Statistical Measures of Center: A Focus on Mean, Median, and Mode

Sushma B

[1]*dept. of computer science JSS college for arts, commerce and science*

---***---

Abstract - In statistics, systematic data collection, organization, analysis, interpretation, and presentation are the main topics of study. Understanding variability, seeing trends, and coming to wise conclusions in a variety of fields all depend on it.

In general, statistics can be separated into:

Descriptive statistics: These statistics include the description and summarisation of datasets.

Inferential statistics: Inferential statistics is the capacity to infer and forecast information about a population from a sample.

Through methods like regression analysis, probability, hypothesis testing, and data visualisation, statistics—which are extensively employed in fields including business, healthcare, social sciences, and science—assist in decision-making. It transforms unstructured data into meaningful knowledge that supports problem-solving and strategic planning.

Key Words: Mean, Median, Mode, Statistics, discrete maths, maths, Descriptive statistics

1.INTRODUCTION

The area of mathematics known as statistics is concerned with the gathering, organizing, analysing, interpreting, and presenting of data. It provides tools and techniques for understanding data and making informed decisions based on it.
The mean, median, and mode are statistical measures of central tendency. We identify the centre place of a set of data while characterising it. It is referred to as the central tendency measure. Every day we come across data. They are found in bank statements, newspapers, articles, and our phone and electricity bills. They are everywhere; the list is endless. This begs the question of whether concentrating on a limited number of data representations will allow us to uncover certain important elements of the data.

This is made feasible by averages or assessments of central tendency, such as the mean, median, and mode. Statistics is the process of converting data into information in order to comprehend demographic characteristics and spot trends. The examination of data is known as statistical analysis and manipulation. The mathematical process of statistics includes the following steps: collecting, analysing, interpreting, presenting, and organizing data. A variety of tools, strategies, and methods are used to help identify patterns, make predictions, and draw conclusions from data.

For example: calculating the mean of the grades each student in the class with a strength of 50 obtained.

Data can be described using statistics, and samples can be used to draw conclusions about populations. Especially when data include extreme values, median values (the 50th percentile) more accurately reflect the centre tendency of data samples than means (averages). Type I (identifying a difference between groups when none exists) and type II (failing to find an actual difference) errors are among the mistakes that arise from the use of inferential statistics when doing classical hypothesis testing. Spurious connections may result from confounding variables, which are ones that fluctuate with both the independent and dependent variables. It's common to overuse and overemphasise traditional hypothesis testing and reporting merely p-values.

.

Describe statistics:

Statistics can summarise and simplify large amounts of numerical data. With statistics, conclusions concerning data can be drawn.

By analysing data, statistics can calculate numerical estimates of "true" values.

Statistics cannot prove anything because estimates are usually stated in probabilistic terms (e.g., we are 95% sure...).
Statistics can't enhance bad data—"garbage in, garbage out."

Why make use of statistics?

Would you like to explain something that we only have a tiny sample of, like species, average grain size, stratigraphic range, or community composition? The "true" parameters must therefore be estimated using statistical methods.

Hidden patterns in data can be revealed by statistics that generally apparent, especially in studies with several variables.
When used properly, statistics may differentiate between the probable and the possible.

In general, statistics fall into two types.

Characteristic statistics: It comprises summarising and characterising the features of a dataset. Metrics such as variance, standard deviation, range, mean, median, and mode are included in this category. Descriptive statistics help organise and arrange data so that its main characteristics may be understood.
Example: a real-world use of descriptive statistics

Consider yourself employed by a smartphone sales company. Information regarding the battery life (in hours) of a new smartphone model has been collected. You need to use descriptive statistics in order to summarise and understand this dataset.
It is possible to analyse this data using descriptive statistics:

1.The Central Tendency:

Mean: In statistics, the mean, also called the average, is a metric of central tendency that provides a single value that represents the data's centre and summarises a dataset.

Median: The median in statistics is a measure of central tendency that indicates the middle value of the dataset when the values are arranged either descending or ascendingly.

Mode: In statistics, the mode is the value that appears most frequently in a dataset. Unlike the mean and median, which are measures of central tendency, the mode identifies the value or values that are most common or frequent in a dataset.

2. Measures of Dispersion:

Range: In statistics, range is a way to quantify how widely distributed or dispersed a dataset is. It is computed as the difference between a dataset's top and lowest values. A straightforward method for determining the degree of data variance is to use the range.

