# STATISTICAL BASICS

## BASIC DEFINITIONS

Let us consider the following frequency distribution table consisting the weights of 40 students.

Table: Frequency distribution table for the weight of 40 students

| Class Interval (Class) | Class Boundary (Original Class) | Class Mark (Mid Value) (x) | Frequency (f) | Cumulative Frequency (F) |
|---|---|---|---|---|
| 54 - 57 | 53.5 - 57.5 | 55.5 | 5 | 5 |
| 58 - 61 | 57.5 - 61.5 | 59.5 | 5 | 10 |
| 62 - 65 | 61.5 - 65.5 | 63.5 | 9 | 19 |
| 66 - 69 | 65.5 - 69.5 | 67.5 | 14 | 33 |
| 70 - 73 | 69.5 - 73.5 | 71.5 | 7 | 40 |

A **frequency distribution** shows us a summarized grouping of data divided into mutually exclusive classes and the number of occurrences in a class. It is a way of showing unorganized data notably to show results of an event considered for a certain interest.

The **frequency distribution** table is an arrangement of the values that one or more variables take in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval, and in this way, the table summarizes the distribution of values in the sample.

The **frequency distribution** is a representation, either in a graphical or tabular format that displays the number of observations within a given interval. The interval size depends on the data being analyzed and the goals of the analyst. The intervals must be mutually exclusive and exhaustive.

In statistics, a **frequency distribution** is a list, table or graph that displays the frequency of various outcomes in a sample. Each entry in the table contains the frequency or count of the occurrences of values within a particular group or interval.

The **class interval** (or **class width**) is the same for all classes. The classes all taken together must cover at least the distance from the lowest value (minimum) in the data to the highest (maximum) value. **Equal class** intervals are preferred in a frequency distribution, while **unequal class** intervals (for example logarithmic intervals) may be necessary for certain

situations to produce a good spread of observations between the classes and avoid a large number of empty, or almost empty classes.

Corresponding to a class interval, the **class limits** may be defined as the minimum value and the maximum value the class interval may contain. The minimum value is known as the **lower-class limit (LCL)** and the maximum value is known as the **upper-class limit (UCL)**.

The **class boundaries** may be defined as the actual class limit of a class interval. For overlapping classification or mutually exclusive classification, the class boundaries coincide with the class limits. This is usually done for a continuous variable. However, for non-overlapping or mutually inclusive classification, we have **lower-class boundaries (LCB)** and **upper-class boundaries (UCB)** will have the following forms.

$$LCB = LCL - \frac{D}{2} \quad \& \quad UCB = UCL + \frac{D}{2}$$

where D is the difference between the LCL of the next class interval and the UCL of the given class interval.

The **class midpoint** (or **class-mark**) is a specific point in the center of the classes in a frequency distribution table. It's also the center of a bar in a histogram. It is defined as the average of the upper and lower class limits. The lower-class limit is the lowest value in a class and the upper-class limits are the highest values that can be in the class. In other words, in a class interval, **class mid-point** may be defined as an arithmetic mean or average of the class limits or the class boundaries.

The **frequency** (or **absolute frequency**) of an event is the number of times the event occurred in an experiment or study. These frequencies are often graphically represented in histograms.

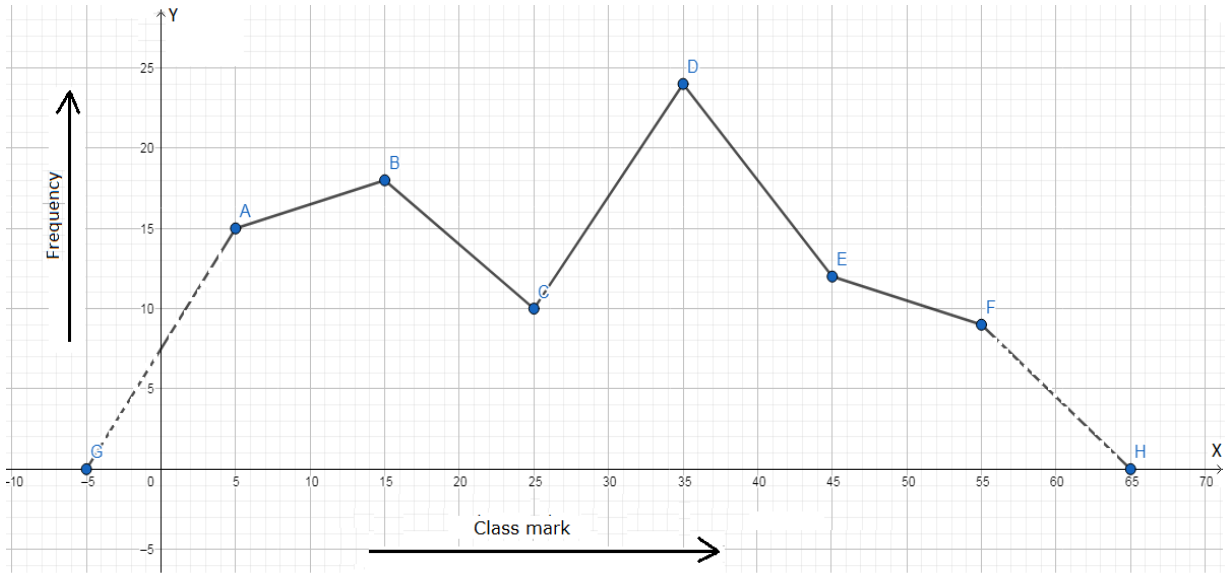
**Cumulative frequency** is defined as a running total of frequencies. The frequency of an element in a set refers to how many of that element there are in the set. Cumulative frequency can also be defined as the sum of all previous frequencies up to the current point.

