)}$$

No outliers.

Comparison Using Boxplots

Side-by-Side Boxplot Description:

- **Class A**: Box from 72.75 to 89.5, median at 81, whiskers from 65 to 100.
- **Class B**: Box from 62.75 to 83.75, median at 71, whiskers from 55 to 92.

Observations:

- **Central Tendency**: Class A has a higher median (81 vs. 71), indicating better overall performance.
- **Spread**: Class A's range (100 - 65 = 35) is slightly smaller than Class B's (92 - 55 = 37), but Class B has a larger IQR (21 vs. 16.75), suggesting greater variability in the middle 50% of scores.
- **Skewness**: Class A's median (81) is closer to Q_1 (72.75) than Q_3 (89.5), indicating a slight right skew. Class B's median (71) is roughly centered between Q_1 (62.75) and Q_3 (83.75), suggesting a more symmetric distribution.

Dataset 3: Cat Weight by Sex

To further illustrate the use of side-by-side boxplots, we'll compare the weights (in kg) of male and female cats.

Data for Male Cats: 3.2, 3.5, 3.8, 4.0, 4.2, 4.5, 4.8, 5.0 ($N = 8$).
Data for Female Cats: 2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0, 4.2 ($N = 8$).

Five-Number Summary for Male Cats

Step 1: Data is already sorted:

$$3.2, 3.5, 3.8, 4.0, 4.2, 4.5, 4.8, 5.0$$

Step 2: Compute the five-number summary.

- **Minimum**: 3.2.
- **Maximum**: 5.0.
- **Median (Q_2)**: Average of the 4th and 5th values:

$$\text{4th value} = 4.0, \quad \text{5th value} = 4.2$$

$$Q_2 = \frac{4.0 + 4.2}{2} = 4.1$$

- Q_1: Position:

$$PL = \frac{25}{100}(8 + 1) = 0.25 \times 9 = 2.25$$

Between the 2nd and 3rd values:

$$\text{2nd value} = 3.5, \quad \text{3rd value} = 3.8$$

$$Q_1 = 3.5 + 0.25 \times (3.8 - 3.5) = 3.5 + 0.25 \times 0.3 = 3.5 + 0.075 = 3.575$$

- Q_3: Position:

$$PL = \frac{75}{100}(8 + 1) = 0.75 \times 9 = 6.75$$

Between the 6th and 7th values:

$$\text{6th value} = 4.5, \quad \text{7th value} = 4.8$$

$$Q_3 = 4.5 + 0.75 \times (4.8 - 4.5) = 4.5 + 0.75 \times 0.3 = 4.5 + 0.225 = 4.725$$

Five-Number Summary for Male Cats:

$$\text{Minimum} = 3.2, \quad Q_1 = 3.575, \quad Q_2 = 4.1, \quad Q_3 = 4.725, \quad \text{Maximum} = 5.0$$

Step 3: Calculate the IQR and whiskers.

$$\text{IQR} = 4.725 - 3.575 = 1.15$$

$$1.5 \times \text{IQR} = 1.5 \times 1.15 = 1.725$$

$$\text{Lower Whisker} = 3.575 - 1.725 = 1.85 \quad \text{(extends to 3.2)},$$

$$\text{Upper Whisker} = 4.725 + 1.725 = 6.45 \quad \text{(extends to 5.0)}$$

No outliers.

Five-Number Summary for Female Cats

Step 1: Data is already sorted:

$$2.8, 3.0, 3.2, 3.4, 3.6, 3.8, 4.0, 4.2$$

Step 2: Compute the five-number summary.

- **Minimum**: 2.8.
- **Maximum**: 4.2.
- **Median (Q_2)**: Average of the 4th and 5th values:

$$\text{4th value} = 3.4, \quad \text{5th value} = 3.6$$

$$Q_2 = \frac{3.4 + 3.6}{2} = 3.5$$

- Q_1: Position:

$$PL = \frac{25}{100}(8 + 1) = 0.25 \times 9 = 2.25$$

Between the 2nd and 3rd values:

$$\text{2nd value} = 3.0, \quad \text{3rd value} = 3.2$$

$$Q_1 = 3.0 + 0.25 \times (3.2 - 3.0) = 3.0 + 0.25 \times 0.2 = 3.0 + 0.05 = 3.05$$

