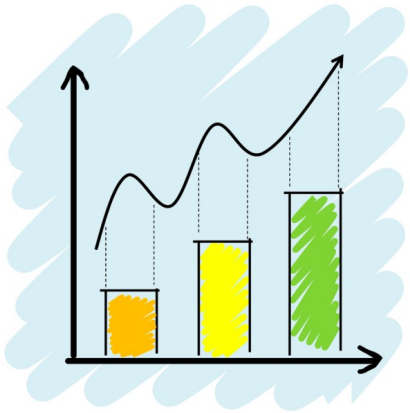


statistics

Descriptive

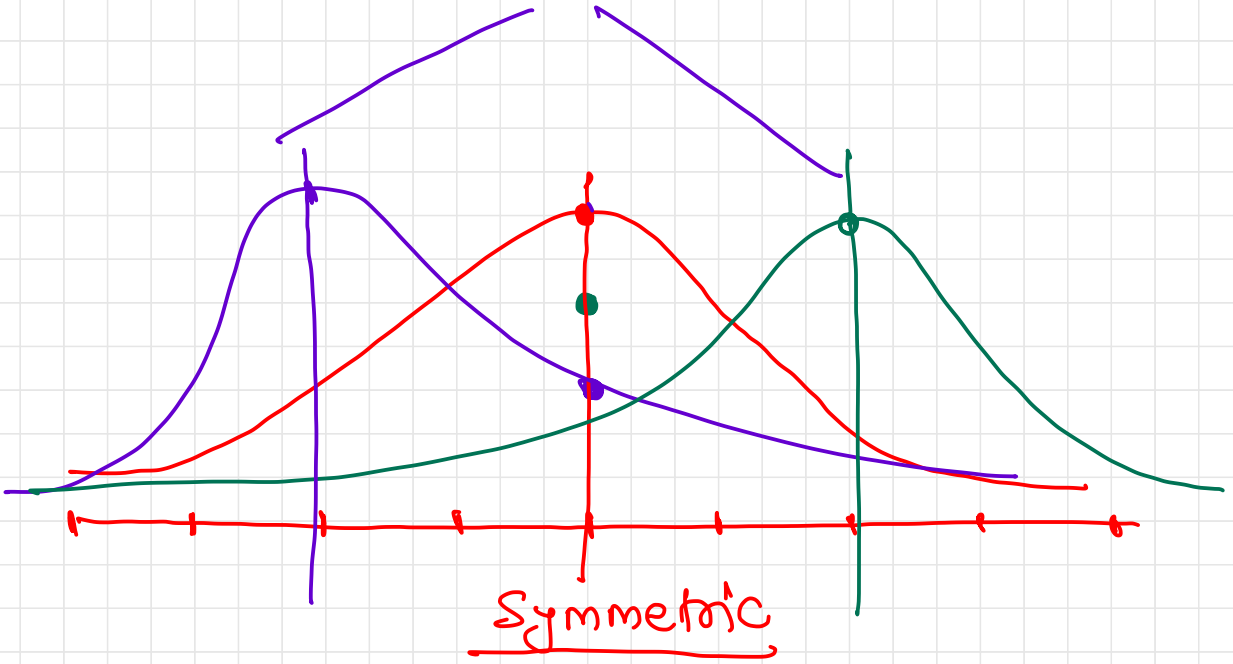
1. measures of central tendency
   - mode
   - median
   - mean
2. measures of asymmetry
   - positive skewness
   - negative skewness
3. measures of spread / dispersion
   - range
   - quartile
   - interquartile range (IQR)
   - variance
   - stadandard deviation
4. measures of relation
   - covariance
   - correlation

Inferential

Measures of Asymmetry

Skew

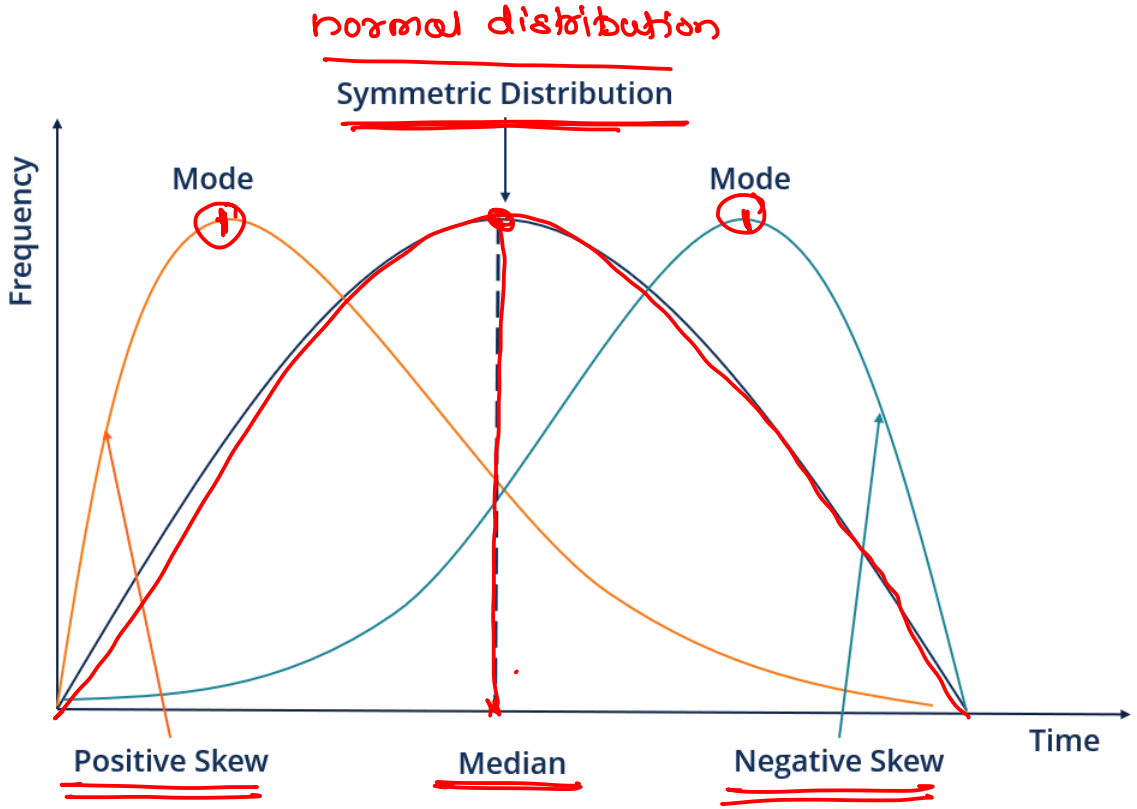
asymmetric



Symmetric

Skewness

- Skewness is a measure of asymmetry or distortion of symmetric distribution
- It measures the deviation of the given distribution of a random variable from a symmetric distribution, such as normal distribution
- A normal distribution is without any skewness, as it is symmetrical on both sides
- Hence, a curve is regarded as skewed if it is shifted towards the right or the left



