Pre-Engagement Session Statistics

MIT-DSML

What is Statistics and Why is it so important?

What is Statistics?

- It is the study of the collection, analysis, interpretation, presentation, and organization of data

Why is it important?

- Availability of large amounts of data from which insights will have to be sifted
- Advances in enormous computing power to effectively process and analyze massive amounts of data
- Large data storage capability that helps businesses and other organizations to solve large scale problems faster than ever

Applications:

- There are many applications of statistics in multiple domains. A few of them are:

  Biostatistics, Quality Control, Environmental statistics etc.

Types of Statistics

1. **Descriptive Statistics** - Summarize the characteristics of data. It is concerned with Data Summarization, Graphs/Charts, and Tables

2. **Inferential Statistics** - Infers properties of a Population from a Sample

Central Tendency

- Single value that reflects the center of the data distribution
- Also called as Measures of Location or Statistical Averages
- Measures of central tendency: mean, median and mode

Mean, Median and Mode

- **Mean:** The sum of all observations in a data set divided by the total number of observations.

  In symbolic form the mean is given by

$$\overline{X} = \frac{\Sigma X}{n}$$

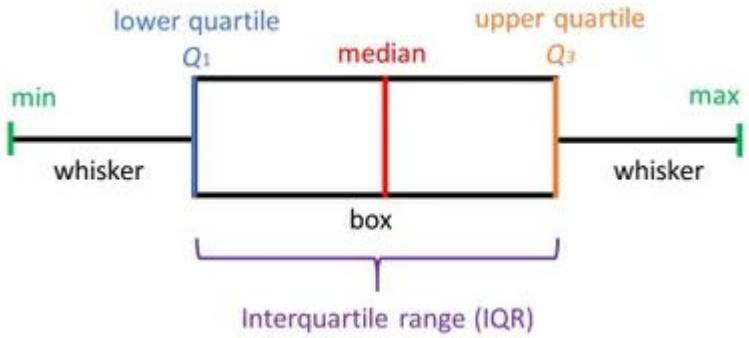
- **n** - The total number of observations
- **Compute the mean**:64, 69, 71, 67, 84
  - Applying the formula,
  - (64+69+71+67+84)/5 = 71
  - The mean is 71
- Arithmetic Mean is affected by **extreme values (outliers)** or fluctuations in sampling. It is not the best average to use when the data set contains extreme values (Very high or very low values)

- **Median** is the middle most observation when you arrange data in ascending order of magnitude. It is the (n+1)/2 th value of ranked data, and has more resistance to outliers than the mean.
- **Compute the median:** 35 30 50 70 80 55 45
- Arranging the data after ranking gives 80 70 55 **50** 45 35 30.
- n = 7, (n+1)/2 => 8/2 =4. Median is 4th value of ranked data
- The median is 50

- **Mode** is that value which occurs most often. It has the maximum frequency of occurrence. The mode also has higher resistance to outliers than the mean.
- **Compute the mode:** 40 50 40 40 20 40 30 30 40 50
- 40 occurs five times.
- The mode is 40

Measures of Dispersion

- Indicate how large the spread of the distribution is around the central tendency
- Measures of Dispersion: Range, Inter-Quartile Range (IQR), Standard Deviation
- **Range:** It is calculated as the difference between maximum and minimum value in the data set.
- **IQR:** Describes the middle 50% of values when ordered from lowest to highest.



Measures of Dispersion - Standard Deviation

- The **standard deviation** is a statistic that measures the amount of variation or dispersion of a dataset relative to its mean.
- It is calculated as the square root of variance by determining each data point deviation relative to the mean.
- A low standard deviation indicates that the values tend to be close to the mean.
- If the data points are further from the mean, there is a higher deviation within the data set.
- That is, the more spread out the data, the higher the standard deviation.

$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

σ	=	population standard deviation
N	=	the size of the population
x_i	=	each value from the population
μ	=	the population mean

Correlation

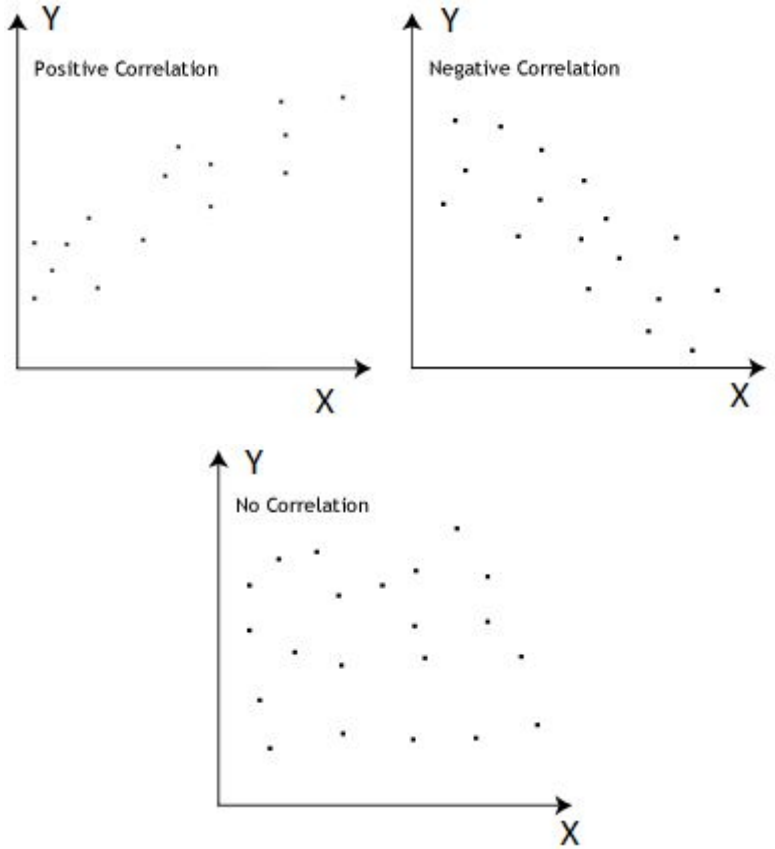
- Correlation denotes association between two variables.
- Correlation coefficients are used to measure how strong a relationship is.
- There are several types of correlation coefficients available. One example is the Pearson's Correlation Coefficient (r).
- Pearson's correlation Coefficient (r) between two variables x and y is given by:

$$r = \frac{N\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2 - (\Sigma x)^2][N\Sigma y^2 - (\Sigma y)^2]}}$$

N - the number of pairs of scores

The formulas return a value between -1 and 1, where:

- 1 indicates a strong positive relationship.
- -1 indicates a strong negative relationship.
- A result of zero indicates no relationship at all



Great Learning

Happy Learning !