Contents

[**1.** **What is descriptive statistics? How is it different from inferential statistics?** 2](#_Toc200740868)

[**2.** **Explain measures of central tendency. What are the commonly used ones?** 2](#_Toc200740869)

[**3.** **What is the difference between mean, median, and mode? When would you use each?** 3](#_Toc200740870)

[**4.** **Define measures of dispersion. Why are they important?** 3](#_Toc200740871)

[**5.** **Explain variance and standard deviation. How are they related?** 3](#_Toc200740872)

[**6.** **What is the interquartile range (IQR), and why is it useful?** 4](#_Toc200740873)

[**7.** **How do you detect outliers using descriptive statistics?** 4](#_Toc200740874)

[**8.** **What are skewness and kurtosis? What do they tell you about data?** 5](#_Toc200740875)

[**9.** **What is the difference between population and sample statistics?** 5](#_Toc200740876)

[**10.** **How do you interpret a box plot?** 6](#_Toc200740877)

[**11.** **Explain the significance of the coefficient of variation.** 6](#_Toc200740878)

[**12.** **What is a frequency distribution? How is it useful?** 6](#_Toc200740879)

[**13.** **What are percentiles and quartiles? How are they different?** 7](#_Toc200740880)

[**14.** **How do you summarize categorical data?** 7](#_Toc200740881)

[**15.** **What are some common visualization techniques used in descriptive statistics?** 8](#_Toc200740882)

[**16.** **Explain how you would summarize a dataset with both numerical and categorical variables.** 8](#_Toc200740883)

[**17.** **What is the difference between parametric and non-parametric statistics?** 8](#_Toc200740884)

[**18.** **How can descriptive statistics help in data cleaning and preprocessing?** 9](#_Toc200740885)

[**19.** **What assumptions underlie the use of mean and standard deviation?** 9](#_Toc200740886)

[**20.** **Explain the term ‘data distribution’. How do you assess if data is normally distributed?** 10](#_Toc200740887)

[**21.** **What is a histogram and what insights does it provide?** 10](#_Toc200740888)

[**22.** **How do you interpret a scatter plot in descriptive analysis?** 10](#_Toc200740889)

[**23.** **What is the difference between univariate, bivariate, and multivariate descriptive statistics?** 11](#_Toc200740890)

[**24.** **How do measures of dispersion change when data is transformed (e.g., scaled or shifted)?** 11](#_Toc200740891)

[**25.** **What are outliers and what impact can they have on descriptive statistics?** 11](#_Toc200740892)

## What is descriptive statistics? How is it different from inferential statistics?

**Answer:**

Descriptive statistics is a branch of statistics that focuses on summarizing and organizing data so it can be easily understood. It uses measures such as **mean, median, mode, standard deviation, variance, and range** to describe the central tendency, dispersion, and overall distribution of a dataset. Visualization tools like **histograms, bar charts, and box plots** also fall under descriptive statistics.

For example, if we collect the exam scores of a class, descriptive statistics would help us calculate the average score, the spread of scores, and display them in a visual format.

On the other hand, **inferential statistics** is used when we want to draw conclusions or make predictions about a larger population based on a sample of data. It involves techniques like **hypothesis testing, confidence intervals, regression analysis, and ANOVA**.

The key difference is:

* **Descriptive statistics** deals with **what the data shows**, and does not go beyond the dataset.
* **Inferential statistics** goes a step further to **make generalizations or inferences** about a population using data from a sample.

In short, descriptive statistics is about **summarizing data**, while inferential statistics is about **interpreting data** to make decisions or predictions.

## Explain measures of central tendency. What are the commonly used ones?

**Answer:**

Measures of central tendency are statistical metrics used to identify the central point or typical value in a dataset. They help summarize a dataset with a single representative value.

The three most commonly used measures are:

1. **Mean** – The arithmetic average, calculated by summing all values and dividing by the count. It's sensitive to outliers.
2. **Median** – The middle value when data is sorted. It’s robust to outliers and skewed distributions.
3. **Mode** – The most frequently occurring value(s) in the dataset. Useful for categorical data.

