Introduction Statistical Methods- Session-1

Statistics is the study of collecting, organizing, analyzing, and interpreting data. It helps to understand data and make decisions. There are two types:

1. **Descriptive Statistics** - Summarizes data (e.g., averages, graphs).
2. **Inferential Statistics** - Makes predictions using data samples.

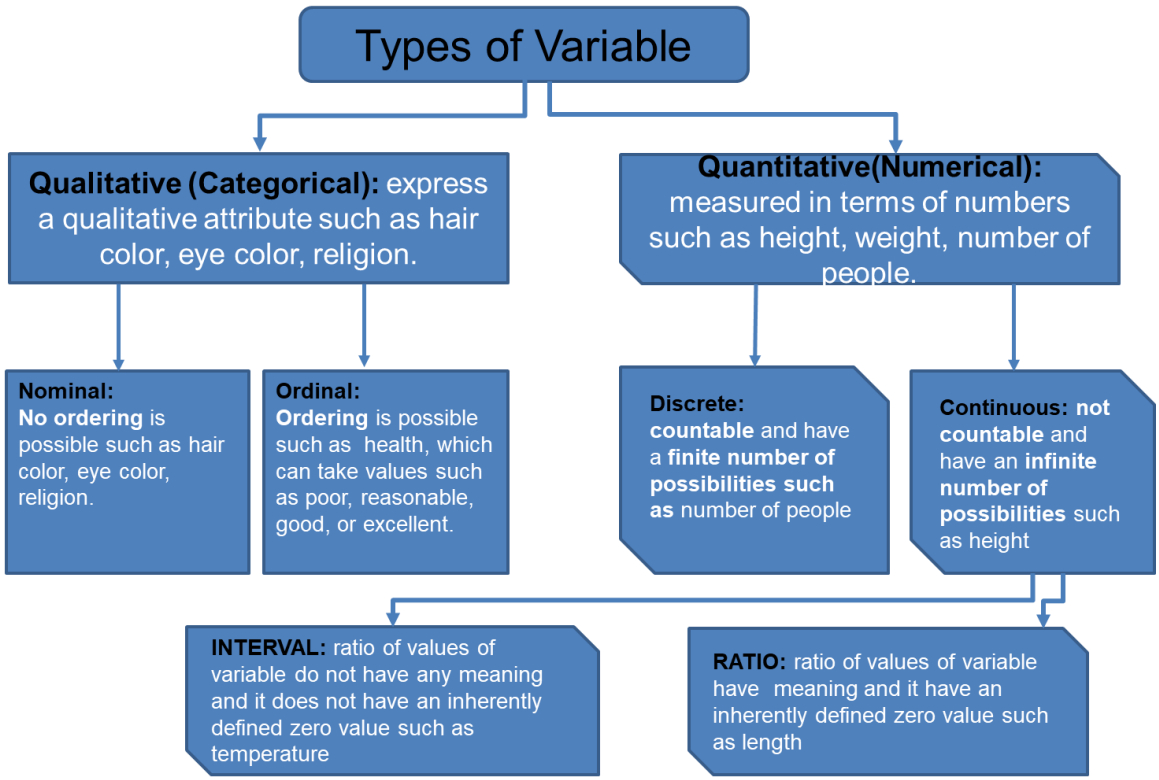
It is used in areas like business, health, and science.

The data into two main types:

1. **Categorical Data**: Non-numerical data, used to describe categories or groups (e.g., colors, types of animals).
2. **Numerical Data**: Data represented in numbers, divided into:
    - **Discrete**: Countable values (e.g., number of students).
    - **Continuous**: Measurable values with infinite possibilities (e.g., height, weight).
    -

There are the **four levels of measurement** in statistics, ranked from lowest to highest:

1. **Nominal (Lowest Level)**:
    - Categories with no order or ranking.
    - Examples: Gender, colors, names.
2. **Ordinal**:
    - Categories with a meaningful order but no consistent difference between values.
    - Examples: Rankings (e.g., 1st, 2nd, 3rd), satisfaction levels (e.g., good, better, best).
3. **Interval**:
    - Numerical data with consistent intervals, but no true zero point.
    - Examples: Temperature (Celsius/Fahrenheit), calendar years.
4. **Ratio (Highest Level)**:
    - Numerical data with a true zero point, allowing for meaningful ratios.
    - Examples: Height, weight, income, distance.



Measures of Central Tendency describe the center or average of a dataset.

The three main measures are:

1. **Mean (Average)**:
   - Sum of all values divided by the number of values.
   - Example: For 2,4,6 Mean = (2+4+6)/3=4
2. **Median**:
   - The middle value when data is arranged in ascending order.
   - Example: For 1,3,7, Median = 3.
3. **Mode**:
   - The value that appears most frequently in the dataset.
   - Example: For 1,2,2,3, Mode = 2.

These measures summarize data and give insights about its distribution.

Mean: Ungrouped Data

Suppose you define the time to get ready as the time (rounded to the nearest minute) from when you get out of bed to when you leave your home. You collect the times shown below for 10 consecutive work days:

Day	1	2	3	4	5	6	7	8	9	10
Time (min)	39	29	43	52	39	44	40	31	44	35

Calculations:

1. The **formula** for the mean:

$$\bar{X} = \frac{\text{Sum of the values}}{\text{Number of values}}$$

2. Substituting the values:

$$\bar{X} = \frac{39 + 29 + 43 + 52 + 39 + 44 + 40 + 31 + 44 + 35}{10}$$

3. The **sum** of the values is 396, and dividing by 10, the mean is:

$$\bar{X} = 39.6$$

Median:

- Arrange the times in ascending order: 29,31,35,39,39,40,43,44,44,52
- Median = Average of the 5th and 6th values.

Median: 39.5minutes

2. **Mode**:
   - Identify the most frequent value(s).

Mode: 39and 44 (both appear most frequently)

Mean: Grouped Scores

Hours Spent Watching TV	*Frequency (f)*	*fY*	*Percentage*	*C%*
1	104	104	31.3	31.3
2	130	260	39.2	70.5
3	98	294	29.5	100.0
Total	332	658	100.0	

$$\bar{Y} = \frac{\sum fY}{N} = \frac{658}{332} = 1.98$$

1. **Hours Spent Watching TV:**
   - **1 hour**: 104 children (31.3%)
   - **2 hours**: 130 children (39.2%)
   - **3 hours**: 98 children (29.5%)
2. **Total Children Surveyed**: 332
3. **Average (Mean) Time Spent Watching TV:**

$$\bar{Y} = \frac{\sum fY}{N} = \frac{658}{332} = 1.98 \text{ hours}$$

Properties of the Mean Explained with Examples:

1. **Mean Measures Stability**:
   - The mean is the most stable measure of central tendency because it takes into account all the values in the dataset. Every score contributes to the mean, making it less affected by small fluctuations in the data.
   - **Example**: If you have scores $10, 15, 20, 25$, the mean is:

$$\text{Mean} = \frac{10 + 15 + 20 + 25}{4} = 17.5$$

     Adding a new score 18 will only slightly change the mean:

$$\text{Mean} = \frac{10 + 15 + 20 + 25 + 18}{5} = 17.6$$

2. **Sensitive to Extreme Scores**:
   - The mean is greatly affected by outliers or extreme values, which can distort the central value.
   - **Example**: For scores $5, 7, 9, 11, 100$, the mean is:

$$\text{Mean} = \frac{5 + 7 + 9 + 11 + 100}{5} = 26.4$$

     Here, the extreme value 100 pulls the mean away from the center of most scores.

3. **Sum of Deviations from the Mean is Zero**:
   - When you calculate the deviations (difference between each score and the mean) and sum them, the result is always zero.
   - **Example**: For scores $4, 6, 8$, the mean is 6. Deviations are:

$$4 - 6 = -2,\ 6 - 6 = 0,\ 8 - 6 = 2$$

     Sum of deviations:

$$-2 + 0 + 2 = 0$$

4. **Applicable to Interval Level of Measurement**:
   - The mean is meaningful only for data measured at the interval or ratio level because it assumes equal intervals between data points.
   - **Example**: Scores on a test (interval data) like $70, 80, 90$ have a mean:

$$\text{Mean} = \frac{70 + 80 + 90}{3} = 80$$

     However, for categorical data like "red, green, blue," a mean cannot be calculated.