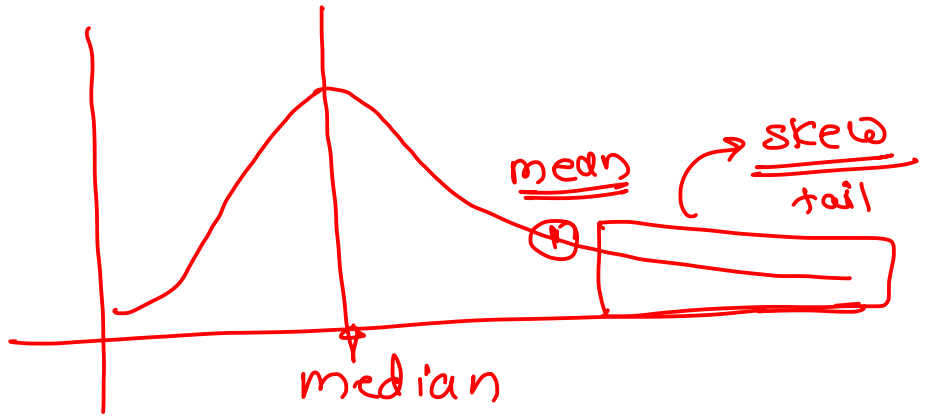
Skewness

- Skewness measures the deviation of a random variable's given distribution from the normal distribution, which is symmetrical on both sides.
- A given distribution can be either be skewed to the left or the right. Skewness risk occurs when a symmetric distribution is applied to the skewed data.
- Investors take note of skewness while assessing investments' return distribution since extreme data points are also considered.

Positive Skewness

mean > median

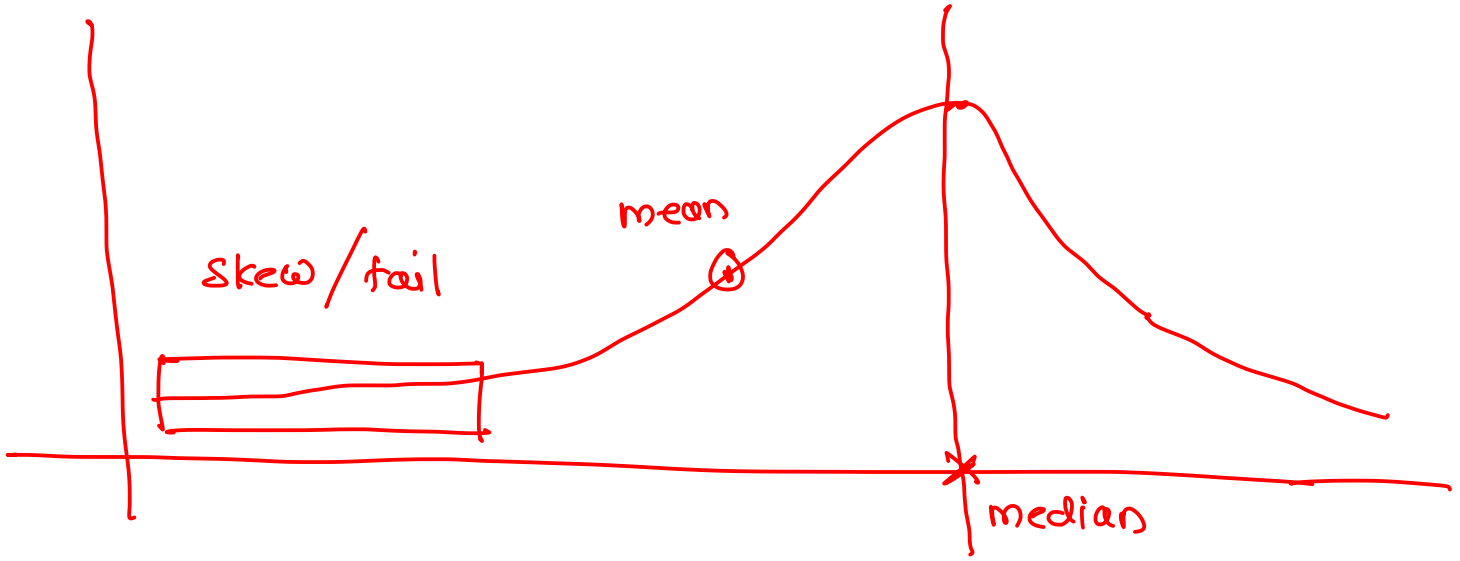
- If the given distribution is shifted to the left and with its tail on the right side, it is a positively skewed distribution
- It is also called the right-skewed distribution
- A tail is referred to as the tapering of the curve differently from the data points on the other side
- As the name suggests, a positively skewed distribution assumes a skewness value of more than zero
- Since the skewness of the given distribution is on the right, the mean value is greater than the median and moves towards the right, and the mode occurs at the highest frequency of the distribution



Negative Skewness

mean < median

- If the given distribution is shifted to the right and with its tail on the left side, it is a negatively skewed distribution. It is also called a left-skewed distribution
- The skewness value of any distribution showing a negative skew is always less than zero
- The skewness of the given distribution is on the left; hence, the mean value is less than the median and moves towards the left, and the mode occurs at the highest frequency of the distribution