Variance: Variance is one statistical measure used to characterise the spread or dispersion of a set of data points. It calculates the degree to which values in a dataset differ from the mean, or average, of the dataset.

Standard Deviation: The standard deviation of a set of data points is a measure of statistical volatility or dispersion. The degree to which individual data points deviate from the dataset mean is indicated.

Data: In statistics, data classification is the process of dividing data into discrete groups or categories based on particular characteristics. This classification aids in the selection of appropriate statistical techniques for analysis and interpretation.
The type of data (qualitative or quantitative) and the degree of measurement are the two primary classification criteria.
1. Qualitative Data
2. Quantitative Data

1. **Qualitative Data:** Qualitative data, sometimes referred to as categorical data, are non-numerical data that depict attributes or categories. It characterises qualities, traits, or traits that fall into groups or categories but cannot be quantified.
   Types of Qualitative Data:

- **Nominal:** Data that depicts categories (such as colours, animal species, and gender) without a set order or ranking.
- **Ordinal:** Information that shows groups with a particular ranking or order but no discernible differences between them (e.g., educational level: high school, bachelor's, master's).

2. Quantitative Data:

Data that can be represented as numbers is referred to as quantitative data, and it is utilised to measure a dataset's characteristics. This kind of data is quantifiable and amenable to arithmetic operations such as division, multiplication, subtraction, and addition. Typically, measurements of height, weight, temperature, and age are made using quantitative data.

Types of Quantitative Data:

- **Discrete Data**: Data that accepts discrete, independent values, typically expressed as whole numbers (e.g., number of automobiles in a parking lot, number of pupils in a class).
- **Continuous Data**: Decimals and fractions are examples of data that can have any value within a specified range (e.g., height, weight, time, temperature).

What is Central Tendency?

A statistical metric known as "central tendency" designates one value as the central or representative point of a collection. It seeks to give a precise overview of the dataset by showing the areas where the majority of the data values are concentrated. In descriptive statistics, central tendency is a crucial idea. **Mean:** In statistics, the mean, sometimes referred to as the average, is a measure of central tendency that is calculated by dividing the total number of data points in a collection by their sum. It offers a single value that encapsulates the information and serves as a focal point for the values' tendency to cluster. The mean is frequently used to convey the "typical" or "central" value in a dataset in a variety of disciplines, such as economics, the social sciences, the natural sciences, and engineering.

Formula for Mean:

For a dataset with n values (x_1, x_2, x_3,…,x_n).

$$\text{Mean} = \frac{\text{Number of values}}{\text{Sum of all values}} = \frac{x_1+x_2+x_3+\dots+x_n}{n}$$

1. **Population Mean** (μ): This is used when you are calculating the mean for the entire population.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Where:

- μ is the population mean.
- N is the total number of data points in the population.
- x_i represents each individual data point.
- ∑ Indicates the sum of all data points

Illustrate: Consider a small population of 5 students and their scores on a test:

Population: 85, 90, 88, 92, 87

Formula for Population Mean:

$$\mu = \frac{\sum X}{N}$$

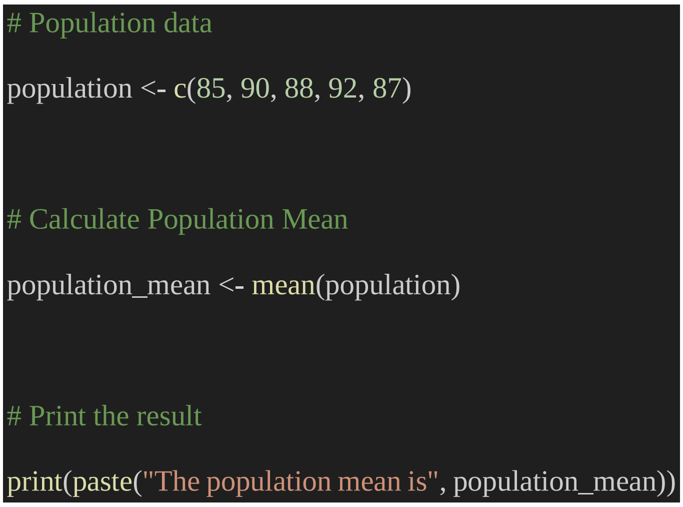
Where:

- μ is the population mean,
- x is each value in the population,
- N is the total number of values in the population.