5. **May Not Be an Actual Score**:
    - The mean is often not a value that appears in the dataset.
    - **Example**: For scores $1, 2, 5$, the mean is:

$$\text{Mean} = \frac{1+2+5}{3} = 2.67$$

      Here, 2.67 is not an actual score in the dataset.

6. **Easy to Compute**:
    - The formula for the mean is simple and straightforward.
    - **Example**: If a student scored $85, 90, 88$ on three exams, the mean is:

$$\text{Mean} = \frac{85+90+88}{3} = 87.67$$

When to Use the Mean

The mean is the most suitable measure of central tendency in the following situations:

---

1. When Other Measures Like Standard Deviation, Coefficient of Variation, and Skewness Are to Be Computed

- The mean serves as the basis for calculating additional statistical measures like:
    - **Standard Deviation**: To measure data dispersion.
    - **Coefficient of Variation**: To compare variability relative to the mean.
    - **Skewness**: To analyze the asymmetry of the data distribution.

Example:

- Dataset: $10, 15, 20, 25, 30$
  - **Mean**:

$$\text{Mean} = \frac{10 + 15 + 20 + 25 + 30}{5} = 20$$

  - **Standard Deviation**: Measures how much data deviates from the mean.

$$SD = \sqrt{\frac{\sum(x - \bar{x})^2}{n}} = 7.07$$

  - **Coefficient of Variation**: Relative variability compared to the mean:

$$CV = \frac{SD}{\text{Mean}} \times 100 = \frac{7.07}{20} \times 100 = 35.35\%$$

  - **Skewness**: If the data is symmetric (e.g., $10, 15, 20, 25, 30$), skewness is 0.

These calculations rely on the mean, making it essential in statistical analysis.

Steps to Calculate the Standard Deviation (SD)

Let's calculate the **Standard Deviation (SD)** for the dataset:
$10, 15, 20, 25, 30$

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

The mean is the average of all values in the dataset.

$$\bar{x} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\bar{x} = \frac{10 + 15 + 20 + 25 + 30}{5} = \frac{100}{5} = 20$$

Step 2: Subtract the Mean from Each Value

For each data point (x_i), subtract the mean ($\bar{x}$):

$$x_i - \bar{x} = \text{Deviation from the mean}$$

x_i	Deviation ($x_i - \bar{x}$)
10	$10 - 20 = -10$
15	$15 - 20 = -5$
20	$20 - 20 = 0$
25	$25 - 20 = 5$
30	$30 - 20 = 10$

Step 3: Square Each Deviation

Square the deviations to eliminate negative values:

$$(x_i - \bar{x})^2$$

x_i	Deviation ($x_i - \bar{x}$)	Squared Deviation ($(x_i - \bar{x})^2$)
10	-10	$(-10)^2 = 100$
15	-5	$(-5)^2 = 25$
20	0	$0^2 = 0$
25	5	$5^2 = 25$
30	10	$10^2 = 100$

Step 4: Calculate the Variance

Variance is the average of the squared deviations.

$$\text{Variance} = \frac{\sum(x_i - \bar{x})^2}{n}$$

Substitute the values:

$$\text{Variance} = \frac{100 + 25 + 0 + 25 + 100}{5} = \frac{250}{5} = 50$$

Step 5: Calculate the Standard Deviation

The standard deviation is the square root of the variance:

$$SD = \sqrt{\text{Variance}}$$

$$SD = \sqrt{50} \approx 7.07$$

Final Answer

The **Standard Deviation (SD)** for the dataset $10, 15, 20, 25, 30$ is approximately:

$$SD \approx 7.07$$

2. When Sampling Stability Is Desired

- The mean is highly stable across samples taken from the same population, making it ideal for consistent comparisons.
- In contrast, the median and mode may vary significantly with sample fluctuations.

Example:

Imagine you are studying the average weight of students in a school. Samples from different classes are taken:

- **Sample 1**: Weights = $50, 55, 60$

$$\text{Mean} = \frac{50 + 55 + 60}{3} = 55$$

- **Sample 2**: Weights = $52, 54, 58$

$$\text{Mean} = \frac{52 + 54 + 58}{3} = 54.67$$

The mean remains consistent across samples, even if individual scores vary slightly. This stability allows reliable inference about the overall population.

When to Avoid the Mean

The mean should not be used if:

- There are extreme outliers in the dataset (e.g., income data where a billionaire skews the average).
- The data is nominal or ordinal (e.g., gender, ranks).

In such cases, median or mode may be more appropriate.

Mode

Key Properties of the Mode:

1. **Definition**: The mode is the data point that appears the most often in a dataset.
2. **Applicability**: The mode can be used for nominal, ordinal, interval, and ratio data.
3. **Uniqueness**: A dataset can have:
   - One mode (**unimodal**)
   - Two modes (**bimodal**)
   - More than two modes (**multimodal**)
   - No mode if no value repeats.

Steps to Identify the Mode:

1. **Collect the Data**: Ensure you have all the data points.
2. **Organize the Data**: Arrange the data in ascending order.
3. **Count Frequencies**: Determine how many times each value appears.
4. **Identify the Mode**: The value(s) with the highest frequency is/are the mode.

Example:

Consider the dataset:
$2, 3, 4, 4, 5, 6, 6, 6, 7, 8$

- **Step 1**: Organize the data:
  $2, 3, 4, 4, 5, 6, 6, 6, 7, 8$
- **Step 2**: Count frequencies:
  - $2 \rightarrow 1$ time
  - $3 \rightarrow 1$ time
  - $4 \rightarrow 2$ times
  - $5 \rightarrow 1$ time
  - $6 \rightarrow 3$ times
  - $7 \rightarrow 1$ time
  - $8 \rightarrow 1$ time
- **Step 3**: Identify the mode:
  The value 6 appears **3 times**, which is the highest frequency.
- **Mode**: 6

Median

Understanding the Median

The **median** is a measure of central tendency that identifies the middle value of a dataset when the numbers are arranged in ascending order. It divides the data into two equal halves—50% of the data points are below it, and 50% are above it.

Steps to Calculate the Median

1. **Arrange Data in Ascending Order**: Start by organizing the dataset from the smallest to the largest value.
2. **Identify the Number of Observations (n)**: Count how many data points there are in the dataset.
3. **Apply the Rules**:
   - **Rule 1**: If n is odd, the median is the middle value in the ordered dataset.
   - **Rule 2**: If n is even, the median is the average of the two middle values.

Examples

Example 1: Odd Number of Observations

Dataset: $12, 18, 24, 30, 36$

- **Step 1**: Arrange data in ascending order (already done).
- **Step 2**: Count $n = 5$ (odd number).
- **Step 3**: The middle value is the 3rd value, which is 24.

Median: 24

Example 2: Even Number of Observations

Dataset: $8, 14, 20, 26, 32, 38$

- **Step 1**: Arrange data in ascending order (already done).
- **Step 2**: Count $n = 6$ (even number).
- **Step 3**: The two middle values are the 3rd (20) and 4th (26) values.
- **Step 4**: Compute the average of these two values:

$$\text{Median} = \frac{20 + 26}{2} = 23$$

Median: 23

Key Characteristics of the Median

1. **Not Sensitive to Outliers**: Unlike the mean, the median is robust to extreme values.
2. **Applicable to Ordinal and Interval/Ratio Data**: It works well for data that can be ranked.
3. **Middle Value or 50th Percentile**: Half of the data lies below it, and half lies above.

This makes the median particularly useful when the data contains outliers or is skewed, as it provides a more accurate measure of central tendency in such cases.