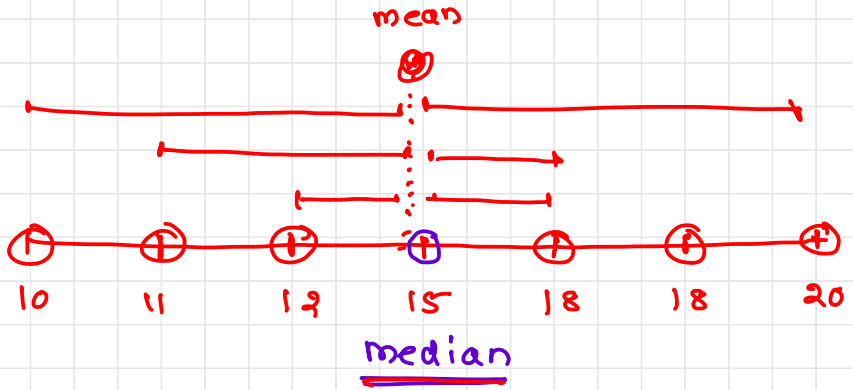


Measures of Dispersion

Spread

10, 15, 18, 11, 12, 18, 20

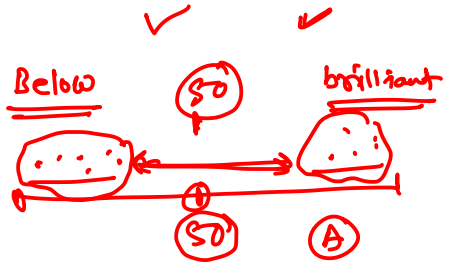
mean = 14.85



Dispersion

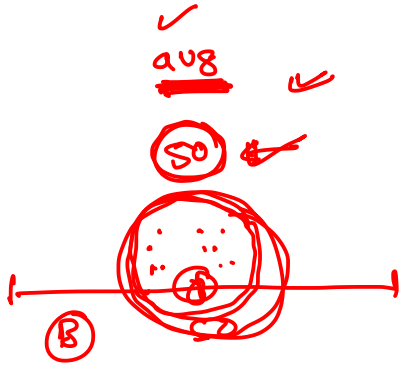
[distance between the data & mean] → spread

- To increase our understanding of the pattern of the data, we must also measure its dispersion—its spread, or variability
- In the diagram, the mean of all three curves is the same, but curve A has less spread (or *variability*) than curve B, and curve B has less variability than curve C



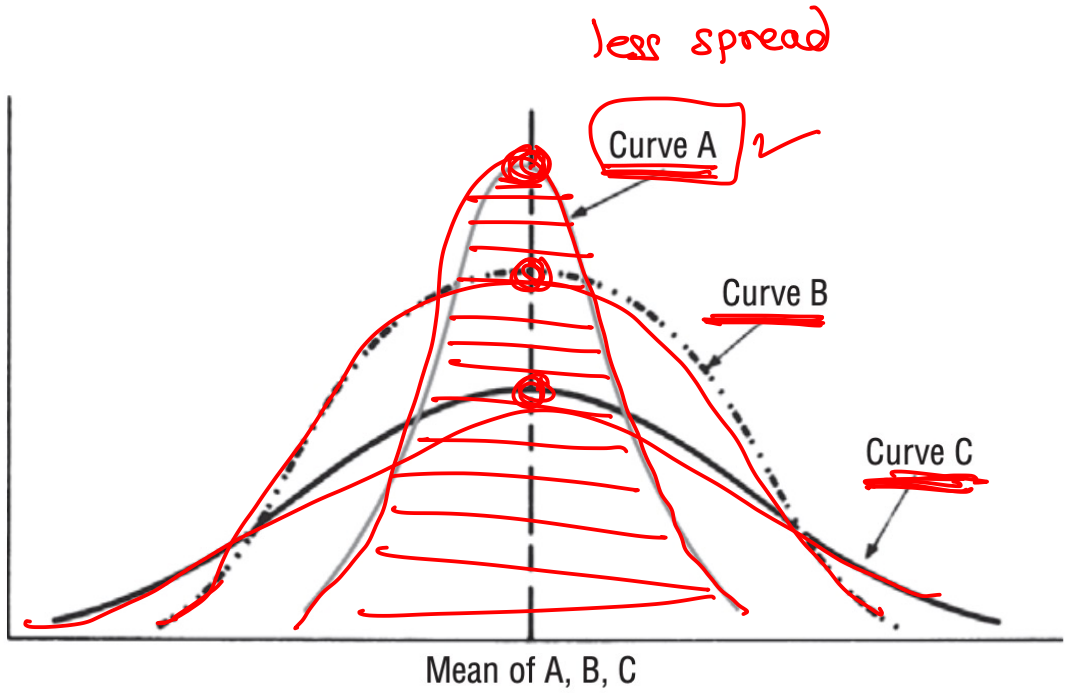
$10 \times 20 = 200$

$\frac{100 + 00}{2} = 200$



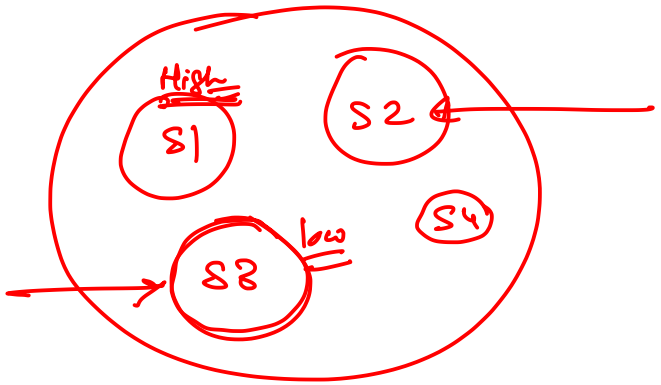
1. Range
2. Quartile
3. IQR
4. variance
5. std deviation

less spread



Why is it important ?

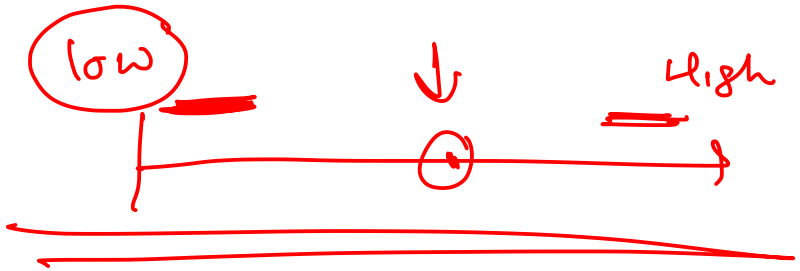
- It gives us additional information that enables us to judge the reliability of our measure of the central tendency
  - If data are widely dispersed, such as those in curve C, the central location is less representative of the data as a whole than it would be for data more closely centered around the mean, as in curve A
- Because there are problems peculiar to widely dispersed data, we must be able to recognize that data are widely dispersed before we can tackle those problems
- We may wish to compare dispersions of various samples. If a wide spread of values away from the center is undesirable or presents an unacceptable risk, we need to be able to recognize and avoid choosing the distributions with the greatest dispersion

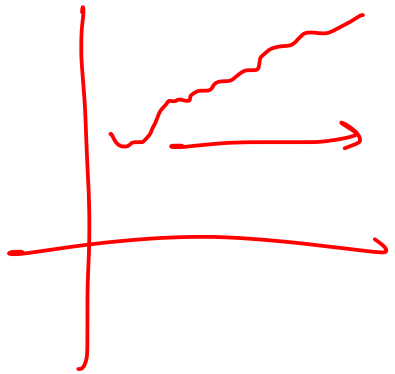


data = samples

Use of dispersion

- Financial analysts are concerned about the dispersion of a extremely high to low or even negative levels—indicate a higher risk to stockholders and creditors than do earnings remaining relatively stable
- Similarly, quality control experts analyze the dispersion of a product's quality levels. A drug that is average in purity but ranges from very pure to highly impure may endanger lives.