- Q_3: Position:

$$PL = \frac{75}{100}(8 + 1) = 0.75 \times 9 = 6.75$$

Between the 6th and 7th values:

$$\text{6th value} = 3.8, \quad \text{7th value} = 4.0$$

$$Q_3 = 3.8 + 0.75 \times (4.0 - 3.8) = 3.8 + 0.75 \times 0.2 = 3.8 + 0.15 = 3.95$$

Five-Number Summary for Female Cats:

$$\text{Minimum} = 2.8, \quad Q_1 = 3.05, \quad Q_2 = 3.5, \quad Q_3 = 3.95, \quad \text{Maximum} = 4.2$$

Step 3: Calculate the IQR and whiskers.

$$\text{IQR} = 3.95 - 3.05 = 0.9$$

$$1.5 \times \text{IQR} = 1.5 \times 0.9 = 1.35$$

$$\text{Lower Whisker} = 3.05 - 1.35 = 1.7 \quad \text{(extends to 2.8)},$$

$$\text{Upper Whisker} = 3.95 + 1.35 = 5.3 \quad \text{(extends to 4.2)}$$

No outliers.

Side-by-Side Boxplot for Cat Weight by Sex

Description:

- **Male Cats**: Box from 3.575 to 4.725, median at 4.1, whiskers from 3.2 to 5.0.
- **Female Cats**: Box from 3.05 to 3.95, median at 3.5, whiskers from 2.8 to 4.2.

Observations:

- **Central Tendency**: Male cats have a higher median weight (4.1 kg vs. 3.5 kg), indicating they are generally heavier.
- **Spread**: Male cats have a larger range (5.0 - 3.2 = 1.8 kg vs. 4.2 - 2.8 = 1.4 kg) and a larger IQR (1.15 vs. 0.9), suggesting greater variability in weights.
- **Skewness**: Both distributions are relatively symmetric, as the medians are roughly centered between Q_1 and Q_3.

2 Comparing Grouped Datasets

We'll compare the test scores of two groups of students (Group 1 and Group 2) using grouped data.

Dataset 1: Group 1 Test Scores

From previous examples:

Class Interval	**Frequency** (f_i)
50–60	5
60–70	10
70–80	15
80–90	12
90–100	8

Five-Number Summary (computed previously):

$$\text{Minimum} = 50, \quad Q_1 = 67.5, \quad Q_2 = 76.67, \quad Q_3 = 86.25, \quad \text{Maximum} = 100$$

Dataset 2: Group 2 Test Scores

Data:

Class Interval	**Frequency** (f_i)
40–50	4
50–60	8
60–70	12
70–80	10
80–90	6

Step 1: Compute the cumulative frequency.

$$N = 4 + 8 + 12 + 10 + 6 = 40$$

Class Interval	Frequency (f_i)	Cumulative Frequency
40–50	4	4
50–60	8	12
60–70	12	24
70–80	10	34
80–90	6	40

Step 2: Compute the five-number summary.

- **Minimum**: Lower boundary of the first class, 40.
- **Maximum**: Upper boundary of the last class, 90.
- **Median (Q_2)**: Position $\frac{50}{100} \times 40 = 20$. Cumulative frequency exceeds 20 at 60–70:

$$Q_2 = 60 + \left(\frac{20 - 12}{12}\right) \times 10 = 60 + \left(\frac{8}{12}\right) \times 10 = 60 + 6.67 = 66.67$$

- Q_1: Position $\frac{25}{100} \times 40 = 10$. Cumulative frequency exceeds 10 at 50–60:

$$Q_1 = 50 + \left(\frac{10 - 4}{8}\right) \times 10 = 50 + \left(\frac{6}{8}\right) \times 10 = 50 + 7.5 = 57.5$$

- Q_3: Position $\frac{75}{100} \times 40 = 30$. Cumulative frequency exceeds 30 at 70–80:

$$Q_3 = 70 + \left(\frac{30 - 24}{10}\right) \times 10 = 70 + \left(\frac{6}{10}\right) \times 10 = 70 + 6 = 76$$

Five-Number Summary for Group 2:

$$\text{Minimum} = 40, \quad Q_1 = 57.5, \quad Q_2 = 66.67, \quad Q_3 = 76, \quad \text{Maximum} = 90$$

Comparison Using Boxplots

Side-by-Side Boxplot Description:

- **Group 1**: Box from 67.5 to 86.25, median at 76.67, whiskers from 50 to 100.
- **Group 2**: Box from 57.5 to 76, median at 66.67, whiskers from 40 to 90.