Symmetrical and Asymmetrical Data in Central Tendency

Central tendency measures (mean, median, mode) can describe data effectively depending on the nature of the dataset—whether it is **symmetrical** or **asymmetrical**. Here's how the median and other measures are applied in these cases:

1. Symmetrical Data

In **symmetrical data**, the left and right sides of the distribution are mirror images. In such cases:

- **Mean, Median, and Mode** are generally equal or very close to each other.
- All three measures can be used to describe the central tendency.

Example of Symmetrical Data

Dataset: $10, 20, 30, 40, 50$

- **Mean**: $\frac{10+20+30+40+50}{5} = 30$
- **Median**: The middle value is 30.
- **Mode**: There is no repeated value, so no mode (or undefined).

Conclusion: For symmetrical data, the mean, median, and mode provide similar results, and any of them can be used as the central measure.

2. Asymmetrical (Skewed) Data

In **asymmetrical (skewed)** data, the distribution is not balanced, and one tail is longer than the other.

- **Median** is preferred as it is less affected by extreme values (outliers) compared to the mean.
- Mean can be distorted due to the skewness, making it a less accurate representation of central tendency.

Example of Asymmetrical Data

Dataset: $10, 20, 30, 40, 500$ (positive skew)

- **Mean:** $\frac{10+20+30+40+500}{5} = 120$
  The mean is heavily influenced by the extreme value 500.
- **Median:** Arrange data in ascending order: $10, 20, 30, 40, 500$. The middle value is 30.
- **Mode:** No repeated value, so no mode (or undefined).

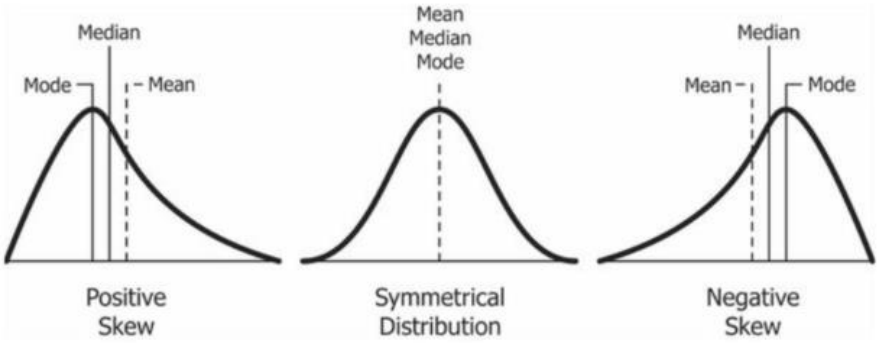
Conclusion: The **median (30)** is a better measure of central tendency than the mean (120) in this case because it is not affected by the extreme value 500.

When to Use Each Measure

1. **Symmetrical Data:**
   - Use **mean** when all data values are relevant.
   - **Median** or **mode** can also be used but offer no additional advantage.
2. **Asymmetrical Data:**
   - Use **median** as it is not influenced by outliers or skewness.
   - Avoid **mean** if there are extreme values in the dataset.
   - Use **mode** only if the most frequent value is important.

By choosing the appropriate central tendency measure based on the data's shape, we ensure that the summary statistic accurately represents the dataset.

Types of Distribution



Empirical Relationship: 3Median = Mode + 2Mean

This **empirical relationship** is used in statistics to approximate the **mode** for moderately skewed data where skewness is between -0.5 to 0.5. This formula works well in cases where the data is moderately skewed but doesn't have extreme asymmetry.

The formula is:

$$3\,\text{Median} = \text{Mode} + 2\,\text{Mean}$$

This relationship allows us to estimate the **mode** when the **mean** and **median** are known.

Example

Let's work with a dataset:

$$\{10, 12, 13, 15, 18, 20, 22, 22, 22, 25, 30\}$$

Step 1: Calculate the Mean

$$\text{Mean} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\text{Mean} = \frac{10 + 12 + 13 + 15 + 18 + 20 + 22 + 22 + 22 + 25 + 30}{11} = \frac{209}{11} \approx 19$$

Step 2: Calculate the Median

Arrange the data in ascending order:

$$\{10, 12, 13, 15, 18, 20, 22, 22, 22, 25, 30\}$$

Since there are 11 values (odd number), the **median** is the middle value:

$$\text{Median} = 20$$

Step 3: Approximate the Mode

The most frequently occurring value in the dataset is:

$$\text{Mode} = 22$$

Verify the Empirical Relationship

Now, use the formula:

$$3\,\text{Median} = \text{Mode} + 2\,\text{Mean}$$

Substitute the values:

$$3(20) = 22 + 2(19)$$

$$60 = 22 + 38$$

$$60 = 60$$

The formula holds true for this moderately skewed dataset.

When to Use This Relationship

1. **Moderately Skewed Data:** Use the formula when the skewness lies between -0.5 to 0.5.
2. **Estimation of Mode:** If the mode is difficult to calculate directly but the mean and median are known, this formula provides an estimate.

This empirical relationship bridges the mean, median, and mode, showing how they interrelate in skewed distributions.

The Range

The **range** is the simplest measure of dispersion, calculated as the difference between the largest and smallest values in a dataset.

Example

Dataset:

$$\{12, 18, 25, 30, 35, 40, 45\}$$

Step 1: Identify the maximum and minimum values

- Maximum value = 45
- Minimum value = 12

Step 2: Apply the formula for the range

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

$$\text{Range} = 45 - 12 = 33$$

Rule of Thumb (Simplified Approximation):

You can estimate skewness based on the relationship between the **mean** and **median**:

$$\text{Skewness Indicator} = \text{Mean} - \text{Median}$$

1. **If Mean > Median**: Positively skewed (right-skewed).
2. **If Mean < Median**: Negatively skewed (left-skewed).
3. **If Mean = Median**: Symmetrical distribution.

Would you like me to calculate the skewness for the given dataset?

What is Standard Deviation?

Standard deviation is a statistical measure that quantifies the amount of variation or dispersion in a set of data values. A low standard deviation means the data points are close to the mean (average), while a high standard deviation indicates the data points are spread out over a wider range of values.

The formula for standard deviation (for a population) is:

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

Where:

- σ is the standard deviation.
- x_i represents each value in the dataset.
- μ is the mean (average) of the dataset.
- N is the number of data points.

For a sample, the formula is slightly modified:

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

Where:

- s is the sample standard deviation.
- $\bar{x}$ is the sample mean.
- n is the number of sample data points.

Example

Imagine you have the following test scores of 5 students:

$70, 75, 80, 85, 90$

Step 1: Calculate the Mean (μ)

$$\mu = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\mu = \frac{70 + 75 + 80 + 85 + 90}{5} = 80$$

Step 2: Find Each Deviation from the Mean ($x_i - \mu$)

$$70 - 80 = -10,\ 75 - 80 = -5,\ 80 - 80 = 0,\ 85 - 80 = 5,\ 90 - 80 = 10$$

Step 3: Square Each Deviation

$$(-10)^2 = 100,\ (-5)^2 = 25,\ 0^2 = 0,\ 5^2 = 25,\ 10^2 = 100$$

Step 4: Find the Average of the Squared Deviations (Variance)

$$\text{Variance} = \frac{\text{Sum of squared deviations}}{\text{Number of values}}$$

$$\text{Variance} = \frac{100 + 25 + 0 + 25 + 100}{5} = \frac{250}{5} = 50$$

Step 5: Take the Square Root of the Variance

$$\sigma = \sqrt{50} \approx 7.07$$

Variance and Standard Deviation for a Sample

1. Variance Formula (Sample):

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

Where:

- s^2 is the sample variance.
- x_i are the data points.
- $\bar{x}$ is the sample mean (average).
- n is the number of data points in the sample.

2. Standard Deviation Formula (Sample):

$$s = \sqrt{s^2}$$

Where s is the standard deviation.