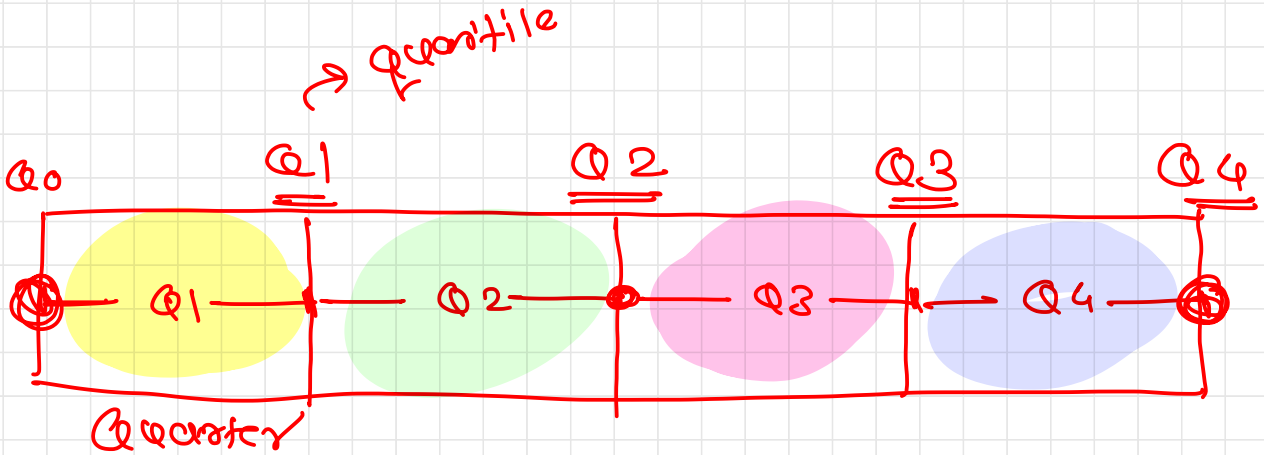
Range (1)

values in the dataset

- The range is the difference between the highest and lowest observed values
- It is easy to understand and to find, but its usefulness as a measure of dispersion is limited
- The range considers only the highest and the lowest values and fails to take account of any other observation in the data set
- As a result it ignores the nature of variation among all the other observations and it is heavily influenced by extreme values
- Because it measures only two values, the range is likely to change drastically from one sample to the next in a given population, even though the values that fall between the highest and lowest values may be quite similar

Range = max - min

Quartile : data must be sorted in asc order



- → Q_0 - minimum 0%.
- → Q_1 - 25%.
- → Q_2 - median → 50%.
- → Q_3 - 75%.
- → Q_4 - maximum 100%.

Quartile : 4 parts

Quintile : 5 parts

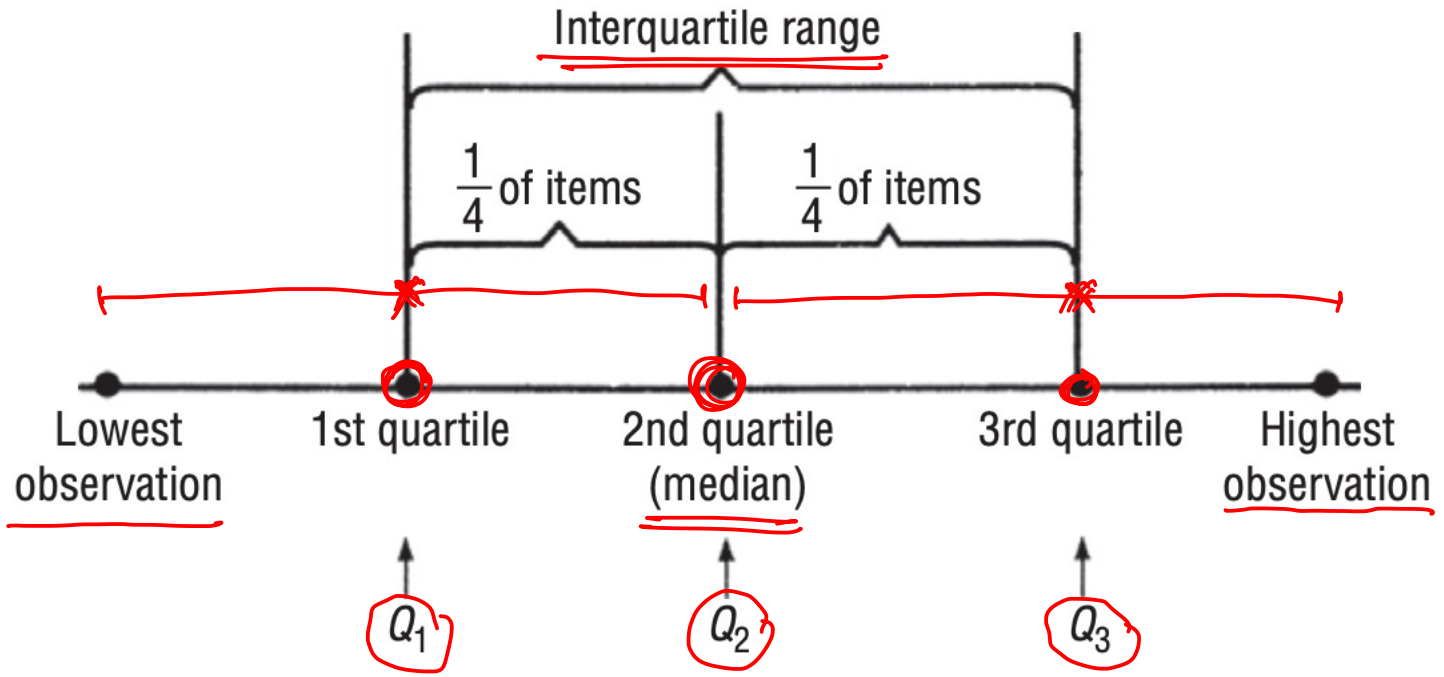
Decile : 10 parts

Percentile : 100 parts

Interquartile Range

- The interquartile range measures approximately how far from the median we must go on either side before we can include one-half the values of the data set
- To compute this range, we divide our data into four parts, each of which contains 25 percent of the items in the distribution
- The quartiles are then the highest values in each of these four parts, and the interquartile range