**Cumulative frequency analysis** is the analysis of the frequency of occurrence of values of a phenomenon less than a reference value. The phenomenon may be time- or space-dependent. Cumulative frequency is also called the frequency of non-exceedance.

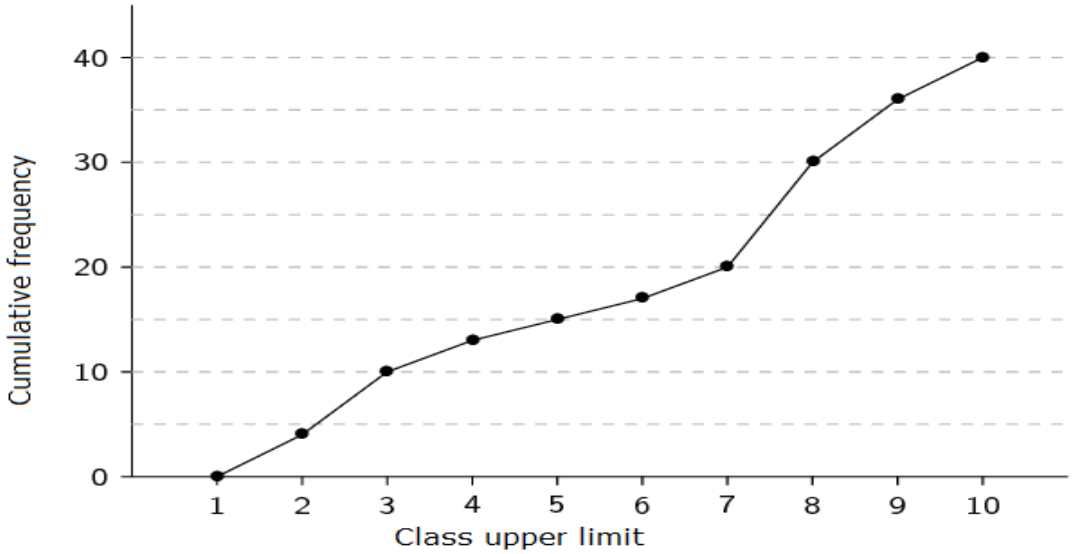
# STATISTICAL BASICS

## STATISTICAL GRAPHS

**Frequency polygon:** They are formed by lines. On the horizontal axis is the independent variable (class marks) and on the vertical axis is the dependent variable (frequency).

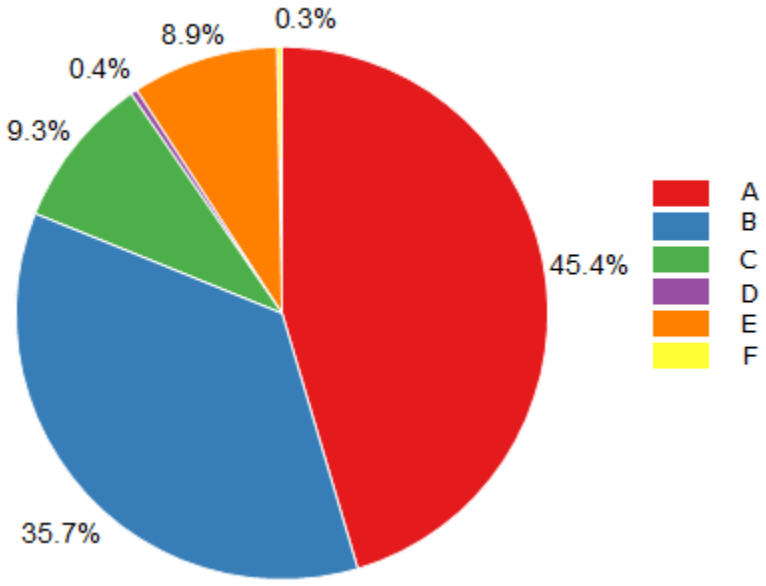



**Cumulative frequency polygon:** They are formed by increasing lines. On the horizontal axis is the independent variable (class upper limit) and on the vertical axis is the dependent variable (cumulative frequency).

