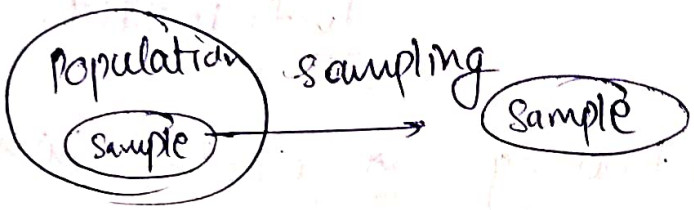
Data science

Population and sample :- sample is subset of population.

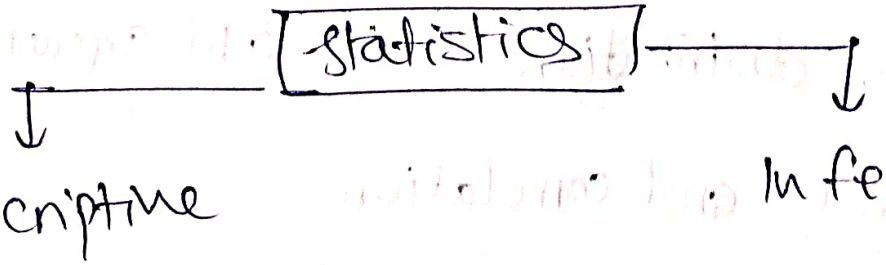
Population :- A population includes all the elements from a set of data. The population is whole set of values or individuals.

Sample :- The sample is a subset of population and is set of value you actually use in your estimation



STATISTICS :-

Statistics is a set of mathematical method and tools that enables us to answer



Descriptive statistics :- this offers method to summarized data by transforming raw observation to meaningful information.

Inferential Statistics :- this offers method to study experiment done on small sample of chalk out the inference to entities.

why inferential ?

↳ ① → suitable for large amount

↳ ② → sample data ∈ inferential stats stats

Descriptive

* measure of central tendency
* measure of variability
* measure of shapes
* percentiles
* frequency distribution
* covariance and correlation

Inferential

* central limit theorem
* hypothesis testing
  - z-test
  - t-test
* chi square test

Measure of central tendancy :-

Mean / Avg :- $\frac{\text{sum of all data}}{\text{total no. of data units}}$

// to read a dataset :-

```
dataset = pd.read_csv("titanic.csv")
```

Median :- null/data → sort → middle values

Mode :- repeating itself (most) → Mode

Range :- Range is difference b/w maximum and minimum value in dataset.

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

Mean absolute deviation :-

The mean absolute deviation :-

$$MAD = \frac{\sum |x_i - \bar{x}|}{n}$$

↳ Idea about variability in dataset

Standard deviation :-

The standard deviation is a measure of the amount of variation or dispersion of a set values. A low standard deviation tend close to mean

$$\sigma = \sqrt{\frac{\Sigma(xi-M)^2}{N}}$$

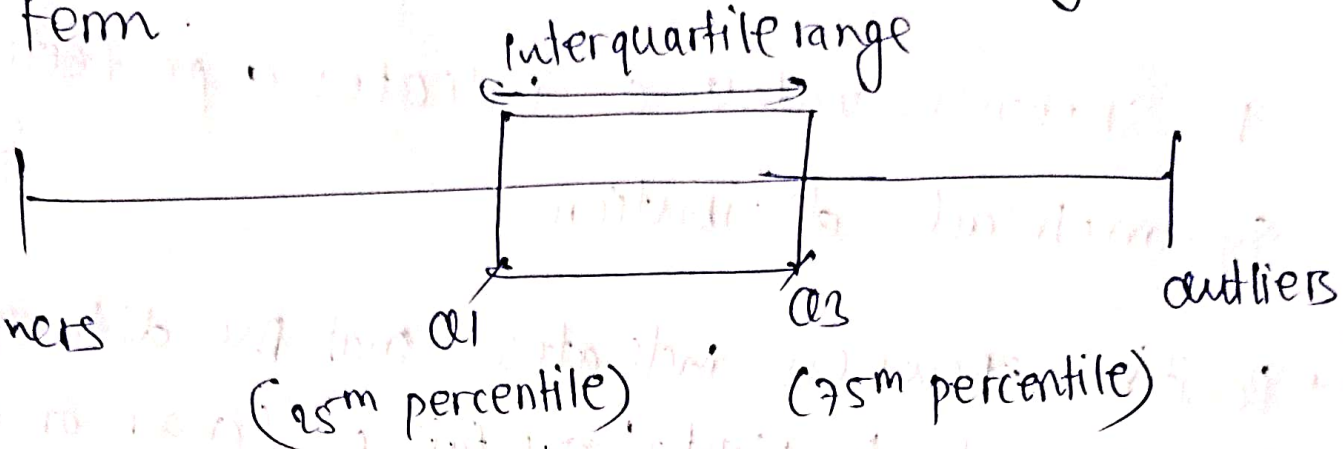
variance :-

variance is measure of how data point diff from mean. Acc to layman a variance is measure of how far a set of data are spread out.

$$\sigma^2 = \frac{\Sigma(xi-\bar{x})^2}{N}$$

Percentile :- Percentiles are used in statistics to gives you a number that describe the value that a given percent of the values are lower than.

Quartile → statistic term dividing into 4 term.



Inter quartile range :- to find outliers.

↳ IQR = Q3 − Q1

Q1 (first quartile) :- value below which 25% data

Q3 (third quartile) :- value below which 75% data

Outliners are defined data points fall below Q1 − 1.5 * IQR or above Q3 + 1.5 * IQR

Skewness :- how much data is spreaded.

↳ Skewness is how much asymmetric distribution

↳ A skewness value 0 indicates a perfectly symmetrical distribution

↳ Positive skewness indicates that the distribution is skewed to right (tail fail is longer on right) negative skewness (tail is longer on left)

$$\text{Skewness} = \frac{\sum (x_i - \bar{x})^3}{(N-1)\sigma^3}$$

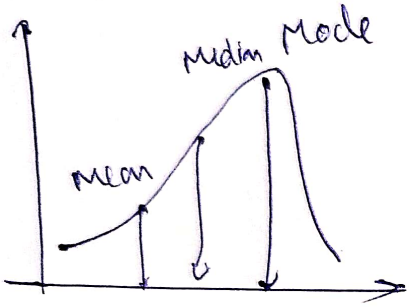
Frequency and cumulative Distribution:

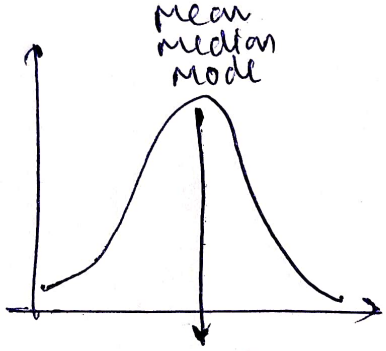
A frequency distribution is a table that shows the numbers of occurances of different value in data set.

A cumulative distribution show the accumulation of frequency up to certain point.

Skewness :-

-ve skewness





+ve skewness