IQR = Q3 - Q1



1 2 4 3 5 6 = Population

$N = 6$

$\mu = 3.5$

$$\text{mean}(\mu) = \frac{21}{6} = 3.5$$

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

$$\sigma^2 = \frac{17.50}{6} = \boxed{2.25}$$

x	$x-\mu$	$(x-\mu)^2$
1	-2.5	6.25
2	-1.5	2.25
4	0.5	0.25
3	-0.5	0.25
5	1.5	2.25
6	2.5	6.25
	0	17.50

N: population count
n: sample count
μ : population mean
$\bar{x}$: sample mean

Range = Highest - lowest

IQR = Q3 - Q1

variance : var()

$$\text{population} = \frac{\Sigma(x-\mu)^2}{N} \quad , \quad \text{sample} = \frac{\Sigma(x-\bar{x})^2}{n-1}$$

standard deviation : std()

$$\text{population} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} \quad , \quad \text{sample} = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$

```
a1 = np.array(...)
a1.var(), a1.std()
```

Population Variance

- Every population has a variance, which is symbolized by σ2 (sigma squared)
- It tells us an average distance of any observation in the data set from the mean of the distribution
- To calculate the population variance, we divide the sum of the squared distances between the mean and each item in the population by the total number of items in the population
- By squaring each distance, we make each number positive and, at the same time, assign more weight to the larger deviations (deviation is the distance between the mean and a value)

μ = mean of population

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

where

- σ^2 = population variance
- x = item or observation
- μ = population mean
- N = total number of items in the population
- Σ = sum of all the values $(x-\mu)^2$ or all the values x^2

Population Standard Deviation

- It tells us an average distance of any observation in the data set from the mean of the distribution
- The population standard deviation, or σ, is simply the square root of the population variance
- Because the variance is the average of the squared distances of the observations from the mean, **the standard deviation is the square root of the average of the squared distances of the observations from the mean**
- While the variance is expressed in the square of the units used in the data, the standard deviation is in the same units as those used in the data

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma(x-\mu)^2}{N}} = \sqrt{\frac{\Sigma x^2}{N} - \mu^2}$$

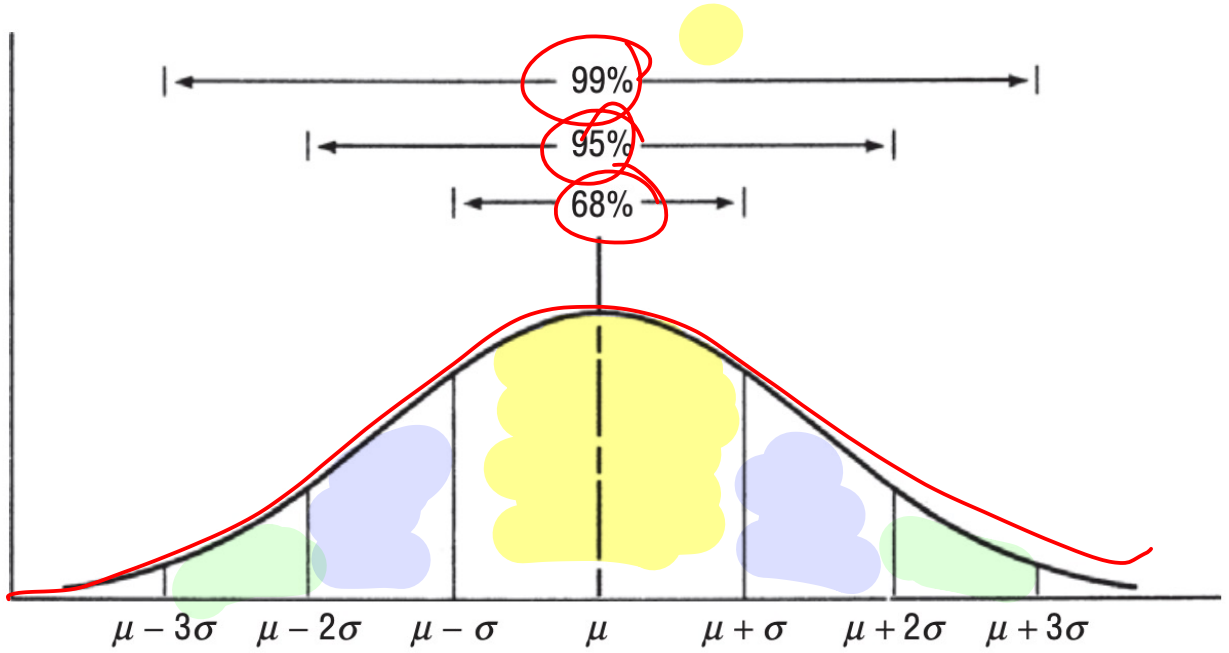
- x = observation
- μ = population mean
- N = total number of elements in the population
- $\sum$ = sum of all the values $(x-\mu,)^2$, or all the values x^2
- σ = population standard deviation
- σ^2 = population variance

Uses of the Standard Deviation

- The standard deviation enables us to determine, with a great deal of accuracy, where the values of a frequency distribution are located in relation to the mean
- We can do this according to a theorem devised by the Russian mathematician P. L. Chebyshev
- Chebyshev's theorem says that no matter what the shape of the distribution
  - at least 75 percent of the values will fall within ± 2 standard deviations from the mean of the distribution
  - and at least 89 percent of the values will lie within ± 3 standard deviations from the mean

Chebyshev's theorem

- We can measure with even more precision the % of items that fall within specific ranges under a symmetrical, bell-shaped curve
- About 68 percent of the values in the population will fall within ± 1 standard deviation from the mean.
- About 95 percent of the values will lie within ±2 standard deviations from the mean
- About 99 percent of the values will be in an interval ranging from 3 standard deviations below the mean to 3 standard deviations above the mean