Solution:

$$\mu = \frac{85+90+88+92+87}{5} = \frac{442}{5} = 88.4$$

Thus, the population mean is **88.4**.



```
# Population data

population <- c(85, 90, 88, 92, 87)


# Calculate Population Mean

population_mean <- mean(population)


# Print the result

print(paste("The population mean is", population_mean))
```

Result: The population mean is :88.4

Sample Mean ($\bar{x}$): This is used when you are calculating the mean for a sample taken from the population.

$$\bar{x} = \frac{\sum X}{n}$$

Where:

- x¯ is the sample mean,
- X is each value in the sample,
- n is the number of values in the sample.
- ∑ Indicates the sum of all data points

let’s take a random sample of 3 students' scores from the above population Sample 90, 88, 92

$$\bar{x} = \frac{\sum X}{n}$$

Where:

- x¯ is the sample mean,
- X is each value in the sample,
- n is the number of values in the sample.

Solution:

$$\bar{x} = \frac{90+88+92}{3} = \frac{270}{3} = 90$$

Thus, the sample mean is **90**.

```
# Sample data

sample <- c(90, 88, 92)


# Calculate Sample Mean

sample_mean <- mean(sample)


# Print the result

print(paste("The sample mean is", sample_mean))
```

Result: The sample mean is: 90

Median: The **median** is a measure of central tendency, which represents the middle value in a dataset when the values are arranged in ascending or descending order. The median is less impacted by outliers and distorted data than the mean.

Steps to Calculate the Median:

Arrange the Data: Sort the data points in ascending order of size.

1) Find the Middle:

- The median is the midpoint value when the number of data points is odd.
- The median is the mean of the two middle values if the number of data points is even.

Median for Odd Number of Values:

If the number of data points (nnn) is odd, the median is the middle value in the ordered data set.

Formula:

$$\text{Median} = X_{\left(\frac{n+1}{2}\right)}$$

Where:

- n is the number of data points.
- $X(2n+1/2)$ is the value in the $(n+1/2)$-th position when the data is sorted in ascending order.

Consider the following dataset: 7,1,9,5,37, 1, 9, 5, 37,1,9,5,3

Steps:

1. **Arrange the data in ascending order**: 1,3,5,7,91, 3, 5, 7, 91,3,5,7,9
2. **Count the number of data points**: n=5n = 5n=5 (odd number of values)
3. **Determine the median's location**: Utilize the calculation for an odd number of values' median.

$$\text{Median} = X_{\left(\frac{n+1}{2}\right)}$$

Here, n=5n = 5n=5, so the position of the median is:

$$\frac{5+1}{2}=\frac{6}{2}=3$$

The value at position three in the ordered dataset is called the median.

Median: The 3rd value in the ordered dataset {1,3,5,7,9}\{1, 3, 5, 7, 9\}{1,3,5,7,9} is **5**.

Thus, the **median** is **5**.

```
# Dataset

data <- c(7, 1, 9, 5, 3)


# Sort the data in ascending order

sorted_data <- sort(data)


# Calculate the median

median_value <- median(sorted_data)


# Print the result

print(paste("The median is", median_value))
```

Result: The median is 5

Median for Even Number of Values:

The median is the mean of the two middle values in the ordered data set if the number of data points (n) is even.

Formula:

$$\text{Median}=\frac{X_{\left(\frac{n}{2}\right)}+X_{\left(\frac{n}{2}+1\right)}}{2}$$

Where:

- n is the number of data points.
- $X_{(n/2)}$ is the value in the (n/2)-th position.
- $X_{(n/2+1)}$ is the value in the (n/2+1) th position.

Consider the following dataset: 12,15,18,20,25,3012, 15, 18, 20, 25, 3012,15,18,20,25,30

Steps:

1. **Arrange the data in ascending order** (already in order): 12,15,18,20,25,3012, 15, 18, 20, 25, 3012,15,18,20,25,30
2. **Count the number of data points**: n=6n = 6n=6 (even number of values)

Find the Two Middle Values: For an even number of data points, the median is the average of the two middle values. These two middle values are at positions n/2=3 and n2+1=4, which correspond to the values at the 3rd and 4th positions.