These measures provide different perspectives on data, and the choice depends on the distribution and nature of the dataset

## What is the difference between mean, median, and mode? When would you use each?

**Answer:**

The **mean** is the arithmetic average of all values; it considers every data point but is sensitive to outliers. The **median** is the middle value in a sorted dataset and is robust to outliers, making it suitable for skewed data. The **mode** is the most frequent value and is useful for categorical or multimodal data.

Use:

* **Mean** for symmetric distributions without outliers.
* **Median** for skewed data or when outliers are present.
* **Mode** for categorical data or to identify frequent occurrences.

## Define measures of dispersion. Why are they important?

**Answer:**

Measures of dispersion describe the spread or variability of data around the central tendency. Common ones include **range**, **variance**, **standard deviation**, and **interquartile range (IQR)**.

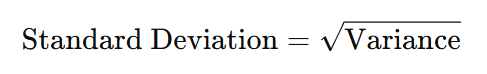
They are important because they show how much the data varies. Two datasets can have the same mean but very different spreads. Dispersion helps assess data consistency, detect outliers, and understand distribution patterns, which is crucial for accurate analysis and decision-making.

## Explain variance and standard deviation. How are they related?

**Answer:**

**Variance** measures the average squared deviation of each data point from the mean, showing how spread out the data is. **Standard deviation** is the square root of variance and represents the average deviation in the same units as the data.

They are directly related:



Standard deviation is more interpretable, while variance is useful in theoretical and intermediate calculations. Both are key to understanding data variability.

## What is the interquartile range (IQR), and why is it useful?

**Answer:**

The **interquartile range (IQR)** is the difference between the third quartile (Q3) and the first quartile (Q1), i.e.,

IQR=Q3−Q1\text{IQR} = Q3 - Q1IQR=Q3−Q1

It represents the middle 50% of the data.

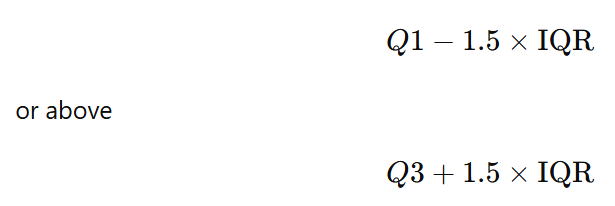
IQR is useful because it measures **data spread while ignoring outliers and extreme values**, making it a robust indicator of variability, especially for skewed distributions. It's also commonly used to detect outliers in box plots.

## How do you detect outliers using descriptive statistics?

**Answer:**

Outliers can be detected using the **interquartile range (IQR)** method or **z-scores**:

* **IQR Method**:  
  An observation is an outlier if it falls below



* **Z-Score Method**:  
  Calculate the z-score for each value (number of standard deviations from the mean). Values with |z| > 3 are typically considered outliers.

These methods help identify data points that deviate significantly from the overall pattern.

## What are skewness and kurtosis? What do they tell you about data?

**Answer:**

**Skewness** measures the **asymmetry** of a distribution.

* **Positive skew** means a longer right tail (more high values).
* **Negative skew** means a longer left tail (more low values).
* **Zero skew** indicates a symmetric distribution.

**Kurtosis** measures the **tailedness** or the concentration of data in the tails versus the center.

* **High kurtosis** (leptokurtic) indicates heavy tails and potential outliers.
* **Low kurtosis** (platykurtic) suggests light tails.
* **Normal kurtosis** (mesokurtic) is around 3.

Together, they help understand data shape and potential outliers.

## What is the difference between population and sample statistics?

**Answer:**

**Population statistics** describe characteristics of the entire dataset or group, while **sample statistics** describe characteristics of a subset drawn from that population.

For example:

* **Population mean (μ)** vs. **Sample mean (x̄)**
* **Population standard deviation (σ)** vs. **Sample standard deviation (s)**

Population parameters are fixed but often unknown. Sample statistics are used to estimate population parameters when studying the entire population is impractical.