Example:

Suppose we have a sample of test scores:
$60, 70, 80, 90, 100$

Step 1: Calculate the Sample Mean ($\bar{x}$)

$$\bar{x} = \frac{\text{Sum of all values}}{\text{Number of values}}$$

$$\bar{x} = \frac{60 + 70 + 80 + 90 + 100}{5} = 80$$

Step 2: Find Each Deviation from the Mean ($x_i - \bar{x}$)

$$60 - 80 = -20,\ 70 - 80 = -10,\ 80 - 80 = 0,\ 90 - 80 = 10,\ 100 - 80 = 20$$

Step 3: Square Each Deviation

$$(-20)^2 = 400,\ (-10)^2 = 100,\ 0^2 = 0,\ 10^2 = 100,\ 20^2 = 400$$

Step 4: Calculate the Variance (s^2)

$$s^2 = \frac{\text{Sum of squared deviations}}{n - 1}$$

$$s^2 = \frac{400 + 100 + 0 + 100 + 400}{5 - 1} = \frac{1000}{4} = 250$$

Step 5: Calculate the Standard Deviation (s)

$$s = \sqrt{s^2} = \sqrt{250} \approx 15.81$$

What Are Degrees of Freedom?

Degrees of freedom (df) refer to the number of independent values or quantities in a statistical calculation that can vary while still estimating a parameter. It is essentially the number of values that are "free to vary" after certain constraints (like the mean) have been applied to the data.

How Does It Work?

When calculating sample statistics, such as variance or standard deviation, one degree of freedom is lost because we use the sample mean ($\bar{x}$) as a constraint. This adjustment compensates for the fact that a sample may not perfectly represent the population, making the estimate more accurate.

For a sample of size n, the degrees of freedom are:

$$df = n - 1$$

This adjustment is why the sample variance formula divides by $n - 1$ instead of n.

Example: Degrees of Freedom in Variance Calculation

Suppose you have a sample of data:

$$x = [4, 6, 8, 10]$$

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

$$\bar{x} = \frac{4 + 6 + 8 + 10}{4} = 7$$

Step 2: Degrees of Freedom:

The sample size is $n = 4$. Since the mean ($\bar{x}$) is calculated from the data and used in further calculations, one degree of freedom is lost. Thus:

$$df = n - 1 = 4 - 1 = 3$$

Step 3: Variance Calculation:

To calculate the variance, we compute the squared deviations from the mean, sum them, and divide by $df = 3$ (not 4):

1. Deviations from the mean:

$$4 - 7 = -3,\ 6 - 7 = -1,\ 8 - 7 = 1,\ 10 - 7 = 3$$

2. Squared deviations:

$$(-3)^2 = 9,\ (-1)^2 = 1,\ 1^2 = 1,\ 3^2 = 9$$

3. Sum of squared deviations:

$$9 + 1 + 1 + 9 = 20$$

4. Variance:

$$s^2 = \frac{\text{Sum of squared deviations}}{df} = \frac{20}{3} \approx 6.67$$

Step 4: Standard Deviation:

$$s = \sqrt{s^2} = \sqrt{6.67} \approx 2.58$$

Descriptive Statistics: Mean (M) and Standard Deviation (s)

Descriptive statistics allow us to summarize a dataset effectively. The **mean** (M) represents the central tendency (average value), and the **standard deviation** (s) quantifies how much the data points deviate from the mean. Together, these two values provide a concise and powerful description of a distribution.

Key Points

1. **The Mean (M):**
   - Describes the "center" of the data.
   - Tells us the typical value around which the data points cluster.
2. **The Standard Deviation (s):**
   - Measures the "spread" or variability of the data.
   - Indicates how far the values tend to deviate from the mean.
   - A larger s indicates a wider spread, while a smaller s suggests data points are closer to the mean.
3. **Reconstructing the Distribution:**
   - Knowing both M and s, we can visualize the distribution of the data, assuming it is roughly normal (bell-shaped).
   - Many data points (~68%) will fall within $M \pm s$, about 95% within $M \pm 2s$, and almost all within $M \pm 3s$.

Example

Imagine the test scores of a class:

$$70, 75, 80, 85, 90$$

Step 1: Calculate the Mean (M)

$$M = \frac{70 + 75 + 80 + 85 + 90}{5} = 80$$

Step 2: Calculate the Standard Deviation (s)

1. Deviations from the mean:
   $$70 - 80 = -10,\ 75 - 80 = -5,\ 80 - 80 = 0,\ 85 - 80 = 5,\ 90 - 80 = 10$$
2. Squared deviations:
   $$(-10)^2 = 100,\ (-5)^2 = 25,\ 0^2 = 0,\ 5^2 = 25,\ 10^2 = 100$$
3. Variance (s^2):
   $$s^2 = \frac{100 + 25 + 0 + 25 + 100}{5} = \frac{250}{5} = 50$$
4. Standard deviation (s):
   $$s = \sqrt{50} \approx 7.07$$

Step 3: Interpret the Results

- The mean $M = 80$: The average test score is 80.
- The standard deviation $s = 7.07$: Most test scores are within 7.07 points of the mean.

Reconstructing the Distribution

Using $M = 80$ and $s = 7.07$:

- Approximately 68% of the scores lie between $M - s = 80 - 7.07 = 72.93$ and $M + s = 80 + 7.07 = 87.07$.
- About 95% of the scores lie within $M \pm 2s = [65.86, 94.14]$.

Five-Number Summary of Data

The five-number summary is a way to describe a dataset by identifying five key values. These values provide a quick overview of the distribution, variability, and range of the data.

Five Components:

1. **Minimum**:
   - The smallest value in the dataset.
2. **Q1 (First Quartile)**:
   - The value below which 25% of the data falls (the 25th percentile).
3. **Median**:
   - The middle value of the dataset when arranged in ascending order (the 50th percentile).
   - If the dataset has an even number of values, it's the average of the two middle values.
4. **Q3 (Third Quartile)**:
   - The value below which 75% of the data falls (the 75th percentile).
5. **Maximum**:
   - The largest value in the dataset.

Steps to Calculate the Five-Number Summary:

1. **Arrange the Data**: Sort the dataset in ascending order.
2. **Identify the Minimum and Maximum**: The smallest and largest values.
3. **Find the Median**: Split the data into two halves. The middle value (or average of the two middle values) is the median.
4. **Calculate Q1 and Q3**:
   - $Q1$: The median of the lower half (values below the median).
   - $Q3$: The median of the upper half (values above the median).

Example:

Consider the dataset:

$$3, 7, 8, 5, 12, 14, 21, 13, 18$$

Step 1: Arrange the data in ascending order:

$$3, 5, 7, 8, 12, 13, 14, 18, 21$$

Step 2: Identify the minimum and maximum:

- **Minimum:** 3
- **Maximum:** 21

Step 3: Find the median:

- The dataset has 9 values (odd), so the middle value is the 5th:

$$\text{Median} = 12$$

Step 4: Calculate $Q1$ and $Q3$:

- **Lower half:** $3, 5, 7, 8$
  - Median of the lower half ($Q1$): $\frac{5+7}{2} = 6$
- **Upper half:** $13, 14, 18, 21$
  - Median of the upper half ($Q3$): $\frac{14+18}{2} = 16$

Five-Number Summary:

1. Minimum: 3
2. $Q1$: 6
3. Median: 12
4. $Q3$: 16
5. Maximum: 21

Box-and-Whisker Plot (Boxplot)

A **box-and-whisker plot**, or simply a boxplot, is a graphical representation of a dataset based on its five-number summary. It helps visualize the distribution, variability, and potential outliers in the data.

Key Components of a Boxplot

1. **Box:**
   - The box spans from $Q1$ (first quartile) to $Q3$ (third quartile), representing the middle 50% of the data (the interquartile range or **IQR**).
   - A line inside the box marks the **median.**
2. **Whiskers:**
   - The whiskers extend from the box to the minimum and maximum values **within 1.5 × IQR** from $Q1$ and $Q3$, respectively.
3. **Outliers:**
   - Data points beyond the whiskers (more than $1.5 \times \text{IQR}$ from $Q1$ or $Q3$) are plotted as individual points.