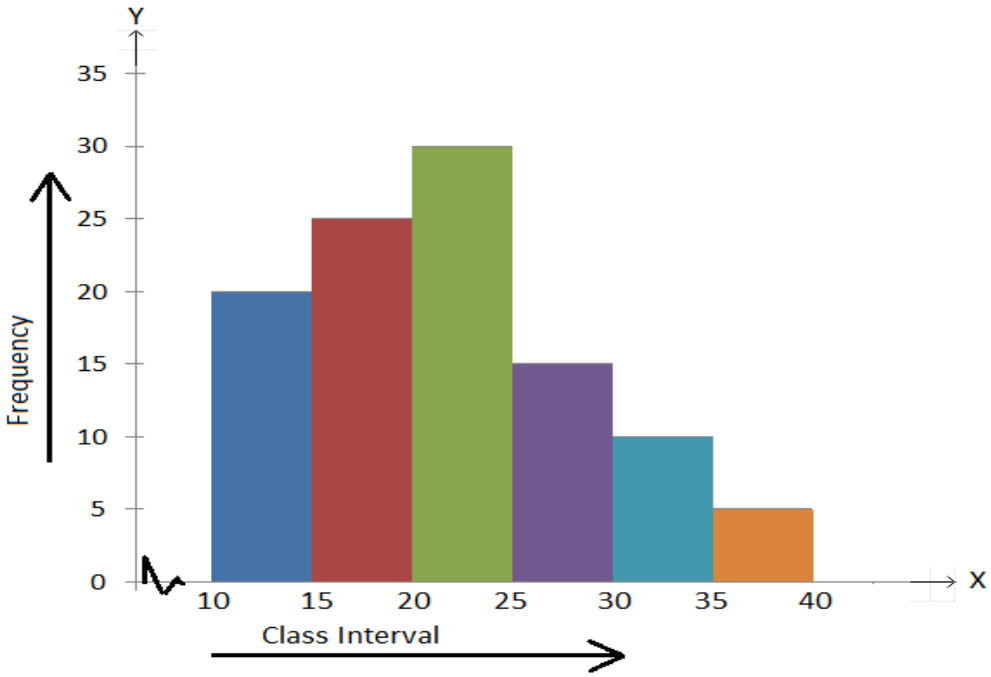


# STATISTICAL BASICS

**Pie chart:** A circle is divided into sectors. The amplitude of each sector is proportional to the corresponding frequency.

