Statistics in Data science

Data science involve the analysis and interpretation of complex datasets to extract valuable insights and support decision-making. Statistics plays a crucial role in data science, providing the foundation for various methods and techniques used in the field.

First topic

Descriptive Statistics (normal distribution)

1- **Descriptive Statistics**

Descriptive Statistics -> Descriptive statistics are a set of techniques used to summarize and describe the main features of a dataset .

Descriptive statistics divided into :

1- central tendency -> **Mean, Median, Mode**

2- measures of dispersion" variability" -> **Range, Variance, Standard Deviation**

1. central tendency -> aim to identify a representative or central value around which the data points cluster. They provide a single value that summarizes the central location of the data.

   1. Mean -> the sum of all values divided by the number of observations. It represents the central point of a dataset.

   2. Median -> The middle value of a dataset when arranged in ascending or descending order. It is less sensitive to extreme values than the mean.

   3. Mode -> The value or values that appear most frequently in a dataset.

2. measures of dispersion” variability”=>quantify the spread, variability, or extent to which data points deviate from the central tendency. They provide information about how "spread out" the values are.

1. **Range ->** The difference between the maximum and minimum values in a dataset.

2. **Variance ->** A measure of how spread ou^{2}t the values in a dataset are from the mean “the unit is the square of original unit => cm “ so the variance difficult to interpret .

3. **Standard Deviation ->** The square root of the variance. It provides a more
ic mean“ .

Population

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

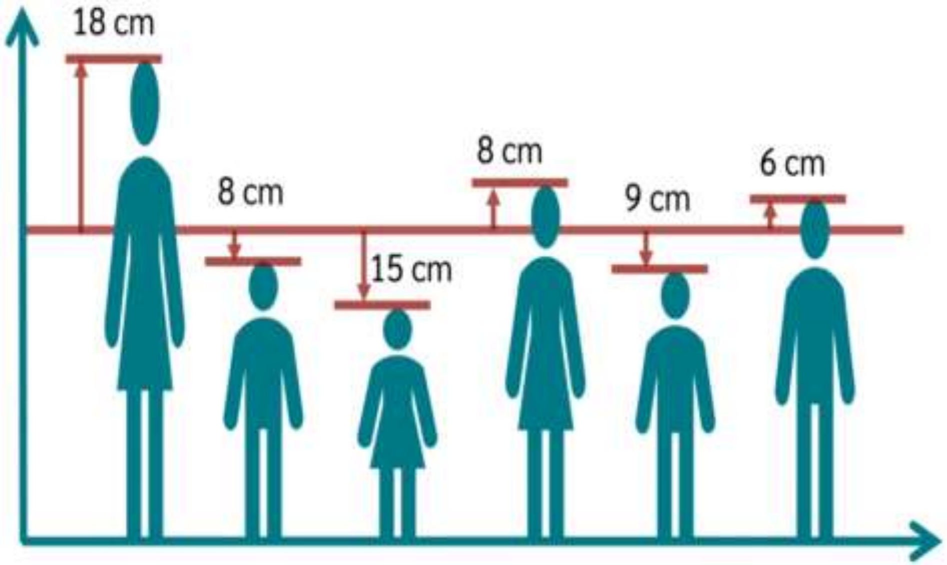
Sample

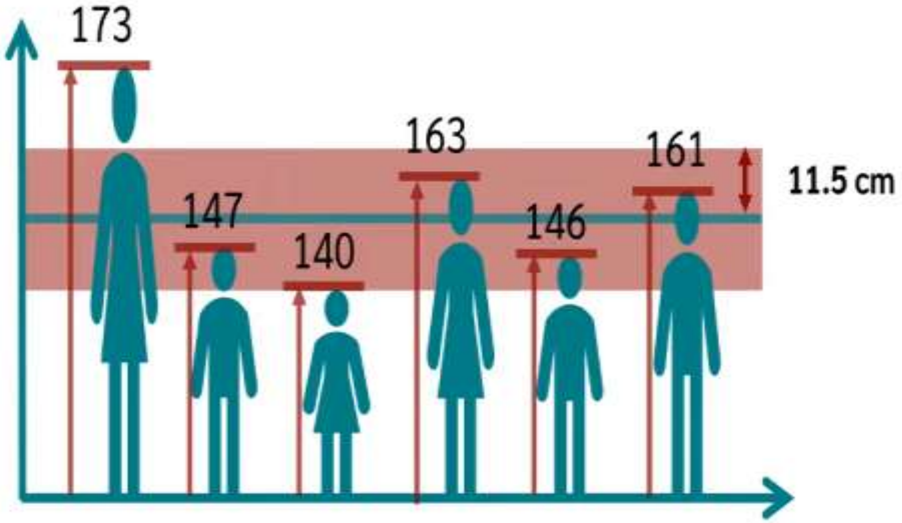
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Dispersion measures provide information about the variability or spread of values in a dataset.

Ex:

we want to know how much the **persons deviate** from the **mean value on average**.





The result is a standard deviation of **11.5 cm**.

$$= \sqrt{\frac{(173-155)^2+(147-155)^2+\cdots+(161-155)^2}{6}} = 11.5$$

central tendency vs measures of dispersion

central tendency => describe the center or average of a dataset,

dispersion => provide information about how the individual data points are spread around that center.

In data analysis => understanding not only the average income (mean) but also the spread of incomes (standard deviation) provides a more complete picture of the economic situation.

Quartiles

Quartiles ➔ are values that divide a dataset into four equal parts.
There are three quartiles: Q1 &Q2 &Q3

step to find quartiles:

1. Arrange the dataset in ascending order
2. Calculate the median (q2)
3. Divide data set two category left and right using origin median of data set
4. Find q1 = median in left and q3 median in right

End