Univariate Analysis in Data Science

1.**Mean**:

Also called Average used to find the central value of a distribution.

Mean=sum of all values/Total no. of values

2.**Median**-

It represents the middle value of a dataset when the values are arranged in order.

If the total number of observations n is odd, the median is the single, unique middle value.

If the total number of observations n is even then median is the average of these two middle values**.**

3.**Mode**:

The most frequently occurring value in a dataset. It can be used with both quantitative (numerical) and qualitative (categorical**)** data.

4.**Percentile**:

Percentile is a statistical measure that indicates the value below which a certain percentage of data points in a dataset fall.

Ex: 90th percentile is the value below which 90% of the observations lie.

5.**IQR**-

The interquartile range (IQR) is a measure that describes the spread of the middle 50% of values in a dataset.

IQR=$Q3-Q1$ where(Q3= 75th percentile and Q1= 25th percentile) Q1=k/100*(n+1) where k=which percentile

Lower Bound=Q1-1.5*IQR

Upper Bound=Q3+1.5*IQR

Any data point less than the lower bound or greater than the upper bound is classified as an **outlier**.

6.**Frequency:**

Frequency in data science refers to the count of how many times a particular value or category appears in a dataset.eg. value_counts().

- Relative Frequency = Frequency of value/Total no of values

- Cumulative Relative Frequency is the running total of the relative frequencies up to a certain point.

7.Histogram:

- A histogram is a graphical representation of Frequency in statistics and data science.
- Horizontal axis represents continuous intervals or bins.
- The vertical axis shows the frequency (count) of data points in each bin.

8.Skewness:

- Skewness in data science is a measure of the asymmetry or symmetry in the distribution of data points around the mean.
- **Skewness=3(Mean-Median)/Standard Deviation.**
- Positive Skew (Right Skew) =Mean > median.
- Negative Skew (Left Skew) =Mean < median.
- Zero Skew (Symmetric Skew) =>Mean = median=mode.

9.Kurtosis:

It is the measure describing the "**tailedness**" of a probability distribution.

- **Mesokurtic**: (normal distribution)kurtosis =3 are considered mesokurtic They have a moderate peak and medium-outliers.
- **Leptokurtic**: Kurtosis > 3 These are sharply peaked with heavy tails and more outliers.
- **Platykurtic**: Kurtosis <3 . These are flatter with light tails, and have rare outliers.

10.Variance: (Measure of Dataspread)

Variance is the average of the squares deviation of each data point. A high variance means data points are widely spread, while a low variance indicates they are closer to the mean.

11.Standard Deviation:

Standard deviation is the square root of variance. Higher standard deviation denotes greater spread, and a lower one shows data points closer to the mean.

12.Normal Distribution:(Gaussian distribution or bell curve)

- perfectly symmetrical around the mean
- mean=median=mode

- Empirical Rule= 68% of data points fall within one S.D of the mean, 95% within two SD, and 99.7% within 3 SD.

13.Probability Density Function: Tells Probability for particular range of values.

14.CDF-cumulative Probability density: It tells cumulative Probability up to certain point or for lesser range values.

15.Standard Normal Distribution:

It is a special case of the normal distribution where the mean (μ) is 0 and the standard deviation (σ) is 1. It is also known as the **Z-distribution**.

16.Z-score:

- Tells how many standard deviations a particular value x is from the mean (μ).
- Z-scores is used to compare data points across different distributions by standardizing them in SND.