## How do you interpret a box plot?

**Answer:**

A **box plot** visually summarizes the distribution of a dataset using five key statistics: **minimum**, **Q1 (first quartile)**, **median (Q2)**, **Q3 (third quartile)**, and **maximum**, along with any **outliers**.

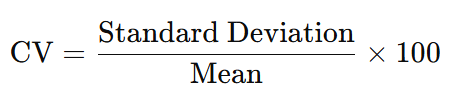
* The **box** spans from Q1 to Q3 (the IQR), showing the middle 50% of data.
* The **line inside the box** is the median.
* The **whiskers** extend to the smallest and largest values within 1.5 × IQR from Q1 and Q3.
* **Points outside the whiskers** are outliers.

It helps assess **central tendency**, **spread**, **symmetry**, and **outliers** at a glance.

## Explain the significance of the coefficient of variation.

**Answer:**

The **coefficient of variation (CV)** is the ratio of the standard deviation to the mean, usually expressed as a percentage:



It measures **relative variability**, allowing comparison of dispersion across datasets with different units or scales. A higher CV indicates greater variability relative to the mean. It's especially useful when comparing the consistency of different processes or investments.

## What is a frequency distribution? How is it useful?

**Answer:**

A **frequency distribution** is a summary that shows how often each value or range of values occurs in a dataset. It can be presented as a table or graph, like a histogram.

It is useful for:

* Understanding the **shape** of the data distribution
* Identifying **patterns**, **modes**, and **outliers**
* Simplifying large datasets for easier analysis and visualization

It provides a clear view of how data values are spread across intervals or categories.

## What are percentiles and quartiles? How are they different?

**Answer:**

**Percentiles** divide a dataset into 100 equal parts, indicating the value below which a certain percentage of observations fall. For example, the 90th percentile means 90% of the data is below that value.

**Quartiles** are a specific type of percentile that divide data into four equal parts:

* **Q1 (25th percentile)**
* **Q2 (50th percentile or median)**
* **Q3 (75th percentile)**

The key difference is **granularity**—percentiles split data into 100 parts, while quartiles split it into 4. Quartiles are commonly used for summary statistics and box plots.

## How do you summarize categorical data?

**Answer:**

Categorical data is summarized using **frequency counts** and **percentages**, often presented in **frequency tables**, **bar charts**, or **pie charts**.

Key methods include:

* **Mode**: Identifies the most common category.
* **Frequency distribution**: Shows how often each category occurs.
* **Proportions/percentages**: Help compare category sizes relative to the whole.

These summaries provide clear insights into category dominance, diversity, and distribution.

## What are some common visualization techniques used in descriptive statistics?

**Answer:**

Common visualization techniques include:

* **Histogram**: Shows the distribution of numerical data.
* **Box Plot**: Displays median, quartiles, and outliers.
* **Bar Chart**: Used for comparing frequencies of categorical data.
* **Pie Chart**: Represents proportions of categorical data.
* **Scatter Plot**: Shows relationships between two numerical variables.
* **Line Chart**: Useful for trends over time.

These visuals help quickly understand data patterns, distribution, and relationships.

## Explain how you would summarize a dataset with both numerical and categorical variables.

**Answer:**

To summarize a mixed dataset:

* For **numerical variables**, use **measures of central tendency** (mean, median), **dispersion** (standard deviation, IQR), and visualizations like **histograms** or **box plots**.
* For **categorical variables**, use **frequency counts**, **percentages**, and visualizations like **bar charts** or **pie charts**.

Additionally, cross-tabulations (e.g., **grouped summaries**) can show relationships between categorical and numerical variables, such as average income by gender or region. This provides a comprehensive view of the data.

## What is the difference between parametric and non-parametric statistics?

**Answer:**

**Parametric statistics** assume that the data follows a specific distribution, usually normal, and rely on fixed parameters like mean and standard deviation. Examples include **t-tests** and **ANOVA**.

**Non-parametric statistics** make no assumptions about the data distribution and are suitable for skewed, ordinal, or non-normal data. Examples include the **Mann-Whitney U test** and **Kruskal-Wallis test**.