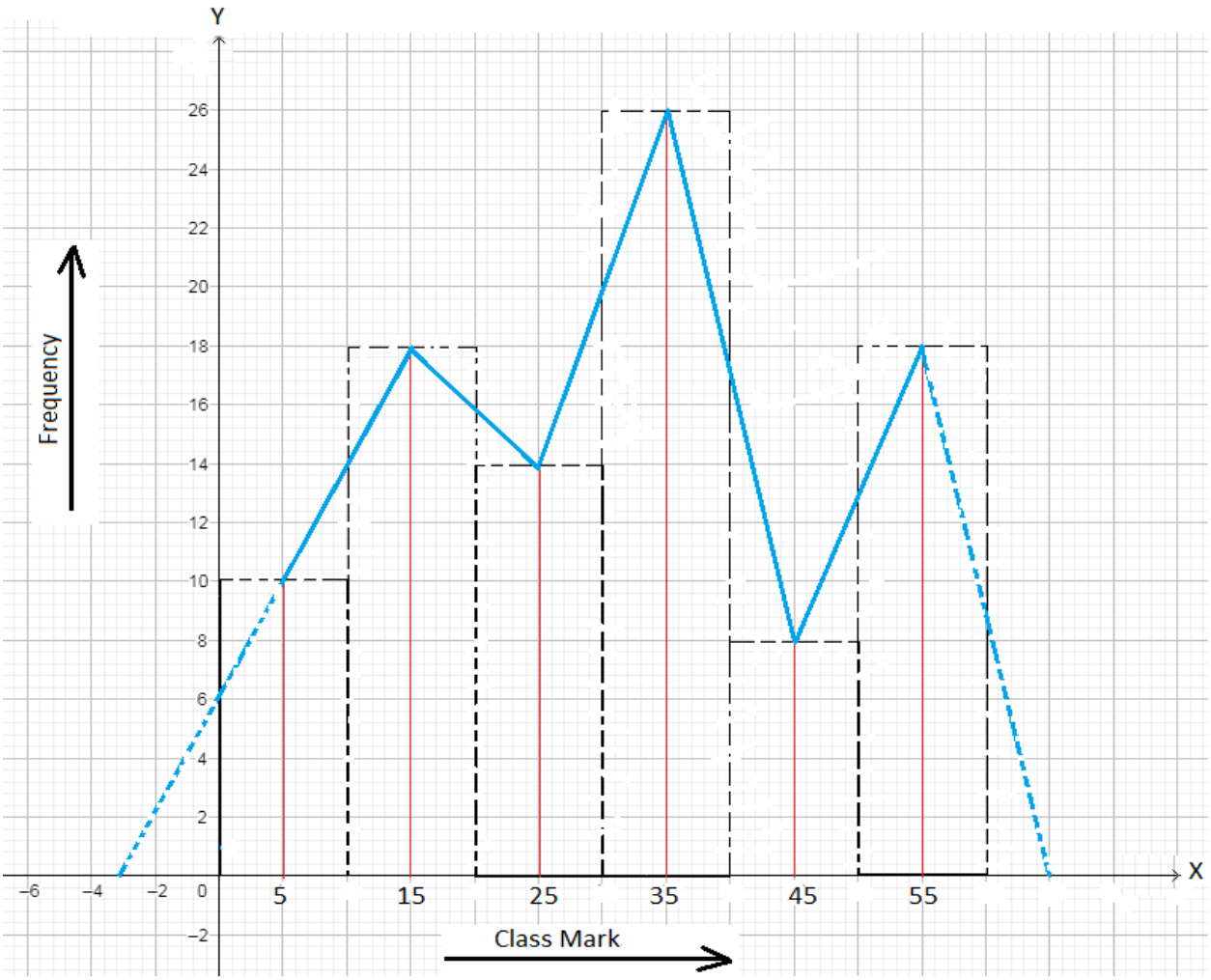



**Histogram:** It is a bar graph in which the height of these bars is proportional to the frequency. There is no space between bars. It is only used if the variable is quantitative and the scale of the values is continuous.



# STATISTICAL BASICS

**The relation between the frequency polygon and the histogram:**




Please, visit the following links for more details.

https://www.statisticshowto.datasciencecentral.com/

https://www.tutorialspoint.com/statistics/index.htm

https://people.richland.edu/james/lecture/m170/ch02-def.html

https://www.scribbr.com/statistics/

https://libguides.library.curtin.edu.au/uniskills/numeracy-skills/statistics

https://www.statisticshowto.com/probability-and-statistics/

https://courses.lumenlearning.com/introstats1/chapter/learning-outcomes/

# STATISTICAL BASICS

## THE MEASURE OF CENTRAL TENDENCY

**Central tendency:** A measure of central tendency is a single value that attempts to describe a set of data by identifying the central position within that set of data. As such, measures of central tendency are sometimes called measures of central location.

The Mean, Mode (Mo) and Median (Me) are all valid measures of central tendency, but under different conditions, some measures of central tendencies, such as Quartile, Decile, and Percentile become more appropriate to use than others.

There are three types of Mean, namely Arithmetic Mean (AM), Geometric Mean (GM), and Harmonic Mean (HM). For $n$ number of classes

$$AM = \bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}$$

$$GM = \left( \prod_{i=1}^{n} {x_i}^{f_i} \right)^{\frac{1}{\sum_{i=1}^{n} f_i}}$$

$$HM = \frac{\sum_{i=1}^{n} f_i}{\sum_{i=1}^{n} \frac{f_i}{x_i}}$$

The working formula for Geometric Mean is

$$GM = \text{Antilog}\left( \frac{\sum_{i=1}^{n} f_i \log x_i}{\sum_{i=1}^{n} f_i} \right) = 10^{\left( \frac{\sum_{i=1}^{n} f_i \log x_i}{\sum_{i=1}^{n} f_i} \right)}$$

For the shift $a$ and scale $h$, the coding formula for Arithmetic Mean is

$$AM = a + \frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i} \times h \; ; \; u_i = \frac{x_i - a}{h}$$

$$Mo = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times C$$

The class at which the highest frequency is present is called the modal class. $L$ is the lower limit of the modal class, $\Delta_1$ is the frequency difference between the modal and pre-modal class, $\Delta_2$ is the frequency difference between the modal and post-modal class, and $C$ is the class size.

# STATISTICAL BASICS

$$Me = L + \frac{\frac{N}{2} - F_{m-1}}{f_m} \times C$$

The class at which $\frac{N}{2} - th$ frequency ($N = \sum_{i=1}^{n} f_i$) is present is called the median class. $L$ is the lower limit of the median class, $F_{m-1}$ is the cumulative frequency pre-median class, $f_m$ is the frequency of the median class, and $C$ is the class size.

$$Q_i = L + \frac{\frac{i \times N}{4} - F_{q-1}}{f_q} \times C \; ; \; i = 1,2,3$$

The class at which $\frac{i \times N}{4} - th$ frequency ($N = \sum_{i=1}^{n} f_i$) is present is called the $i - th$ quartile class. $L$ is the lower limit of the quartile class, $F_{q-1}$ is the cumulative frequency pre-quartile class, $f_q$ is the frequency of the quartile class, and $C$ is the class size.

$$D_i = L + \frac{\frac{i \times N}{10} - F_{d-1}}{f_d} \times C \; ; \; i = 1,2, \ldots \ldots, 9$$

The class at which $\frac{i \times N}{10} - th$ frequency ($N = \sum_{i=1}^{n} f_i$) is present is called the $i - th$ decile class. $L$ is the lower limit of the decile class, $F_{d-1}$ is the cumulative frequency pre-decile class, $f_d$ is the frequency of the decile class, and $C$ is the class size.

