**1.Summary of mean, median and mode.**

what is measures of central tendency?

A measure of central tendency is a single value that summarizes a set of data by identifying the middle or center of the data. The three most common measures of central tendency are the mean, median, and mode

Uses of measures of central tendency

* To identify outliers or extreme values that may significantly influence the data's overall distribution
* To provide a reference point for decision-making
* To assist in forecasting and predictive modeling

**Mean**:

In measures of central tendency, the **mean** (or **arithmetic mean**) is the average of a set of numbers. It is calculated by adding up all the values in the data set and then dividing by the number of values.

**Formula for Mean:**

Mean=N∑X​

Where:

* ∑X = Sum of all values in the data set
* N = Number of values in the data set

**Example:**

If we have the numbers: **5, 10, 15, 20, 25**, the mean is:

5+10+15+20+25​=75/5​=15

The mean is one of the three main measures of central tendency

The most commonly used measure of central tendency

* The average of a set of numbers
* Calculated by adding all the numbers in a set and dividing by the number of values in the set

**Median:**

The middle value of an ordered dataset. If the number of values is odd, the median is the middle number. If even, it is the average of the two middle numbers.

The **median** is one of the three main measures of **central tendency** (the others being **mean** and **mode**). It represents the middle value in an **ordered** data set, meaning half of the data points are **above** it and half are **below** it.

**How to Calculate the Median:**

1. **Arrange** the data in ascending order (smallest to largest).
2. **Find the middle value**:
   * If the number of observations (**n**) is **odd**, the median is the middle number.
   * If **n** is **even**, the median is the average of the two middle numbers.

### Example:

#### **Odd Number of Data Points**

Data: **3, 7, 9, 12, 15**

* The middle value is **9**, so the **median = 9**.

#### **Even Number of Data Points**

Data: **2, 5, 8, 11, 14, 17**

* The two middle numbers are **8** and **11**.
* Median = (8 + 11) / 2 = **9.5**.

### Why is the Median Important?

* It is **not affected by outliers** (unlike the mean).
* It is a **better measure** of central tendency when data is skewed.
* The middle number in an ordered set of numbers
* The value that divides a distribution into two equal parts

**Mode:**

The most frequently occurring value(s) in a dataset. A dataset may have one mode (unimodal), multiple modes (bimodal or multimodal), or no mode if all values occur with the same frequency.

In measures of central tendency, the **mode** is the value that appears most frequently in a data set. It represents the most common observation.

**Key Points About Mode:**

* A data set can have **one mode (unimodal)**, **more than one mode (bimodal or multimodal)**, or **no mode** if no value repeats.
* The mode is useful for categorical, ordinal, and numerical data.
* Unlike the mean and median, the mode is not affected by extreme values (outliers).

**Examples:**

1. **Unimodal Data Set**  
   Data: **2, 3, 4, 4, 5, 6**  
   Mode: **4** (because it appears most frequently)
2. **Bimodal Data Set**  
   Data: **1, 2, 2, 3, 4, 4, 5**  
   Modes: **2 and 4** (both appear twice)
3. **No Mode**  
   Data: **1, 2, 3, 4, 5**  
   Mode: **None** (no number repeats)

* The most frequently occurring value in a set of numbers

Each measure is useful in different situations:

* **Mean** is sensitive to extreme values (outliers).
* **Median** is useful when data has outliers or is skewed.
* **Mode** is helpful in identifying the most common value.

**2.Summary of Percentile**

A **percentile** is a measure of location in a dataset that indicates the value below which a given percentage of observations fall. It is commonly used in statistics to compare individual data points to a larger dataset.

**Understanding Percentiles**

The **p-th percentile** of a dataset is the value that separates the lowest **p%** of the data from the highest **(100 - p)%**.

For example:

The **25th percentile (Q1)** is the first quartile, meaning 25% of the data falls below it.

The **50th percentile (Q2)** is the median, meaning 50% of the data falls below it.

The **75th percentile (Q3)** is the third quartile, meaning 75% of the data falls below it.

The **90th percentile** means that 90% of the data falls below this value.

**Formula for Finding Percentiles**

For an **ordered dataset** with nnn values, the position of the **p-th percentile** is given by:

L=100p​×(n+1)

If **L is an integer**, the percentile value is the data point at that position.

If **L is not an integer**, round it to the nearest ranks and interpolate between the closest values.

**Example Calculation**

Suppose we have a dataset:  
10,20,30,40,50,60,70,80,90,100 (sorted data)

To find the **25th percentile**:

L=10025​×(10+1)=2.75

Since 2.75 is not an integer, we interpolate between the 2nd (20) and 3rd (30) values:

P25​=20+(0.75×(30−20))=27.5

Thus, the **25th percentile is 27.5**.

**Uses of Percentiles**

Standardized test scores (e.g., SAT, IQ tests)

Medical data analysis (e.g., growth charts for children)

Performance evaluations (e.g., employee rankings)

Economic studies (e.g., income percentiles)

**3.What is IQR in measures of location of data?**

In measures of location of data, "IQR percentile" refers to the range of data values encompassed by the interquartile range (IQR), which is calculated by taking the difference between the 75th percentile (the third quartile) and the 25th percentile (the first quartile); essentially, it represents the middle 50% of the data distribution

The **Interquartile Range (IQR)** is a measure of statistical dispersion that describes the spread of the middle 50% of a dataset. It is based on percentiles, specifically:

IQR=Q3−Q1IQR = Q\_3 - Q\_1IQR=Q3​−Q1​

where:

* **Q1 (1st quartile, 25th percentile)**: The value below which 25% of the data falls.
* **Q3 (3rd quartile, 75th percentile)**: The value below which 75% of the data falls.

Thus, the **IQR percentile range** is **between the 25th percentile (Q1) and the 75th percentile (Q3)**.

**Why 1.5 value is used in IQR percentile?**

The value **1.5** in the **Interquartile Range (IQR) method** for detecting outliers is based on statistical reasoning and empirical observations.

The interquartile (IQR) method of outlier detection uses 1.5 as its scale to detect outliers because it most closely follows Gaussian distribution. As a result, the method dictates that any data point that's 1.5 points below the lower bound quartile or above the upper bound quartile is an outlier.

### ****Understanding the IQR Method****

The **Interquartile Range (IQR)** is the difference between the third quartile (**Q3**) and the first quartile (**Q1**): IQR=Q3−Q1

To detect outliers, we define **lower** and **upper** bounds:

Lower Bound=Q1−1.5×IQR

Upper Bound=Q3+1.5×IQR

Any data point outside these bounds is considered an **outlier**.

Reasons why 1.5 value is used in IQR percentile

1. **Empirical Justification (Tukey’s Rule)**

* The 1.5 multiplier was suggested by **John Tukey**, who found that it effectively identifies unusually large or small values in a dataset.
* In a **normal distribution**, about **99.3%** of the data lies within **1.5 × IQR**. This means that only a small portion of the data (less than 1%) is flagged as outliers.

2. **Works Well for Many Distributions**

* The IQR method with **1.5 × IQR** is **not dependent on the assumption of normality**, making it robust for various distributions.
* Unlike standard deviation-based methods (e.g., Z-scores), which assume normality, the IQR method works even for skewed data.

3. **Balancing Sensitivity and Specificity**

* If the multiplier is **too small**, many normal points are incorrectly flagged as outliers (false positives).
* If the multiplier is **too large**, true outliers might be missed (false negatives).
* The value **1.5** balances these concerns and is a practical choice.

4.Finding outliers assignment = DONE

5.Preprocessing the chronic kidney disease raw dataset = DONE

6.Explaining Probability Density Function = DONE