Uses of the Standard Deviation

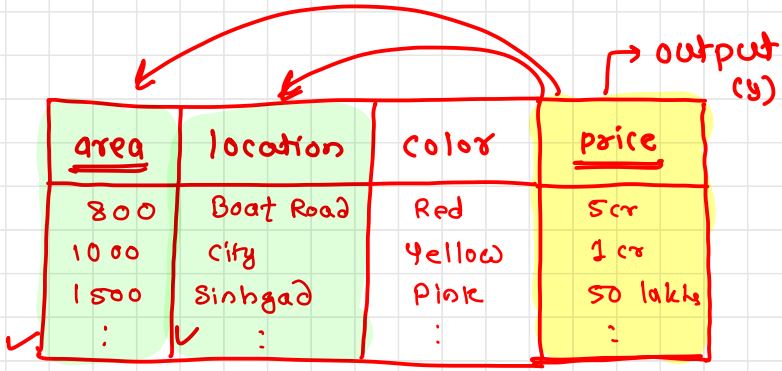
- The standard deviation is also useful in describing how far individual items in a distribution depart from the mean of the distribution
- A measure called the standard score gives us the number of standard deviations a particular observation lies below or above the mean

$$\text{Population standard score} = \frac{x - \mu}{\sigma}$$

where

- x = observation from the population
- μ = population mean
- σ = population standard deviation

Covariance and Correlation



real estate → 1

predict house price with area = 550

1 cr

independent input — dependent output

x	y
1	1
2	4
3	9
4	16
5	25
10	?

$y = x^2$

formula

model

$y = 10^2 = 100$

→ covariance.

Relationship = No direction correlation

(exp, salary) , = (salary, exp)

→ regression

Causes / causal Relationship = direction

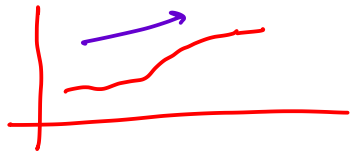
Salary ← Exp ✓

Exp ← Salary ✗

Terminology

- **Univariate** (vector)
  - This type of data consists of only one variable
  - It does not deal with causes or relationships
  - the main purpose of the analysis is to describe the data and find patterns that exist within it

temperature = [27, 30, 32, 35 ...]



- **Bivariate**
  - This type of data involves two different variables (unidirectional)
  - The analysis of this type of data deals with causes and relationships
  - the analysis is done to find out the relationship among the two variables

exp	salary
1.5	15
2	20
3	25
3.5	28

Salary ← exp

Salary ~ exp (depends)

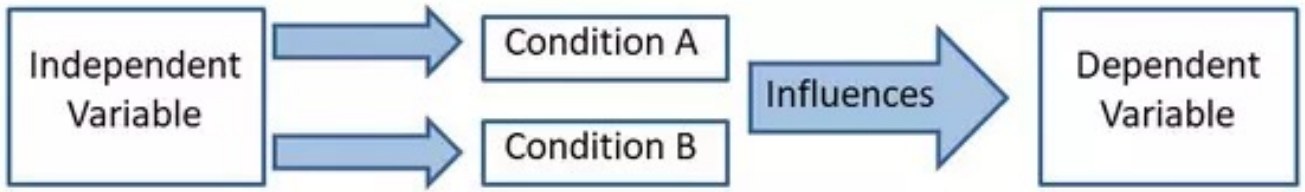
- **Multivariate**
  - When the data involves three or more variables
  - It is similar to bivariate but contains more than one dependent variable

area	location	color	price

Terminology

- **Independent variable(s)**
  - A variable that represents a quantity that is being manipulated in an experiment
  - Represents input
  - Also known as regressors in a statistical context.
  - x is often the variable used to represent the independent variable in an equation

- **Dependent variable**
  - A quantity whose value *depends* on how the independent variable is manipulated
  - Represents output
  - y is often the variable used to represent the independent variable in an equation



Covariance

- A measure of the relationship between two random variables
- The metric evaluates how much – to what extent – the variables change together
- A positive covariance would indicate a positive linear relationship between the variables
- A negative covariance would indicate the opposite

$$cov(x,y) = \frac{\sum(X_i - \mu)(Y_i - \mu)}{N}$$

$$cov(x,y) = \frac{\sum(X_i - \overline{x})(Y_i - \overline{y})}{n-1}$$

- Where
  - X_i – the values of the X-variable
  - Y_i – the values of the Y-variable
  - $\overline{x}$ – the mean (average) of the X-variable
  - $\overline{y}$ – the mean (average) of the Y-variable
  - n – the number of the data points in sample
  - N – the number of the data points in Population

x	y	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$
10	18			
15	20			
18	30			
20	35			
25	40			

$$\bar{x} = \frac{\Sigma x_i}{n} \qquad \bar{y} = \frac{\Sigma y_i}{n}$$

$cov(x,y) = \underline{50.8}$, $cov(y,x) = \underline{50.8}$

$$cov(x,y) = \begin{bmatrix} cov(x,x) & cov(x,y) \\ cov(y,x) & cov(y,y) \end{bmatrix}$$

$$corrcoef(x,y) = \begin{bmatrix} (x,x) & (x,y) \\ (y,x) & (y,y) \end{bmatrix}$$

$$cov(x,y) = cov(y,x),$$

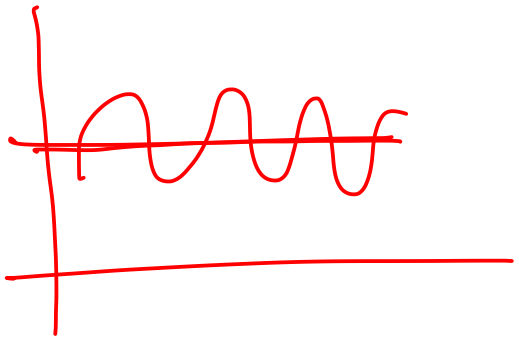
$$corrcoef(x,y) = corrcoef(y,x)$$

Correlation

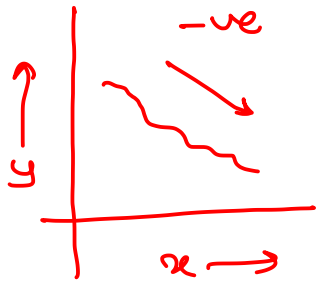
→ strong relation (+ or -ve)
→ weak relation (+ or -ve)
→ No relation (0)

np.corrcoef(..) , cor(..)