Steps to Create a Boxplot

1. **Calculate the Five-Number Summary:**
   - Minimum, $Q1$, Median, $Q3$, and Maximum.
2. **Determine the IQR:**

$$\text{IQR} = Q3 - Q1$$

3. **Identify Whisker Boundaries:**
   - Lower whisker: $Q1 - 1.5 \times \text{IQR}$
   - Upper whisker: $Q3 + 1.5 \times \text{IQR}$
4. **Identify Outliers:**
   - Any values outside the whisker boundaries.
5. **Plot the Box and Whiskers:**
   - Draw a box from $Q1$ to $Q3$ with a line at the median.
   - Extend whiskers to the smallest and largest values within the boundaries.
   - Plot outliers as individual points.

Example

Consider the dataset:

$$4, 7, 8, 9, 10, 15, 21, 22, 24$$

Step 1: Arrange the data in ascending order:

$$4, 7, 8, 9, 10, 15, 21, 22, 24$$

Step 2: Find the Five-Number Summary:

1. **Minimum**: 4
2. **Maximum**: 24
3. **Median**:
   - Middle value: 10
4. $Q1$ (lower half: $4, 7, 8, 9$):
   - Median of lower half: $\frac{7+8}{2} = 7.5$
5. $Q3$ (upper half: $15, 21, 22, 24$):
   - Median of upper half: $\frac{21+22}{2} = 21.5$

Step 3: Calculate the IQR:

$$\text{IQR} = Q3 - Q1 = 21.5 - 7.5 = 14$$

Step 4: Identify Whisker Boundaries:

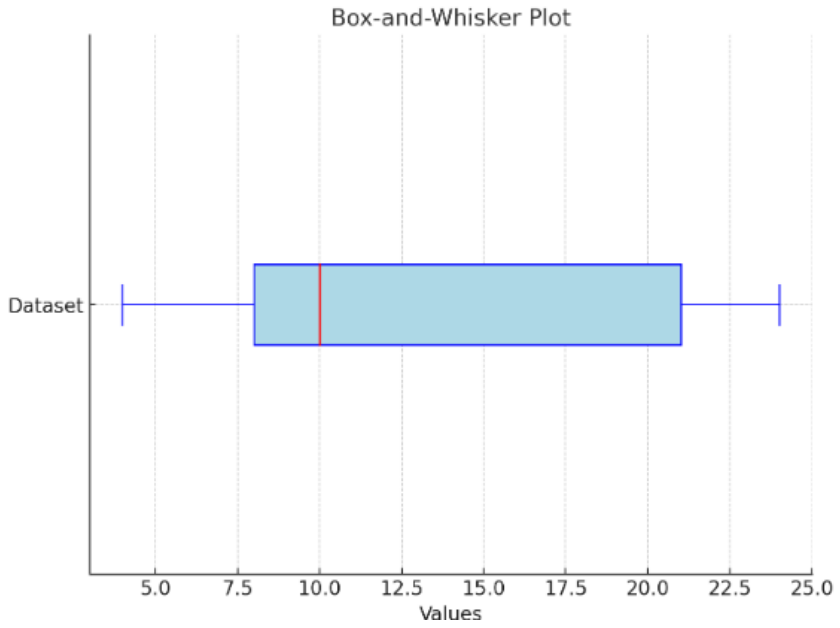
- **Lower boundary**: $Q1 - 1.5 \times \text{IQR} = 7.5 - (1.5 \times 14) = -13.5$
  (No values below this, so the lower whisker ends at 4).
- **Upper boundary**: $Q3 + 1.5 \times \text{IQR} = 21.5 + (1.5 \times 14) = 42.5$
  (No values above this, so the upper whisker ends at 24).

Step 5: Identify Outliers:

- No outliers, as all data points are within the whisker boundaries.

Boxplot Interpretation

- The **box** represents the interquartile range, from 7.5 to 21.5.
- The **whiskers** extend to the minimum (4) and maximum (24).
- The **median** (10) divides the box into two parts, showing the center of the data.

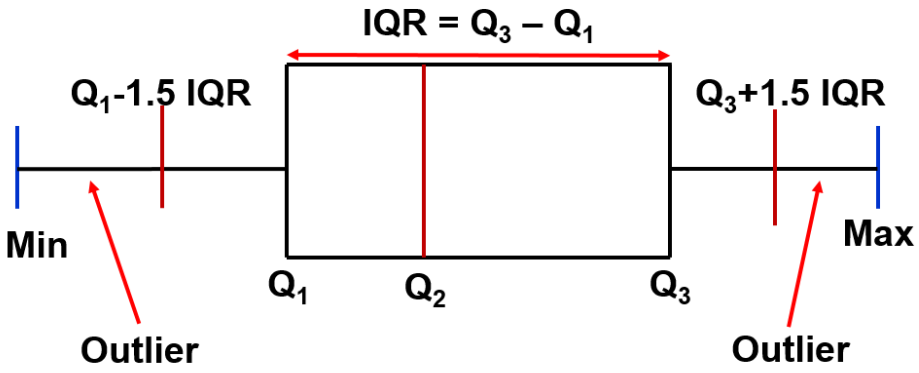


Explanation of the Plot:

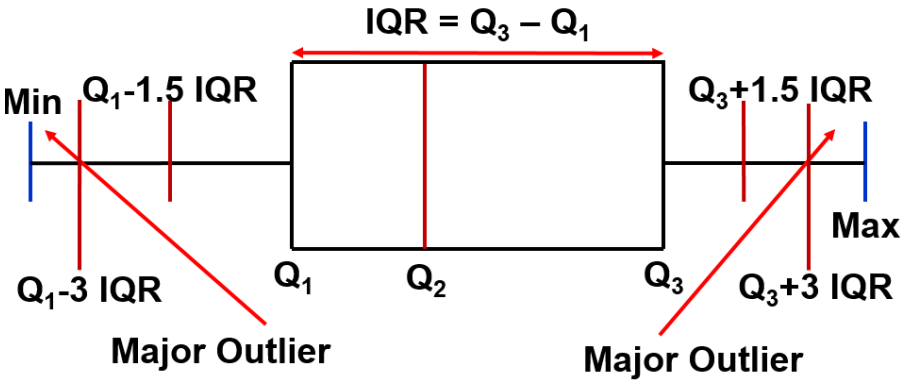
1. **Box:**
   - The box represents the interquartile range (IQR), spanning from $Q1 = 7.5$ to $Q3 = 21.5$.
   - The red line inside the box is the **median** (10).
2. **Whiskers:**
   - The whiskers extend to the **minimum** value (4) and the **maximum** value (24) within the 1.5 × IQR boundaries.
3. **No Outliers:**
   - There are no outliers, as all data points fall within the whisker boundaries.