Observations:

- **Central Tendency**: Group 1 has a higher median (76.67 vs. 66.67), indicating better overall performance.
- **Spread**: Group 1's range (100 - 50 = 50) is larger than Group 2's (90 - 40 = 50), but Group 2 has a slightly smaller IQR (76 - 57.5 = 18.5 vs. 86.25 - 67.5 = 18.75), indicating slightly less variability in the middle 50%.
- **Skewness**: Group 1's median is closer to Q_1 (right skew), while Group 2's median is more centered (more symmetric).

3 Benefits of a Boxplot

Boxplots offer several advantages for data analysis and visualization, making them a valuable tool in statistics:

- **Easy Way to See the Distribution of Data**: Boxplots provide a clear visual summary of the data's spread, central tendency, and variability, all in one graph.
- **Tells About Skewness of Data**: The position of the median within the box and the lengths of the whiskers can indicate whether the data is skewed (e.g., if the median is closer to Q_1, the data is right-skewed).
- **Can Identify Outliers**: Data points outside the whiskers (beyond $Q_1 - 1.5 \times \text{IQR}$ or $Q_3 + 1.5 \times \text{IQR}$) are plotted as outliers, making them easy to spot.
- **Compare Two Categories of Data**: Side-by-side boxplots allow for quick comparison of distributions between two or more groups, as demonstrated with the test scores and cat weight examples.

Practical Notes

- **Educational Insights**: Comparing test scores of two classes or groups helps educators identify performance gaps and tailor teaching strategies (e.g., Group 2 may need additional support).
- **Skewness Analysis**: Boxplots reveal skewness, which can inform further analysis (e.g., if Group 1 is right-skewed, there may be a few high performers pulling up the median).
- **Veterinary Applications**: Comparing cat weights by sex can help veterinarians understand typical weight ranges and identify potential health issues (e.g., female cats are generally lighter, so a very heavy female cat might warrant further investigation).
- **Limitations**: For grouped data, the five-number summary is an approximation, and individual outliers cannot be identified. Side-by-side boxplots also don't show the full distribution shape (e.g., bimodality).

Covariance and Correlation

Covariance and correlation are fundamental statistical measures used to analyze the relationship between two variables. This document explores what covariance and correlation are, how they are calculated, their interpretations, and their limitations, with detailed examples and practical notes.

1 Covariance

What Problem Does Covariance Solve?

Covariance measures how two variables vary together

Covariance addresses the problem of understanding how two variables change in relation to each other. For example, if one variable increases, does the other variable also increase, decrease, or remain unchanged? Covariance quantifies this relationship by measuring the joint variability of two variables.

Consider the following illustrative examples (described textually, as images are not included):

- **Diagram 1**: Data points at $(-1,-1)$, $(0,0)$, and $(1,1)$ on a coordinate plane with axes labeled x (horizontal) and y (vertical). The points form a line with a positive slope, indicating that as x increases, y also increases. The variance of x-values is calculated as:

$$\frac{(-1)^2+0^2+1^2}{3}=\frac{1+0+1}{3}=\frac{2}{3}$$

  This suggests a positive covariance, as the variables move in the same direction.

- **Diagram 2**: Data points at $(-1,1)$, $(0,0)$, and $(1,-1)$. The points form a line with a negative slope, indicating that as x increases, y decreases. The variance of x-values is:

$$\frac{(-1)^2+0^2+(-1)^2}{3}=\frac{1+0+1}{3}=\frac{2}{3}$$

  This suggests a negative covariance, as the variables move in opposite directions.

Additional diagrams illustrate the concept further:

- **Second Diagram**: Several x marks plotted on a coordinate plane, generally increasing in both x and y directions, labeled "covariance +ve".
- **Third Diagram**: x marks plotted with high y values but no clear trend in x, labeled "covar ≈ 0".

What is Covariance and How is it Interpreted?

Covariance is a statistical measure that describes the degree to which two variables are linearly related. It measures how much two variables change together:

- If one variable increases, does the other also increase (positive covariance)?

- If one variable increases, does the other decrease (negative covariance)?
- If there's no consistent pattern, the covariance is zero (no linear relationship).

Interpretation:

- **Positive Covariance**: The variables tend to move in the same direction (e.g., as x increases, y increases).
- **Negative Covariance**: The variables tend to move in opposite directions (e.g., as x increases, y decreases).
- **Zero Covariance**: The variables are not linearly related.

How is Covariance Calculated?