$$P_i = L + \frac{\frac{i \times N}{100} - F_{p-1}}{f_p} \times C \; ; \; i = 1,2, \ldots \ldots, 99$$

The class at which $\frac{i \times N}{100} - th$ frequency ($N = \sum_{i=1}^{n} f_i$) is present is called the $i - th$ percentile class. $L$ is the lower limit of the percentile class, $F_{p-1}$ is the cumulative frequency pre-percentile class, $f_p$ is the frequency of the percentile class, and $C$ is the class size.

$$AM - Mo = 3(AM - Me)$$

$$Mo = 3Me - 2Me$$

# STATISTICAL BASICS

## MEASURE OF DISPERSION

**Dispersion:** Dispersion in statistics is a way of describing how to spread out a set of data is. When a data set has a large value, the values in the set are widely scattered; when it is small the items in the set are tightly clustered.

The spread of a data set can be described by a range of descriptive statistics including Mean Deviation (MD), Standard Deviation (SD), and Interquartile Range. Those are called the absolute measures of dispersion. Also, there are some relative measures of dispersion, such as co-efficient of Mean Deviation, co-efficient of Standard Deviation, and co-efficient of Interquartile Range.

There are three types of Mean Deviation, estimated from Arithmetic Mean, Mode, and Median, respectively.

$$MD_{AM} = \frac{\sum_{i=1}^{n} f_i |x_i - AM|}{\sum_{i=1}^{n} f_i}$$

$$MD_{Mo} = \frac{\sum_{i=1}^{n} f_i |x_i - Mo|}{\sum_{i=1}^{n} f_i}$$

$$MD_{Me} = \frac{\sum_{i=1}^{n} f_i |x_i - Me|}{\sum_{i=1}^{n} f_i}$$

Co-efficient of Mean Deviation is

$$CMD = \frac{MD_{Base}}{Base} \times 100\% \; ; \; Base = AM, Mo, Me$$

Variance and Standard Deviation of the statistical data

$$V = \sigma^2 = \frac{\sum_{i=1}^{n} f_i (x_i - AM)^2}{\sum_{i=1}^{n} f_i}$$

$$SD = \sigma = \sqrt{\frac{\sum_{i=1}^{n} f_i (x_i - AM)^2}{\sum_{i=1}^{n} f_i}}$$

The working formula for Standard Deviation is

$$SD = \sqrt{\frac{\sum_{i=1}^{n} f_i {x_i}^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i}\right)^2}$$

## STATISTICAL BASICS

The coding formula for Standard Deviation is

$$SD = h \times \sqrt{\frac{\sum_{i=1}^{n} f_i {u_i}^2}{\sum_{i=1}^{n} f_i} - \left(\frac{\sum_{i=1}^{n} f_i u_i}{\sum_{i=1}^{n} f_i}\right)^2} \; ; \; u_i = \frac{x_i - a}{h}$$