This plot visually summarizes the data's spread, central tendency, and variability. Let me know if you'd like further clarification or an example with outliers! [›-]

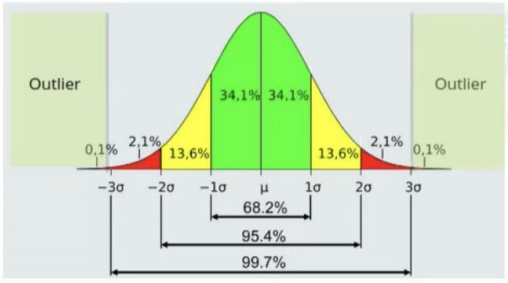
Box and Whisker plot

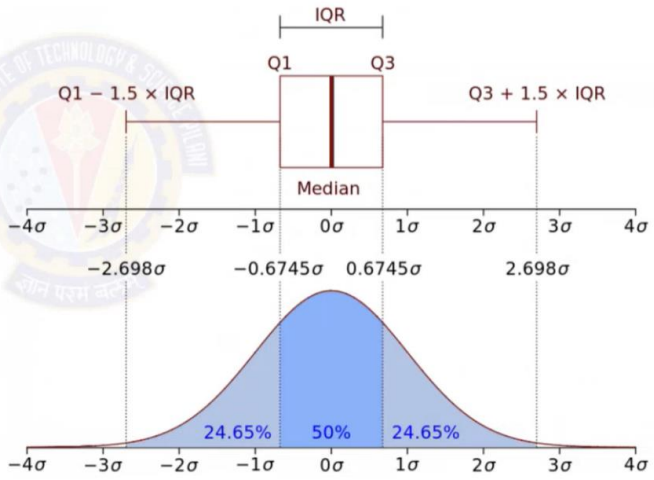


Box-and-Whisker plot



Outlier Detection





Practice Problems:

Q.1 A sample of 77 individuals working at a particular office was selected and the noise level (dBA) experienced by everyone is the following data:

55.3, 55.3, 55.3, 55.9, 55.9, 55.9, 55.9, 56.1, 56.1, 56.1, 56.1, 56.1, 56.1, 56.8, 56.8, 57.0, 57.0, 57.0, 57.8, 57.8, 57.8, 57.9, 57.9, 57.9, 58.8, 58.8, 58.8, 59.8, 59.8, 59.8, 62.2, 62.2, 63.8, 63.8, 63.8, 63.9, 63.9, 63.9, 64.7, 64.7, 64.7, 65.1, 65.1, 65.1, 65.3, 65.3, 65.3, 65.3, 67.4, 67.4, 67.4, 67.4, 68.7, 68.7, 68.7, 68.7, 69.0, 70.4, 70.4, 71.2, 71.2, 71.2, 73.0, 73.0, 73.1, 73.1, 74.6, 74.6, 74.6, 74.6, 79.3, 79.3, 79.3, 79.3, 83.0, 83.0, 83.0.

Find a) Arithmetic Mean, SD, variance, and IQR

b) Draw box and whisker plot

c) Comment on the outliers, if any.

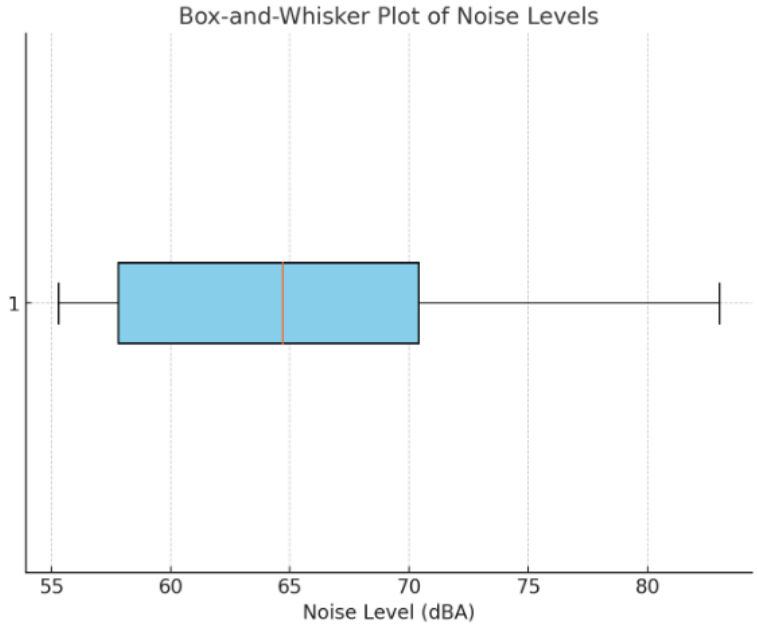
Part A: Calculate the Arithmetic Mean, Standard Deviation, Variance, and Interquartile Range (IQR)

1. Arithmetic Mean:
   - Formula:
     $$\text{Mean} = \frac{\sum x_i}{n}$$
     Where x_i are the noise level values, and n is the number of values.
   - We summed up all the noise levels (e.g., $55.3 + 55.3 + \ldots + 83.0$) and divided by the total count ($n = 77$).
   - **Result**: $\text{Mean} = 64.89\,\text{dBA}$.
2. Standard Deviation (SD):
   - Formula (for sample SD):
     $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$
     Where $\bar{x}$ is the mean.
   - Subtract each value from the mean, square the differences, sum them, divide by $n - 1$ (76), and take the square root.
   - **Result**: $s = 7.80\,\text{dBA}$.
3. Variance:
   - Formula:
     $$\text{Variance} = s^2$$
     The square of the standard deviation.
   - **Result**: Variance = $60.88\,\text{dBA}^2$.

4. **Interquartile Range (IQR):**
   - Formula:
     $$\text{IQR} = Q_3 - Q_1$$
     Where Q_3 is the 75th percentile and Q_1 is the 25th percentile.
   - From the data:
     - $Q_1 = 58.8$
     - $Q_3 = 71.4$
   - **Result:** $\text{IQR} = 71.4 - 58.8 = 12.60\,\text{dBA}$.

Part B: Create a Box-and-Whisker Plot

1. **Box Plot Elements:**
   - **Box:** Represents the interquartile range (IQR) from Q_1 to Q_3.
   - **Line in the Box:** Indicates the median (middle value).
   - **Whiskers:** Extend to the minimum and maximum values that are not outliers.
   - **Outliers:** Any points beyond $1.5 \times \text{IQR}$ from the quartiles.
2. **Plot Construction:**
   - The minimum value (55.3) and maximum value (83.0) fall within the whisker range. No extreme outliers were identified.
   - The plot visually displays the distribution of the noise levels.



Part C: Identify and Comment on Outliers

1. **Outlier Rule:**
   - Any value beyond the range:
     $$[Q_1 - 1.5 \times \text{IQR},\ Q_3 + 1.5 \times \text{IQR}]$$
     Using $\text{IQR} = 12.60$:
     - Lower limit: $Q_1 - 1.5 \times 12.60 = 58.8 - 18.9 = 39.9$
     - Upper limit: $Q_3 + 1.5 \times 12.60 = 71.4 + 18.9 = 90.3$
   - All data points (55.3 to 83.0) fall within this range.
2. **Observation:**
   - No outliers exist, as all values are within the acceptable range.

Practice Problems:

Q.2 The data given below is the total fat, in grams per serving, for a sample of 20 chicken sandwiches from fast-food chains.

7 8 4 5 16 20 20 24 19 30 23 30 25 19 29 29 30 30 40 56

a. Compute the mean, median, first quartile, and third quartile.

b. Compute the variance, standard deviation, range, interquartile range, Are there any outliers? Explain.

c. Are the data skewed? If so, how?

d. Based on the results of (a) through (c), what conclusions can you reach concerning the total fat of chicken sandwiches?

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

2. **Variance:**
$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$
3. **Standard Deviation (SD):**
$$s = \sqrt{\text{Variance}}$$
4. **Interquartile Range (IQR):**
$$\text{IQR} = Q3 - Q1$$
5. **Outliers:** Outliers are identified using:
$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR}$$
$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR}$$

Step A: Compute Mean, Median, Q1, and Q3

1. Mean

The **mean** represents the average fat content in the dataset.

$$\text{Mean} = \frac{\text{Sum of all data points}}{\text{Number of data points}}$$

For this dataset:

$$\text{Mean} = \frac{7+8+4+5+16+20+20+24+19+30+23+30+25+19+29+29+30+30+40+56}{20} = 23.2$$

- **Interpretation:** On average, chicken sandwiches have $23.2\,\text{grams}$ of fat per serving.

2. Median

The **median** is the middle value of the dataset when arranged in ascending order.

- When n is even ($n = 20$), the median is the average of the 10th and 11th values.
  Sorted data:
  $4, 5, 7, 8, 16, 19, 19, 20, 20, 23, 24, 25, 29, 29, 30, 30, 30, 30, 40, 56$
  The 10th and 11th values are 23 and 24.

$$\text{Median} = \frac{23+24}{2} = 23.5$$

- **Interpretation:** Half of the sandwiches have fat content below $23.5\,\text{grams}$, and the other half above it.