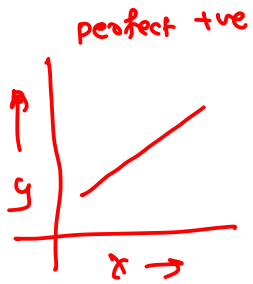
- Measures the strength of the relationship between variables
- Correlation is the scaled measure of covariance
- It is dimensionless: the correlation coefficient is always a pure value and not measured in any units

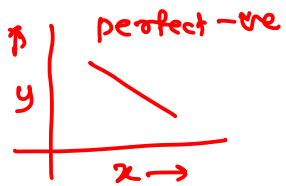


$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$



- Where
  - ρ(X,Y) – the correlation between the variables X and Y
  - cov(X,Y) – the covariance between the variables X and Y
  - σ_X – the standard deviation of the X-variable
  - σ_Y – the standard deviation of the Y-variable





Correlation coefficient

- Following are the ways to calculate correlation coefficient
  - Karl Pearson's coefficient of correlation
  - Spearman's Rank correlation
  - Scatter diagram
  - Coefficient of concurrent duration

- Correlation (r)
  - -1 <= r <= 1
  - r = 1 (perfect correlation)
  - r = -1 (perfect negative correlation)
  - r > 0 (positive correlation)
  - r < 0 (negative correlation)
  - r = 0 (no correlation)

Karl Pearson's coefficient of correlation

- Also known product moment coefficient of correlation

$$r = \frac{N\sum XY - \sum X \sum Y}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}}$$

Spearman's Rank correlation

- Find the correlation using the rank

$$r = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

$$r = 1 - \frac{6 \left[\sum d^2 + \frac{\sum (m^3 - m)}{12}\right]}{n(n^2 - 1)}$$

Covariance vs Correlation

Covariance	Correlation
It shows the extent to which two variables are dependent on each other	It measures the strength of two variables considering other conditions are constant
Range is $-\infty\ to\ +\ \infty$	Range is -1 to +1
It is affected by scale of variables	It is not affected by scale
It has definite units as it is derived by multiplying two numbers and they units	It is unit less

Regression Analysis

What is regression analysis

- Linear regression is a basic and commonly used type of predictive analysis
- The dictionary meaning of the word Regression is ‘Stepping back’ or ‘Going back’
- Set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables
- It attempts to establish the functional relationship between the variables and thereby provide a mechanism for prediction or forecasting
- The overall idea of regression is to examine two things
  - does a set of predictor variables do a good job in predicting an outcome (dependent) variable?
  - Which variables in particular are significant predictors of the outcome variable, and in what way do they– indicated by the magnitude and sign of the beta estimates–impact the outcome variable?
- These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables

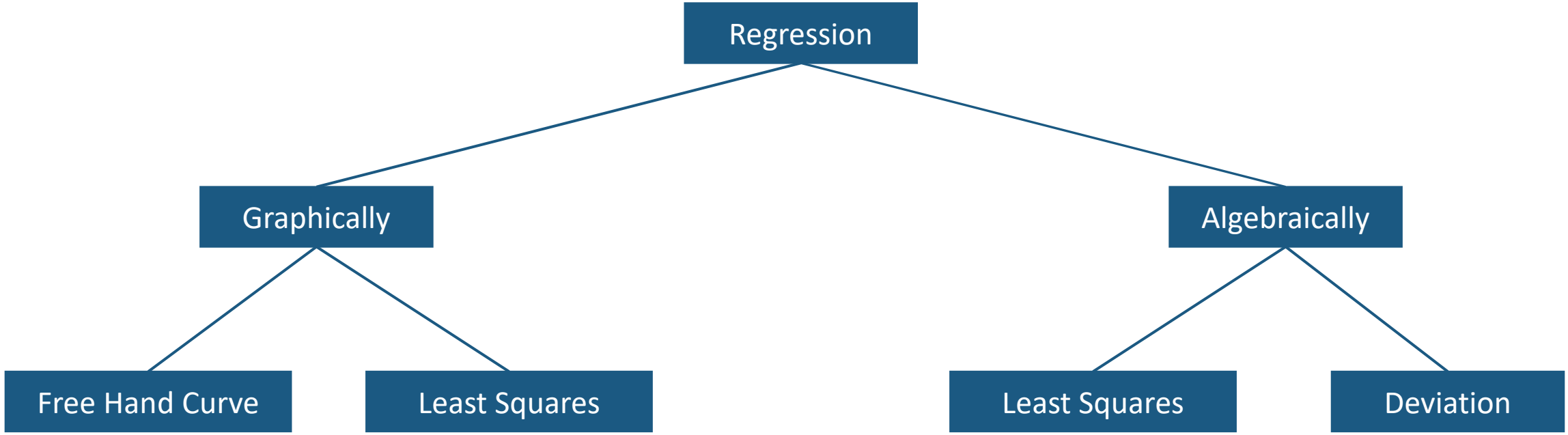
Correlation vs Regression

- Correlation is a statistical measure which determines co-relationship or association of two variables while Regression describes how an independent variable is numerically related to the dependent variable
- Correlation is used to represent linear relationship between two variables while regression is used to fit a best line and estimate one variable on the basis of another variable.
- Cov(x, y) = Cov(y, x)
- but reg(x,y) != reg(y, x)