Co-efficient of Standard Deviation is

$$CSD = \frac{SD}{AM} \times 100\%$$

Interquartile Range

$$IQR = Q_3 - Q_1$$

Co-efficient of Interquartile Range

$$CIQ = \frac{Q_3 - Q_1}{Q_3 + Q_1} \times 100\%$$

$$MD = \frac{4}{5} SD$$

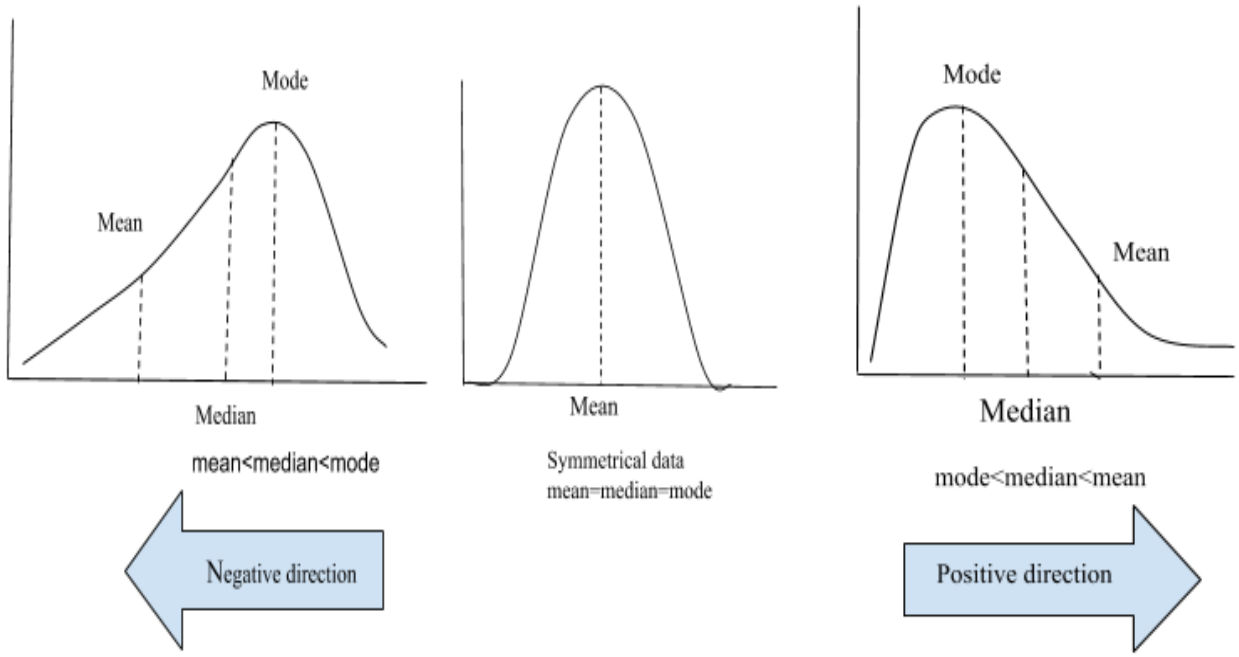
$$IQR = \frac{2}{3} SD$$

## MOMENTS, SKEWNESS AND KURTOSIS

# STATISTICAL BASICS

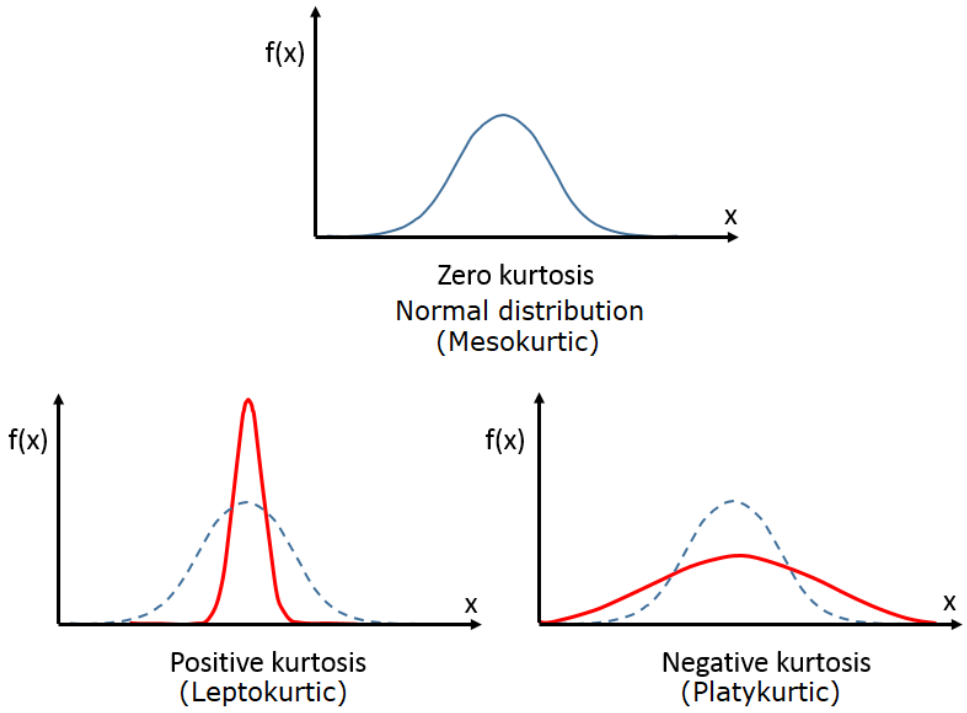
Two distributions may have the same Mean and Standard Deviation but may differ in their shape of the distribution. Further description of their characteristics is necessary that is provided by measures of skewness and kurtosis. Moments are popularly used to describe the characteristics of a distribution. They represent a convenient and unifying method for summarizing many of the most commonly used descriptive statistical measures such as central tendency, variation, Skewness, and Kurtosis.

The term ‘skewness’ refers to a lack of symmetry or departure from symmetry, e.g., when a distribution is not symmetrical (or is asymmetrical) it is called a skewed distribution. The measures of skewness indicate the difference between the manners in which the observations are distributed in a particular distribution compared with the symmetrical (or normal) distribution. The concept of skewness gains importance from the fact that statistical theory is often based upon the assumption of the normal distribution. A measure of skewness is, therefore, necessary in order to guard against the consequence of this assumption.




In statistics, kurtosis refers to the degree of flatness or peakedness in the region about the mode of a frequency curve. The degree of kurtosis of a distribution is measured relative to the flatness

or peakedness of a normal curve, it is called "Platykurtic" or "Leptokurtic". The normal curve itself is known as "Mesokurtic".