3. First Quartile (Q1)

The **first quartile (Q1)** is the median of the lower half of the sorted data.
Lower half:
$4, 5, 7, 8, 16, 19, 19, 20, 20, 23$
The median of this subset (5th and 6th values):

$$Q1 = \frac{16+19}{2} = 18.25$$

- **Interpretation:** 25% of sandwiches have fat content below $18.25\,\text{grams}$.

4. **Third Quartile (Q3)**

The **third quartile** (Q3) is the median of the upper half of the sorted data.

Upper half:

$24, 25, 29, 29, 30, 30, 30, 30, 40, 56$

The median of this subset (5th and 6th values):

$$Q3 = \frac{30 + 30}{2} = 30.0$$

- **Interpretation:** 75% of sandwiches have fat content below $30.0\,\text{grams}$, and 25% have more.

Step B: Compute Variance, Standard Deviation, Range, and IQR

1. **Range**

The **range** is the difference between the maximum and minimum values:

$$\text{Range} = \text{Maximum} - \text{Minimum} = 56 - 4 = 52$$

- **Interpretation:** The total fat content varies by up to $52\,\text{grams}$ between the least and most fatty sandwiches.

2. **Variance**

The **variance** measures how spread out the data is around the mean. Formula:

$$\text{Variance} = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Where x_i are the data points, $\bar{x}$ is the mean, and n is the sample size.

For this dataset:

$$\text{Variance} = 153.43\,\text{grams}^2$$

- **Interpretation:** On average, the squared deviation from the mean is $153.43\,\text{grams}^2$.

3. **Standard Deviation (SD)**

The **standard deviation** is the square root of the variance:

$$\text{SD} = \sqrt{\text{Variance}} = \sqrt{153.43} = 12.39\,\text{grams}$$

- **Interpretation:** Most sandwiches have fat content within $12.39\,\text{grams}$ of the mean.

4. Interquartile Range (IQR)

The IQR is the difference between Q3 and Q1:

$$\text{IQR} = Q3 - Q1 = 30.0 - 18.25 = 11.75$$

- **Interpretation:** The middle 50% of the data lies within an $11.75\ \text{gram}$ range.

5. Outliers

Outliers are identified as values beyond:

$$\text{Lower Bound} = Q1 - 1.5 \times \text{IQR} = 18.25 - 1.5 \times 11.75 = 0.625$$

$$\text{Upper Bound} = Q3 + 1.5 \times \text{IQR} = 30.0 + 1.5 \times 11.75 = 47.625$$

- Any value < 0.625 or > 47.625 is an outlier.
- The value 56 is the only outlier.

Step C: Check for Skewness

- **Mean** (23.2) < **Median** (23.5): Indicates a **left-skewed** distribution.
- **Interpretation:** The data have a few low values that pull the mean slightly below the median.

Step D: Conclusions

1. The total fat content of chicken sandwiches has a wide range ($52\ \text{grams}$), with a mean of $23.2\ \text{grams}$ and most values concentrated between 18.25 and $30.0\ \text{grams}$ (IQR).
2. There is one outlier ($56\ \text{grams}$) with significantly higher fat content than the rest.
3. The data are slightly left-skewed, meaning a few low-fat sandwiches influence the average.
4. The variation (SD = $12.39\ \text{grams}$) suggests substantial differences in fat content between sandwiches.

Rule of Thumb (Simplified Approximation):

You can estimate skewness based on the relationship between the **mean** and **median**:

$$\text{Skewness Indicator} = \text{Mean} - \text{Median}$$

1. **If Mean > Median**: Positively skewed (right-skewed).
2. **If Mean < Median**: Negatively skewed (left-skewed).
3. **If Mean = Median**: Symmetrical distribution.

Would you like me to calculate the skewness for the given dataset?

Practice Problems:

Q.3 The following data represent the battery life (in shots) for three pixel digital cameras:

300 180 85 170 380 460 260 35 380 120 110 240

List the Five-point summary.

To calculate the five-number summary, we need to identify the following:

1. **Minimum** (smallest value)
2. **First Quartile (Q1)** (25th percentile)
3. **Median** (50th percentile)
4. **Third Quartile (Q3)** (75th percentile)
5. **Maximum** (largest value)

Given data:

300, 180, 85, 170, 380, 460, 260, 35, 380, 120, 110, 240

Step 1: Arrange the data in ascending order:

35, 85, 110, 120, 170, 180, 240, 260, 300, 380, 380, 460

Step 2: Identify the five-number summary:

1. **Minimum**: The smallest value is **35**.
2. **Maximum**: The largest value is **460**.
3. **Median** (Q2): Since there are 12 data points, the median will be the average of the 6th and 7th values.
   - The 6th value is **180** and the 7th value is **240**.
   - Median = $\frac{180+240}{2} = 210$.
4. **First Quartile (Q1)**: This is the median of the lower half of the data (values before the median). The lower half is: **35, 85, 110, 120, 170, 180**.
   - The median of this subset is the average of the 3rd and 4th values: **110** and **120**.
   - Q1 = $\frac{110+120}{2} = 115$.
5. **Third Quartile (Q3)**: This is the median of the upper half of the data (values after the median). The upper half is: **240, 260, 300, 380, 380, 460**.
   - The median of this subset is the average of the 3rd and 4th values: **300** and **380**.
   - Q3 = $\frac{300+380}{2} = 340$.

Five-point summary:

- **Minimum** = 35
- **Q1** = 115
- **Median** = 210
- **Q3** = 340
- **Maximum** = 460

Practice Problems:

Q.4 For the data set below:

82	45	64	80	82	74	79	80	80	78	80	80	48	73	80	79	81	70	78	73

a. Obtain and interpret the quartiles.

b. Determine and interpret the interquartile range.

c. Find and interpret the five-number(point) summary.

d. Identify potential outliers, if any.

e. Construct and interpret a boxplot.

Step 1: **Sort the data in ascending order**

Sorted Data:

45, 48, 64, 70, 73, 73, 74, 78, 78, 79, 79, 80, 80, 80, 80, 80, 80, 81, 82, 82

a. **Obtain and interpret the quartiles.**

The **quartiles** divide the data into four equal parts, which are the 25th percentile (Q1), the 50th percentile (median), and the 75th percentile (Q3).

- **Median (Q2):**
  The median is the middle value of the data set. Since we have 20 values (an even number), the median is the average of the 10th and 11th values in the sorted data:
  10th value = 79, 11th value = 79.
  Median $Q2 = \frac{79+79}{2} = 79$.
- **First Quartile (Q1):**
  This is the median of the lower half of the data (values before the median). The lower half of the data is:
  45, 48, 64, 70, 73, 73, 74, 78, 78, 79.
  The median of this subset is the average of the 5th and 6th values:
  73 and 73.
  $Q1 = \frac{73+73}{2} = 73$.
- **Third Quartile (Q3):**
  This is the median of the upper half of the data (values after the median). The upper half of the data is:
  79, 80, 80, 80, 80, 80, 80, 81, 82, 82.
  The median of this subset is the average of the 5th and 6th values:
  80 and 80.
  $Q3 = \frac{80+80}{2} = 80$.

b. **Determine and interpret the interquartile range (IQR).**

The **interquartile range (IQR)** is the difference between the third quartile (Q3) and the first quartile (Q1):

$$IQR = Q3 - Q1 = 80 - 73 = 7$$

Interpretation:
The IQR tells us that the middle 50% of the data has a spread of 7 units, meaning the central half of the data points fall within 7 units of each other.

c. **Find and interpret the five-number summary.**

The **five-number summary** consists of:

1. **Minimum:** The smallest value in the data set, which is **45.**
2. **Q1 (First Quartile):** The 25th percentile, which is **73.**
3. **Median (Q2):** The middle value, which is **79.**
4. **Q3 (Third Quartile):** The 75th percentile, which is **80.**
5. **Maximum:** The largest value in the data set, which is **82.**

So, the five-number summary is:

- **Minimum** = 45
- **Q1** = 73
- **Median** = 79
- **Q3** = 80
- **Maximum** = 82

Interpretation:
The five-number summary provides a compact overview of the distribution of the data. The range between the minimum and maximum values is 37, and the middle 50% of the data is concentrated between 73 and 80, with the median (79) near the upper middle of this range.

d. **Identify potential outliers, if any.**

To detect potential outliers, we can use the **1.5 IQR rule**. According to this rule:

- Any data point below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ is considered a potential outlier.