3rd position: **18**

4th position: **20**

Calculate the median: The average of the two middle values is known as the median.

$$\text{Median}=\frac{18+20}{2}=19$$

Thus, the **median** is **19**.

```
# Dataset

data <- c(12, 15, 18, 20, 25, 30)


# Sort the data in ascending order (if not already sorted)

sorted_data <- sort(data)


# Calculate the median (R's median function handles even cases automatically)

median_value <- median(sorted_data)


# Print the result

print(paste("The median is", median_value))
```

Result: The median is 19

Mode: The **mode** is a measure of central tendency that identifies the most frequently occurring value(s) in a dataset. Unlike the mean and median, which are based on the numerical values, the mode depends purely on frequency and can be used with both numerical and categorical data.

- **Unimodal**: A dataset with one mode (one most frequent value).
- **Bimodal**: A dataset with two modes (two most frequent values).
- **Multimodal**: A dataset with more than two modes.
- **No Mode**: A dataset where no value repeats.

Steps to Find the Mode:

1. **List all the values** in the dataset.
2. **Count the frequency** of each value.
3. **Identify the value(s)** with the highest frequency.

For Ungrouped Data (Simple Dataset):

To find the mode for an ungrouped dataset (e.g., raw data), the formula is based on counting the frequencies of each value:

- Determine which value occurs most frequently within the dataset.

Formula:

If $x_1, x_2, ..., x_n$ are the values of the dataset, the mode M is the value x_i that appears the most.

M=Value with the highest frequency

For example, for the dataset {3, 5, 7, 3, 9, 3, 4, 7, 5}, the mode is 3, as it appears 3 times, which is the most frequent.

For Grouped Data (Data in Intervals):

For **grouped data** (when the data is organized into classes or intervals), the mode is calculated using the **modal class** (the class with the highest frequency), and the following formula is used:

$$M = L + \left(\frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times h$$

Where:

- M is the mode,
- L is the lower boundary of the modal class,
- f_1 is the frequency of the modal class,
- f_0 is the frequency of the class before the modal class,
- f_2 is the frequency of the class after the modal class, h is the class width.

Example (Grouped Data):

Consider the following frequency distribution:

Class Interval	Frequency
0-10	3
10-20	7
20-30	12
30-40	5
40-50	2

The modal class is $20 - 30$ (since it has the highest frequency of 12).

Using the formula:

$$M = 20 + \left(\frac{12 - 7}{2 \times 12 - 7 - 5} \right) \times 10$$

After plugging in the values:

$$M = 20 + \left(\frac{5}{12} \right) \times 10$$

$$M = 20 + \frac{50}{12}$$

$$M \approx 20 + 4.17 = 24.17$$

Thus, the **mode** is approximately **24.17**.

```
# Custom function to calculate mode for ungrouped data
get_mode <- function(x) {

  uniq_x <- unique(x)  # Find unique values

  freq <- table(x)     # Count the frequency of each value

  mode_val <- uniq_x[freq == max(freq)]  # Find the value(s) with the highest frequency

  return(mode_val)

}



# Example ungrouped dataset

data_ungrouped <- c(3, 5, 7, 3, 9, 3, 4, 7, 5)



# Get the mode for ungrouped data

mode_ungrouped <- get_mode(data_ungrouped)



# Print the result

print(paste("The mode of the ungrouped dataset is", mode_ungrouped))
```

Result: The mode of the ungrouped dataset is 3

Conclusion:

In statistics, the mean, median, and mode are crucial indicators of central tendency, each providing a distinct perspective on the properties of a dataset:

1. **Mean** provides the arithmetic average, making it highly sensitive to extreme values (outliers) and best suited for datasets without significant skewness.
2. **Median** represents the middle value, offering a robust measure of central tendency for skewed or non-normal data distributions.
3. **Mode** identifies the most frequently occurring value, making it particularly useful for categorical data or datasets with repeating elements.

When combined, these metrics aid in describing and summarising the primary location of data, facilitating more in-depth pattern recognition and well-informed decision-making. The type of data and the particular objectives of the study will determine which metric is most suitable.

References:

1. "Statistics for Business and Economics" by Paul Newbold, William L. Karla, and Betty Thorne
2. "The Art of Statistics: How to Learn from Data" by David Spiegelhalter
3. "Elementary Statistics" by Mario F. Triola
4. "Naked Statistics: Stripping the Dread from the Data" by Charles Wheelan