The $r-th$ raw moment about an arbitrary point $A$ is

$$m'_r = \frac{\sum_{i=1}^{n} f_i (x_i - A)^r}{\sum_{i=1}^{n} f_i} \; ; A \neq \bar{x}$$

The coding formula for the $r-th$ raw moment is

$$m'_r = h^r \times \frac{\sum_{i=1}^{n} f_i u_i^r}{\sum_{i=1}^{n} f_i} \; ; \; u_i = \frac{x_i - A}{h}$$

The $r-th$ central moment about the arithmetic mean $\bar{x}$ is

$$m_r = \frac{\sum_{i=1}^{n} f_i (x_i - \bar{x})^r}{\sum_{i=1}^{n} f_i}$$

Estimation of the central moments from the raw moments

$$m_1 = 0$$

# STATISTICAL BASICS

$$m_2 = m_2' - {m_1'}^2$$

$$m_3 = m_3' - 3m_2'm_1' + 2{m_1'}^3$$

$$m_4 = m_4' - 4m_3'm_1' + 6m_2'{m_1'}^2 - 3{m_1'}^4$$

Here, $m_2$ is the variance of the data set and hence $SD = \sqrt{m_2}$.

Co-efficient of Skewness

$$\gamma_3 = \frac{m_3}{\sqrt{m_2^3}}$$

If $\gamma_3 < 0$, the provided data set is called negatively skewed.

If $\gamma_3 = 0$, the provided data set is called non-skewed (Normal).

If $\gamma_3 > 0$, the provided data set is called positively skewed.

Co-efficient of Kurtosis

$$\gamma_4 = \frac{m_4}{m_2^2}$$

If $\gamma_4 < 3$, the provided data set is called platykurtic (flattered).

If $\gamma_4 = 3$, the provided data set is called mesokurtic (balanced).

If $\gamma_4 > 3$, the provided data set is called leptokurtic (peaked).

Corrected central momnets due to class size $c$ os chosen as the round figure

$$m_1^{(corrected)} = 0$$

$$m_2^{(corrected)} = m_2 - \frac{1}{12}c^2$$

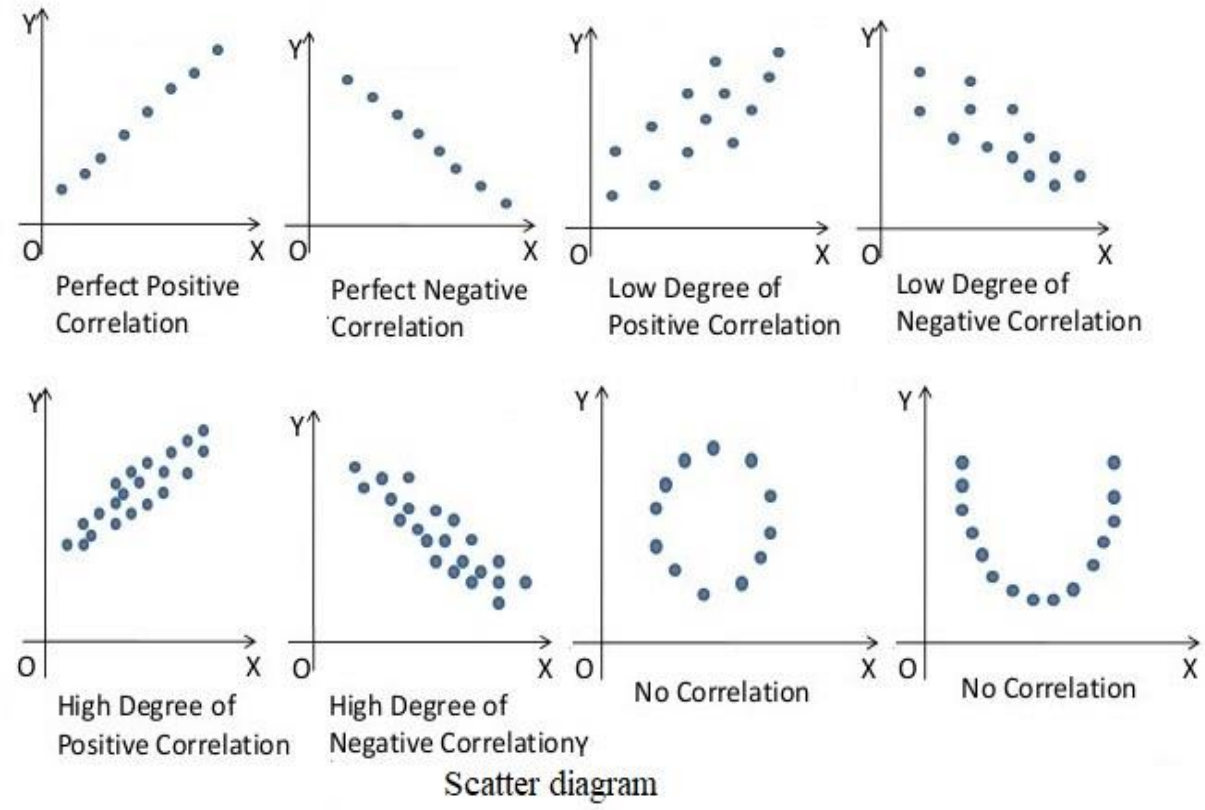
$$m_3^{(corrected)} = m_3$$

$$m_4^{(corrected)} = m_4 - \frac{1}{2}m_2c^2 - \frac{7}{240}c^4$$

# CORRELATION AND REGRESSION

# STATISTICAL BASICS

**Correlation Analysis:** Correlation analysis is applied in quantifying the association between two continuous variables, for example, a dependent and independent variable or among two independent variables.