First, calculate the threshold values for outliers:

- $Q1 - 1.5 \times IQR = 73 - 1.5 \times 7 = 73 - 10.5 = 62.5$
- $Q3 + 1.5 \times IQR = 80 + 1.5 \times 7 = 80 + 10.5 = 90.5$

Any values below **62.5** or above **90.5** would be potential outliers.

Data points:

- The smallest value is **45**, which is less than 62.5, so **45** is a potential outlier.
- The largest value is **82**, which is less than 90.5, so there are no upper outliers.

Potential outlier:
45 is a potential outlier (it is much lower than most of the other values).

e. **Construct and interpret a boxplot.**

A **boxplot** visually displays the five-number summary and identifies potential outliers.

Steps for constructing the boxplot:

- Draw a number line that covers the range of the data (at least from 45 to 82).
- Draw a box from **Q1 (73)** to **Q3 (80)**.
- Draw a line inside the box at the **median (79)**.
- Extend "whiskers" from the box to the **minimum (45)** and the **maximum (82)** values (unless there are outliers).
- Mark the potential outlier (**45**) as a point outside the whisker.

Interpretation:

- The boxplot will show a relatively symmetric distribution, but with a slight skew towards the higher values because Q3 (80) is a bit higher than Q1 (73).
- There is one potential outlier at 45, which is far from the rest of the data.

Summary:

- **Quartiles:**
  Q1 = 73, Median = 79, Q3 = 80.
- **IQR:**
  7 (interpreted as the spread of the middle 50% of the data).
- **Five-number summary:**
  Minimum = 45, Q1 = 73, Median = 79, Q3 = 80, Maximum = 82.
- **Potential outlier:**
  The value **45** is a potential outlier.
- **Boxplot interpretation:**
  The data is fairly symmetric, with a slight skew towards higher values and a potential outlier at 45.

Practice Problems:

Q.5 A bank branch located in a commercial place of a city has developed an improved process for serving customers during the noon-to-1:00 p.m. lunch period. The waiting time, in minutes (defined as the time the customer enters the line to when he or she reaches the teller window), of a sample of 15 customers during this hour is recorded over a period of one week. The results are: 4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.54, 3.20, 4.50, 6.10, 0.38, 5.12, 6.46, 6.19, 3.79.

Another branch, located in a residential area, is also concerned with the noon-to-1 p.m. lunch hour. The waiting time, in of a sample of 15 customers during this hour is recorded over a period of one week. The results are listed below: 9.66, 5.90, 8.02, 5.79, 8.73, 3.82, 8.01, 8.35, 10.49, 6.68, 5.64, 4.08, 6.17, 9.91, 5.47.

a. List the five-number summaries of the waiting times at the two bank branches.

b. Construct box-and-whisker plots and describe the shape of the distribution of each for the two bank branches.

c. What similarities and differences are there in the distributions of the waiting time at the two bank branches?

Data for Bank 1 (Commercial Location):

Waiting times (minutes):

4.21, 5.55, 3.02, 5.13, 4.77, 2.34, 3.54, 3.20, 4.50, 6.10, 0.38, 5.12, 6.46, 6.19, 3.79

Data for Bank 2 (Residential Location):

Waiting times (minutes):

9.66, 5.90, 8.02, 5.79, 8.73, 3.82, 8.01, 8.35, 10.49, 6.68, 5.64, 4.08, 6.17, 9.91, 5.47

Step 1: Calculate the Five-Number Summary

To calculate the five-number summary, we need to:

- **Sort the data** in ascending order.
- Find the **Minimum**, **Q1 (First Quartile)**, **Median (Q2)**, **Q3 (Third Quartile)**, and **Maximum**.

Bank 1 (Commercial Location)

Sorted Data:

0.38, 2.34, 3.02, 3.20, 3.54, 3.79, 4.21, 4.50, 4.77, 5.12, 5.13, 5.55, 6.10, 6.19, 6.46

- **Minimum** = 0.38
- **Maximum** = 6.46

To find the **quartiles**, we can use the following approach:

1. **Median (Q2):**
   The median is the middle value. With 15 data points, the median is the 8th value:
   Median (Q2) = 4.50
2. **Q1 (First Quartile):**
   The first quartile is the median of the lower half of the data (the first 7 values):
   0.38, 2.34, 3.02, 3.20, 3.54, 3.79, 4.21
   The median of these 7 values is the 4th value:
   Q1 = 3.20
3. **Q3 (Third Quartile):**
   The third quartile is the median of the upper half of the data (the last 7 values):
   4.50, 4.77, 5.12, 5.13, 5.55, 6.10, 6.19, 6.46
   The median of these 7 values is the 4th value:
   Q3 = 5.13

So, the five-number summary for **Bank 1** is:

- **Minimum** = 0.38
- **Q1** = 3.20
- **Median (Q2)** = 4.50
- **Q3** = 5.13
- **Maximum** = 6.46

Bank 2 (Residential Location)

Sorted Data:

3.82, 4.08, 5.47, 5.64, 5.79, 5.90, 6.17, 6.68, 8.01, 8.02, 8.35, 8.73, 9.66, 9.91, 10.49

- **Minimum** = 3.82
- **Maximum** = 10.49

Again, let's find the quartiles:

1. **Median (Q2):**
   The median is the 8th value (with 15 data points):
   Median (Q2) = 6.68
2. **Q1 (First Quartile):**
   The first quartile is the median of the lower half of the data (the first 7 values):
   3.82, 4.08, 5.47, 5.64, 5.79, 5.90, 6.17
   The median of these 7 values is the 4th value:
   Q1 = 5.64
3. **Q3 (Third Quartile):**
   The third quartile is the median of the upper half of the data (the last 7 values):
   6.68, 8.01, 8.02, 8.35, 8.73, 9.66, 9.91, 10.49
   The median of these 7 values is the 4th value:
   Q3 = 8.35

So, the five-number summary for **Bank 2** is:

- **Minimum** = 3.82
- **Q1** = 5.64
- **Median (Q2)** = 6.68
- **Q3** = 8.35
- **Maximum** = 10.49

Step 3: Similarities and Differences

Similarities:

- Both branches have a **central concentration of data** around the median, meaning that the majority of customers have waiting times around the middle value.

Differences:

- **Bank 1 (Commercial):**
  - The **waiting times are more concentrated** around the lower end (Q1 = 3.20, Median = 4.50), with a **smaller range** (0.38 to 6.46 minutes). This suggests that customers at this branch tend to wait less.
  - **Skewness:** The distribution appears **slightly skewed** to the right due to a few customers with higher waiting times (6.19, 6.46), but overall, most customers wait in the range of 2-5 minutes.
- **Bank 2 (Residential):**
  - The **waiting times are more spread out** (3.82 to 10.49 minutes), indicating **greater variability**. The **upper quartile (Q3 = 8.35)** and **median (6.68)** are higher than in Bank 1, indicating that many customers have longer waiting times.
  - **Skewness:** The distribution seems **slightly skewed** to the right as well, with a few longer waiting times, but the overall distribution indicates a broader spread compared to Bank 1.

Summary of Key Points:

1. **Bank 1** has **lower, more consistent waiting times**, with a **smaller range** (0.38 to 6.46 minutes).
2. **Bank 2** has **higher and more variable waiting times**, with a **larger range** (3.82 to 10.49 minutes).
3. Both distributions are **slightly right-skewed** but Bank 2's data has a **greater spread**.