The covariance between two variables X and Y can be calculated for both population and sample data:

Population	**Sample**
$\sigma_{xy} = \frac{\sum[(X-\mu_x)(Y-\mu_y)]}{N}$	$s_{xy} = \frac{\sum[(X-\bar{x})(Y-\bar{y})]}{(n-1)}$
X, Y : Values of X and Y in the population μ_x, μ_y : Population means of X and Y N : Total number of observations	X, Y : Values of X and Y in the sample $\bar{x}, \bar{y}$: Sample means of X and Y n : Total number of observations

Example 1: Positive Covariance (Experience vs. Salary)

Data: The following table shows years of experience (X) and salary (Y, in thousands) for 5 employees:

Exp (X)	**Salary** (Y)	$X - \bar{x}$	$Y - \bar{y}$	$(X - \bar{x})(Y - \bar{y})$
2	1	-6	-5	30
5	2	-3	-4	12
8	5	0	-1	0
12	12	4	6	24
13	10	5	4	20

Means:

$$\bar{x} = \frac{2+5+8+12+13}{5} = \frac{40}{5} = 8, \quad \bar{y} = \frac{1+2+5+12+10}{5} = \frac{30}{5} = 6$$

Calculation:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 30 + 12 + 0 + 24 + 20 = 86$$

$$n - 1 = 5 - 1 = 4$$

$$\text{Cov}(X, Y) = \frac{86}{4} = 21.5$$

Interpretation: The covariance is positive (21.5), indicating that as experience increases, salary tends to increase.

Scatter Plot Description: A scatter plot with the horizontal axis labeled "experience" (0 to 16) and the vertical axis labeled "salary" (0 to 14) shows points at approximately $(2, 1)$, $(5, 2)$, $(8, 5)$, $(12, 12)$, and $(13, 10)$. The points generally trend upward, confirming a positive relationship.

Example 2: Negative Covariance (Backlogs vs. Package)

Data: The following table shows the number of backlogs (X) and package (Y, in lakhs) for 5 students:

Backlogs (X)	**Package** (Y)	$X - \bar{x}$	$Y - \bar{y}$	$(X - \bar{x})(Y - \bar{y})$
2	10	-6	4	-24
5	12	-3	6	-18
8	5	0	-1	0
12	2	4	-4	-16
13	1	5	-5	-25

Means: $\bar{x} = 8$, $\bar{y} = 6$.

Calculation:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = -24 + (-18) + 0 + (-16) + (-25) = -83$$

$$n - 1 = 4$$

$$\text{Cov}(X, Y) = \frac{-83}{4} = -20.75$$

Interpretation: The covariance is negative (-20.75), indicating that as the number of backlogs increases, the package tends to decrease.

Scatter Plot Description: A scatter plot with the horizontal axis (0 to 16) and vertical axis (0 to 12) shows points at approximately $(2, 10)$, $(5, 12)$, $(8, 5)$, $(12, 2)$, and $(13, 1)$. Dashed lines at $x = 8$ and $y = 6$ divide the plot into quadrants, showing that the product of deviations has different signs, contributing to the negative covariance.

Example 3: Zero Covariance (Backlogs vs. Package with Constant Y)

Data: The following table shows backlogs (X) and a constant package (Y):

Backlogs (X)	**Package** (Y)	$X - \bar{x}$	$Y - \bar{y}$	$(X - \bar{x})(Y - \bar{y})$
2	10	-6	0	0
5	10	-3	0	0
8	10	0	0	0
12	10	4	0	0
13	10	5	0	0

Calculation: Since Y is constant, $Y - \bar{y} = 0$, so:

$$\text{Cov}(X, Y) \approx 0$$

Interpretation: There is no linear relationship between backlogs and package when the package is constant.

Graph Description: A coordinate system with the horizontal axis labeled with values 2, 5, 8, 12, and 13 shows a horizontal line of data points at $y = 10$, indicating zero covariance.

Disadvantages of Using Covariance

One major limitation of covariance is that it does not indicate the **strength** of the relationship between two variables. The magnitude of covariance is affected by the scale of the variables, making it difficult to compare relationships across datasets with different units or scales.

Covariance of a Variable with Itself

The formula for sample covariance is:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

If we consider the covariance of a variable with itself ($Y = X$), then $y_i = x_i$ and $\bar{y} = \bar{x}$. Substituting these into the formula:

$$\text{Cov}(X, X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

This is the formula for the **sample variance**. Therefore, the covariance of a variable with itself is equal to its variance.

2 Correlation

What Problem Does Correlation Solve?