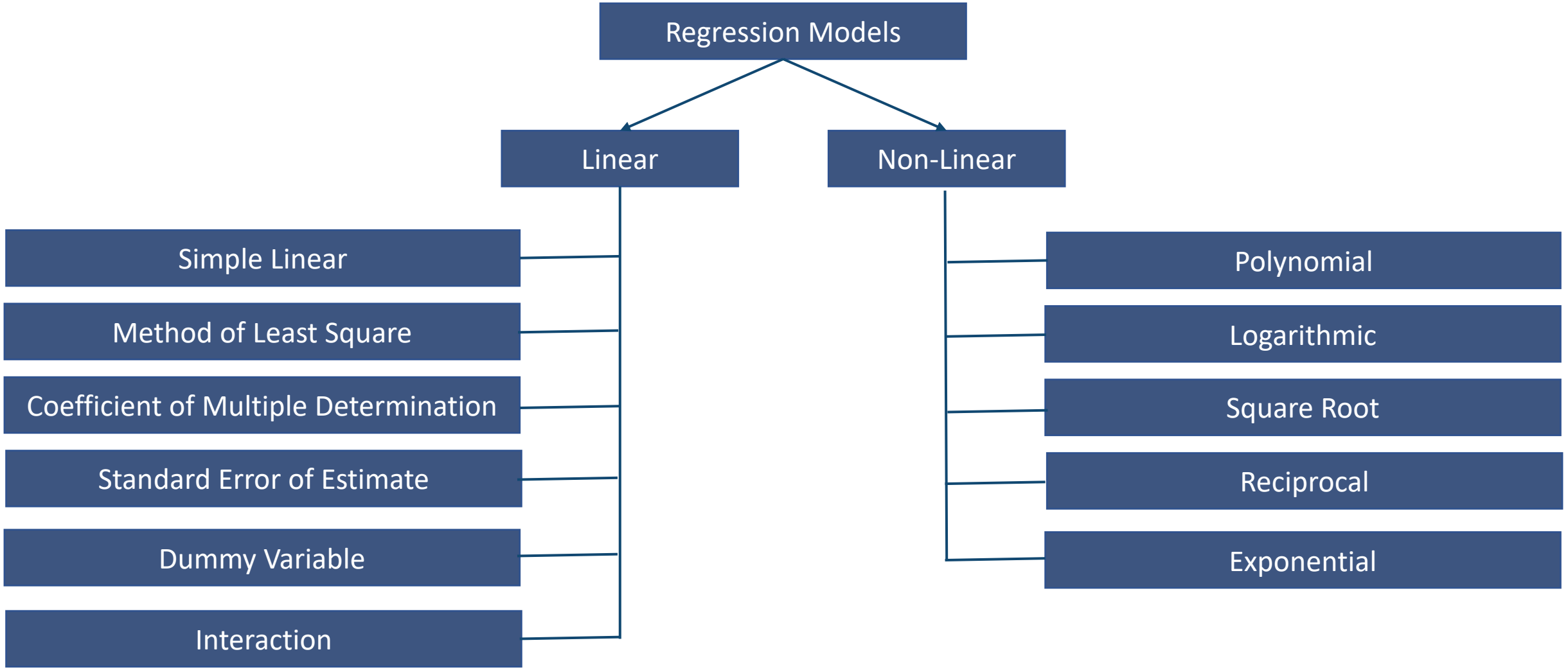
Applications of regression analysis

- It helps in the formulation and determination of functional relationship between two or more variables
- It helps in establishing a cause and effect relationship between two variables in economics and business research
- It helps in predicting and estimating the value of dependent variable as price production sales etc
- It helps to measure the variability or spread of values of a dependent variable with respect to the regression line
- In the field of business regression is widely used by businessmen in
  - Predicting future production
  - Investment analysis
  - Forecasting on sales etc.

Methods of studying regression



Regression Models



Types of regression

- Linear Regression
- Polynomial Regression
- Logistic Regression
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Support Vector Regression
- Quantile Regression
- Principle Component Regression
- Partial Least Square Regression
- Ordinal Regression
- Poisson Regression
- Negative Binomial Regression
- Cox Regression

Regression Equation

Adv	Sales
100	500
90	400
80	450
95	510
150	??

Sales depends on advertisement
Y depends on X [Y on X]

$(Y - \bar{Y}) = b_{yx} (X - \bar{X})$

$$b_{yx} = \frac{n \sum XY - \sum X \sum Y}{n \sum X^2 - (\sum X)^2}$$

$$b_{yx} = r \frac{\sigma y}{\sigma x}$$

$$b_{yx} = \frac{cov(x, y)}{(\sigma x)^2}$$

Regression Equation

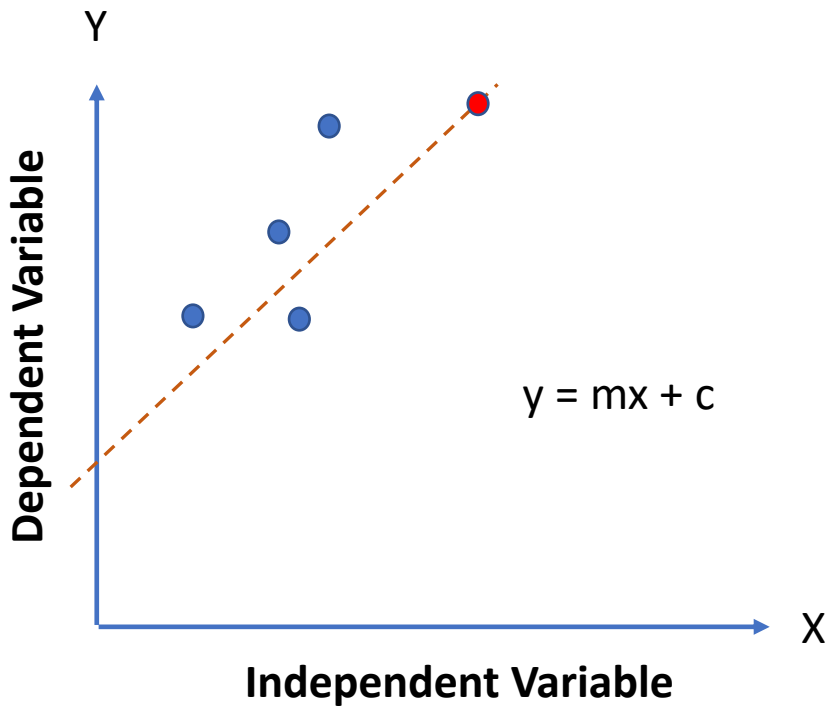
X	Y
3	11
4	12
8	9
7	3
2	5

What likely to be the value of Y if X = 10

Linear Regression

Overview

- The data in Linear Regression is modelled using a straight line
- It is used with continuous variable
- It gives a future value as an output
- To calculate accuracy following methods are used
  - R-squared
  - Adjusted R-squared



Least Square Method

- A form of mathematical regression analysis used to determine the line of best fit for a set of data, providing a visual demonstration of the relationship between the data points
- Each point of data represents the relationship between a known independent variable and an unknown dependent variable
- The least squares method provides the overall rationale for the placement of the line of best fit among the data points being studied
- It aims to create a straight line that minimizes the sum of the squares of the errors that are generated by the results of the associated equations, such as the squared residuals resulting from differences in the observed value, and the value anticipated, based on that model
- It begins with a set of data points to be plotted on an x- and y-axis graph
- An analyst using the least squares method will generate a line of best fit that explains the potential relationship between independent and dependent variables.

R-squared

- R-squared value is a statistical measure of how close the data are to the fitted regression line
- It is also known as coefficient of determination or coefficient of multiple determination

$$R^2 = \frac{\sum (\hat{Y} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}$$