The main difference is that parametric methods are more powerful when assumptions are met, while non-parametric methods are more flexible and robust to violations of those assumptions.

## How can descriptive statistics help in data cleaning and preprocessing?

**Answer:**

Descriptive statistics help identify **data quality issues** during cleaning and preprocessing by:

* Highlighting **missing values**, unusual **zeros**, or **outliers** using summary metrics.
* Revealing **inconsistencies** or **invalid entries** through frequency counts for categorical variables.
* Detecting **skewed distributions** or **unexpected ranges** in numerical data using measures like mean, median, and standard deviation.
* Guiding **normalization**, **transformation**, or **imputation** strategies based on data distribution.

This ensures the dataset is accurate, consistent, and ready for analysis or modeling.

## What assumptions underlie the use of mean and standard deviation?

**Answer:**

The use of **mean** and **standard deviation** assumes that:

1. **Data is continuous and interval or ratio-scaled**.
2. **The distribution is symmetric or approximately normal**—especially important for standard deviation.
3. **No extreme outliers**, as both metrics are sensitive to them.
4. **Observations are independent** of each other.

If these assumptions are violated, alternative measures like **median** and **IQR** may be more appropriate.

## Explain the term ‘data distribution’. How do you assess if data is normally distributed?

**Answer:**

**Data distribution** shows how values in a dataset are spread—whether they're mostly in the center, evenly spread out, or skewed to one side.

To check if data is **normally distributed** (bell-shaped), you can:

* **Plot a histogram** and look for a symmetrical shape.
* Check **skewness** (should be near 0) and **kurtosis** (should be near 3).

These checks help decide if you can use methods that assume normality.

## What is a histogram and what insights does it provide?

**Answer:**

A **histogram** is a graphical representation of the distribution of a numerical dataset. It groups data into **bins** (intervals) and uses bars to show the **frequency** of values within each bin.

It provides insights into:

* **Shape** of the distribution (normal, skewed, bimodal, etc.)
* **Spread** and **range** of the data
* **Central tendency**
* Presence of **outliers** or **gaps**

Histograms are useful for quickly understanding how data is distributed.

## How do you interpret a scatter plot in descriptive analysis?

**Answer:**

A **scatter plot** displays the relationship between two numerical variables using dots on a 2D graph.

Interpretation includes:

* **Direction**: Positive, negative, or no correlation.
* **Strength**: How closely points follow a line or curve.
* **Pattern**: Linear, non-linear, or clustered.
* **Outliers**: Points that fall far from the trend.

It helps identify trends, associations, and potential correlations in the data.

## What is the difference between univariate, bivariate, and multivariate descriptive statistics?

**Answer:**

* **Univariate** descriptive statistics analyze a single variable, focusing on its distribution, central tendency, and dispersion (e.g., mean, median, standard deviation).
* **Bivariate** statistics examine the relationship between two variables, using tools like scatter plots, correlation, and cross-tabulations.
* **Multivariate** statistics involve more than two variables, analyzing complex relationships using methods like multiple regression or principal component analysis.

The key difference is the **number of variables** analyzed and the **complexity of relationships** explored.

## How do measures of dispersion change when data is transformed (e.g., scaled or shifted)?

**Answer:**

When data is **shifted** (e.g., adding/subtracting a constant), measures of dispersion like **range**, **variance**, and **standard deviation** remain **unchanged**, because the spread doesn't change.

When data is **scaled** (e.g., multiplied/divided by a constant):

* **Range** and **standard deviation** scale by the **absolute value of the constant**.
* **Variance** scales by the **square of the constant**.

This shows dispersion is sensitive to scaling but not to shifting.

## What are outliers and what impact can they have on descriptive statistics?

**Answer:**

**Outliers** are data points that differ significantly from the rest of the dataset. They may result from variability, data entry errors, or rare events.

Outliers can:

* **Skew the mean**, making it unrepresentative.
* **Inflate the standard deviation and variance**, overstating data spread.
* **Distort visualizations** and mislead interpretation.

Detecting and handling outliers is crucial to ensure accurate and reliable statistical analysis.