Correlation addresses the problem of quantifying the strength and direction of the linear relationship between two variables. While covariance indicates whether variables move together, correlation provides a standardized measure that allows for comparison across different datasets.

Scatter Plots Illustrating Correlation (described textually):

- **Strong Positive Correlation**: Points closely clustered around an upward-sloping line.
- **Weak Positive Correlation**: Points loosely scattered around an upward-sloping line.
- **Strong Negative Correlation**: Points closely clustered around a downward-sloping line.
- **Weak Negative Correlation**: Points loosely scattered around a downward-sloping line.
- **Moderate Negative Correlation**: Points moderately clustered around a downward-sloping line.
- **No Correlation**: Points randomly scattered with no clear linear trend.

The question posed is: *Can we quantify this weak and strong relationship?* Correlation provides the answer by introducing a standardized measure.

What is Correlation?

Correlation refers to a statistical relationship between two or more variables, specifically measuring the degree to which two variables change together. It is quantified using the **correlation coefficient**, which ranges from -1 to 1:

- A correlation coefficient of 1 indicates a **perfect positive correlation**.
- A correlation coefficient of -1 indicates a **perfect negative correlation**.
- A correlation coefficient of 0 indicates **no correlation**.

The formula for the correlation coefficient (Pearson's r) is:

$$\text{Correlation} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

where $\text{Cov}(X, Y)$ is the covariance between X and Y, and σ_X and σ_Y are the standard deviations of X and Y, respectively.

Number Line Representation: A number line from -1 to 1, with points marked at -1, 0, and 1. An arrow from 0 to 1 indicates positive correlation, and an arrow from 0 to -1 indicates negative correlation. A mark between 0 and 1 represents a positive correlation value (e.g., 0.5).

Example: Correlation Between Experience and Salary

Using the data from "Example 1: Positive Covariance (Experience vs. Salary)":

Step 1: Use the Covariance: From the earlier calculation:

$$s_{xy} = 21.5$$

Step 2: Calculate Standard Deviations:

For X (Experience):

$$\sigma_X = \sqrt{\frac{\sum(X - \bar{x})^2}{n-1}} = \sqrt{\frac{(-6)^2 + (-3)^2 + 0^2 + 4^2 + 5^2}{4}} = \sqrt{\frac{36 + 9 + 0 + 16 + 25}{4}} = \sqrt{\frac{86}{4}} = \sqrt{21.5} \approx 4.64$$

For Y (Salary):

$$\sigma_Y = \sqrt{\frac{\sum(Y - \bar{y})^2}{n-1}} = \sqrt{\frac{(-5)^2 + (-4)^2 + (-1)^2 + 6^2 + 4^2}{4}} = \sqrt{\frac{25 + 16 + 1 + 36 + 16}{4}} = \sqrt{\frac{94}{4}} = \sqrt{23.5} \approx 4.8$$

Step 3: Compute Correlation:

$$\text{Correlation} = \frac{21.5}{4.64 \times 4.85} \approx \frac{21.5}{22.5} \approx 0.956$$

Interpretation: A correlation of 0.956 indicates a strong positive linear relationship between experience and salary.

Correlation and Causation

The phrase *"correlation does not imply causation"* means that an association between two variables does not necessarily mean one causes the other. Correlation only measures the strength and direction of a linear relationship, not causality.

Example: Suppose there is a positive correlation between the number of firefighters at a fire and the amount of damage caused by the fire. One might conclude that firefighters cause more damage. However, a third variable—the severity of the fire—explains the correlation: more severe fires require more firefighters and cause more damage.

Establishing causality requires additional evidence, such as experiments, randomized controlled trials, or well-designed observational studies.

Practical Notes and Key Points

- **Uses**:
  - Covariance shows the direction of the relationship between variables.
  - Correlation quantifies the strength and direction, making it easier to compare relationships across datasets.
- **HR Analytics**: The positive covariance (21.5) and strong correlation (0.956) between experience and salary suggest that more experienced employees tend to earn higher salaries, which can inform compensation strategies.
- **Education Insights**: The negative covariance (-20.75) between backlogs and package indicates that students with more backlogs tend to secure lower packages, highlighting the importance of academic performance.
- **Limitations**:
  - Covariance is affected by the scale of the variables, making comparisons difficult.
  - Correlation only measures linear relationships and does not prove causation.
- **Correlation vs. Causation**: Always consider potential confounding variables when interpreting correlations to avoid incorrect causal conclusions.
- **Next Steps**: Use regression analysis or experimental designs to explore causation and further understand relationships between variables.