Descriptive Statistics

1. What do you mean by descriptive statistics?

Descriptive statistics refers to a set of methods used to summarize and describe the main features of a dataset, such as its central tendency, variability, and distribution. These methods provide an overview of the data and help identify patterns and relationships.

2. What is descriptive statistics?

Descriptive statistics are methods used to summarize and describe the main features of a dataset. Examples include measures of central tendency, such as mean, median, and mode, which provide information about the typical value in the dataset. Measures of variability, such as range, variance, and standard deviation, describe the spread or dispersion of the data. Descriptive statistics can also include graphical methods, including histograms, box plots, and scatter plots, to visually represent the data.

3. What are the four types of descriptive statistics?

The four types of descriptive statistics are:

- **Measures of central tendency**
- **Measures of variability**
- **Standards of relative position**
- **Graphical methods**

Measures of central tendency describe the typical value in the dataset and include mean, median, and mode. Measures of variability represent the spread or dispersion of the data and include range, variance, and standard deviation. Measures of relative position describe the location of a specific value within the dataset, such as percentiles. Graphical methods use charts, histograms, and other visual representations to display data.

4. What is the main purpose of descriptive statistics?

The primary objective of descriptive statistics is to effectively summarize and describe the main features of a dataset, providing an overview of the data and helping to identify patterns and relationships within it. Descriptive statistics provide a useful starting point for analyzing data, as they can help to identify outliers, summarize key characteristics of the data, and inform the selection of appropriate statistical methods for further analysis. They are commonly used in multiple fields, including social sciences, business, and healthcare.

Measures of Centrality

Central Tendency / Location

- Measures of location are designed to provide the analyst with some quantitative values of where the center, or some other location, of data is located.
- We need one value to represent the data.

Measures of Centrality

The following measures are of primary importance:

1. **Mean**
2. **Median**
3. **Mode**

The **mean**, also sometimes referred to as the **arithmetic mean**, is the most useful and most commonly used measure of centrality. The **median** is the second most used, and the **mode** is the least used measure of centrality.

Measures of Centrality – Mean

Population Mean

The population mean is denoted by the Greek letter **μ** (read as *meu*), for a finite population with **N** equally likely values.

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

Sample Mean "Sample Average"

The sample mean is the average value of all observations in the data set. Usually, these data are a sample of observations that have been selected from some larger population of observations.

It is denoted by **x̄** (read as *x bar*), for **n** observations.

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example 1:

The eight observations are:

$x_1 = 12.6$, $x_2 = 12.9$, $x_3 = 13.4$, $x_4 = 12.3$, $x_5 = 13.6$, $x_6 = 13.5$, $x_7 = 12.6$, and $x_8 = 13.1$.

The sample mean is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{8} x_i}{8}$$

$$= \frac{12.6 + 12.9 + \cdots + 13.1}{8} = \frac{104}{8} = 13.0$$

Measures of Centrality – Median

Here is a rule for finding the median:

1. Arrange all observations in order of size, from smallest to largest.
2. Find the location (rank) of the median of a data set of size **n**.
3. Find the value of the observation corresponding to the rank of the median found in step 2, i.e.,
   - **If n is odd:** (x_{(n+1)/2})
   - **If n is even:** ((x*{n/2}* + *x*{(n/2)+1}) / 2)

$$\text{Rank} = \begin{cases} (n+1)/2 & \text{if } n \text{ odd} \\ n/2 \quad \text{and} \quad n/2+1 & \text{if } n \text{ even} \end{cases}$$

Example 1:

The nine observations are:

$x_1 = 12.6, x_2 = 12.9, x_3 = 13.4, x_4 = 12.3, x_5 = 13.6, x_6 = 13.5, x_7 = 12.6, x_8 = 13.1$ and $x_9 = 13.2$.

Arrange all observations in order of size $n = 9$:

Order	**12.3**	**12.6**	**12.6**	**12.9**	**13.1**	**13.2**	**13.4**	**13.5**	**13.6**
Ranks	1	2	3	4	5	6	7	8	9

Measures of Centrality – Mode (1/4)

The mode of a data set is the value that occurs most frequently. There may be no mode, or conversely, there may be multiple modes.

- If all values are unique without repetition, then there is **no mode**.
- If more than one value has the same highest frequency, then there are **multiple modes**.

Example 1:

Find the mode for the following data set

3, 8, 5, 6, 10, 17, 19, 20, 3, 2, 11

In the data set of this example, each value occurs once except the value 3, which occurs twice. Thus, *the mode* = 3.

Measures of Variability

1. Standard Deviation (SD)

Standard deviation (SD) is a measure of the dispersion or spread of a dataset. It quantifies how much individual data points deviate from the mean.

- A low SD indicates that the data points are close to the mean.
- A high SD indicates that the data points are more spread out.

The formula for the population standard deviation is:
σ = sqrt(Σ (xi - μ)^2 / N)

For a sample, the formula is:
s = sqrt(Σ (xi - x̄)^2 / (n-1))

2. Variance

Variance measures the average squared deviation from the mean. It is the square of the standard deviation.

- Population variance: σ² = Σ (xi - μ)^2 / N
- Sample variance: s² = Σ (xi - x̄)^2 / (n-1)

3. Range

The range is the simplest measure of dispersion. It is the difference between the maximum and minimum values in a dataset.

Range = Max - Min

4. Quartiles and Interquartile Range (IQR)

Quartiles divide the data into four equal parts:

- **Q1 (First Quartile):** 25% of the data falls below this value.
- **Q2 (Second Quartile or Median):** 50% of the data falls below this value.
- **Q3 (Third Quartile):** 75% of the data falls below this value.

Interquartile Range (IQR)

The interquartile range (IQR) measures the spread of the middle 50% of the data:
IQR = Q3 - Q1

IQR is useful for detecting outliers because it ignores extreme values and focuses on the central portion of the dataset.