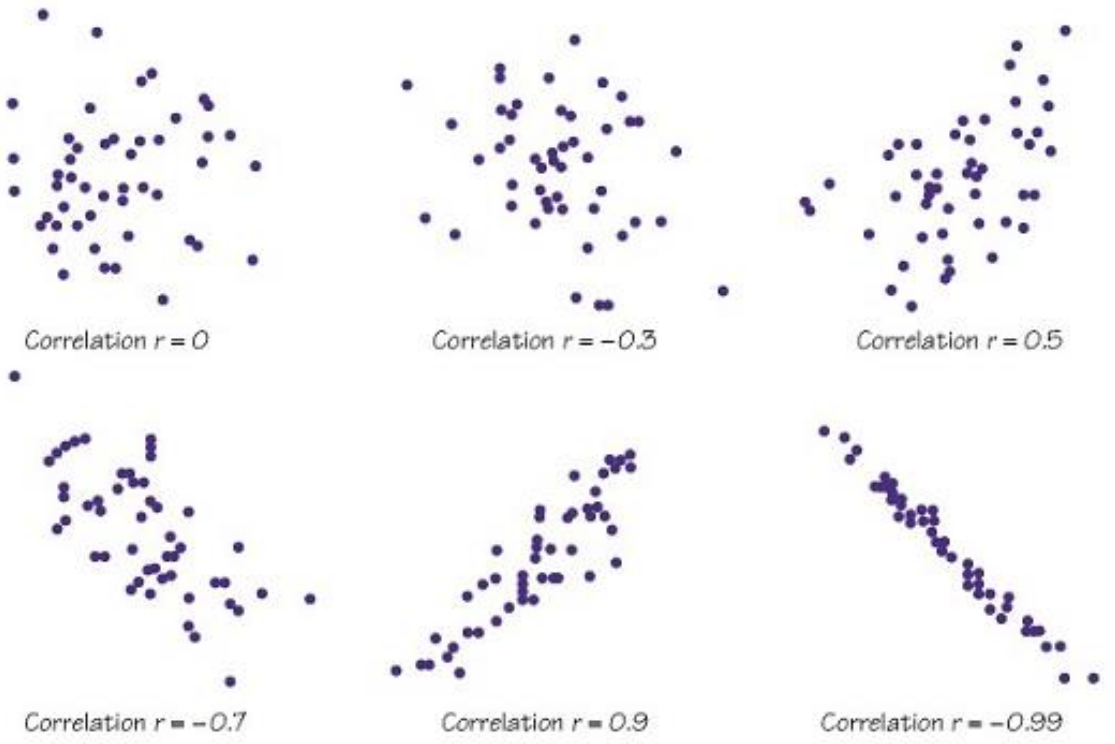


Scatter diagram

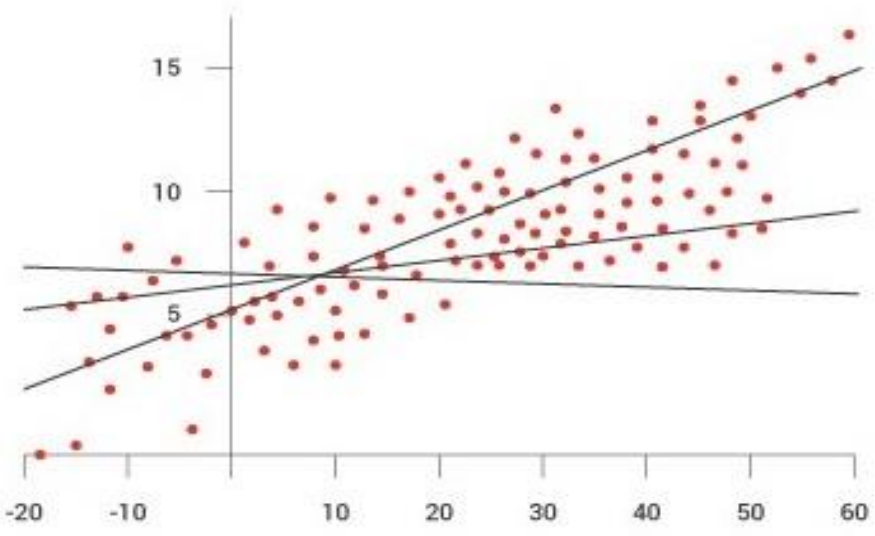
The sign of the coefficient of correlation shows the direction of the association. The magnitude of the coefficient shows the strength of the association. The sample of a correlation coefficient is estimated in the correlation analysis.

It ranges between $-1$ and $+1$, denoted by r and quantifies the strength and direction of the linear association among two variables. The correlation among two variables can either be positive, i.e., a higher level of one variable is related to a higher level of another or negative, i.e., a higher level of one variable is related to a lower level of the other.

# STATISTICAL BASICS

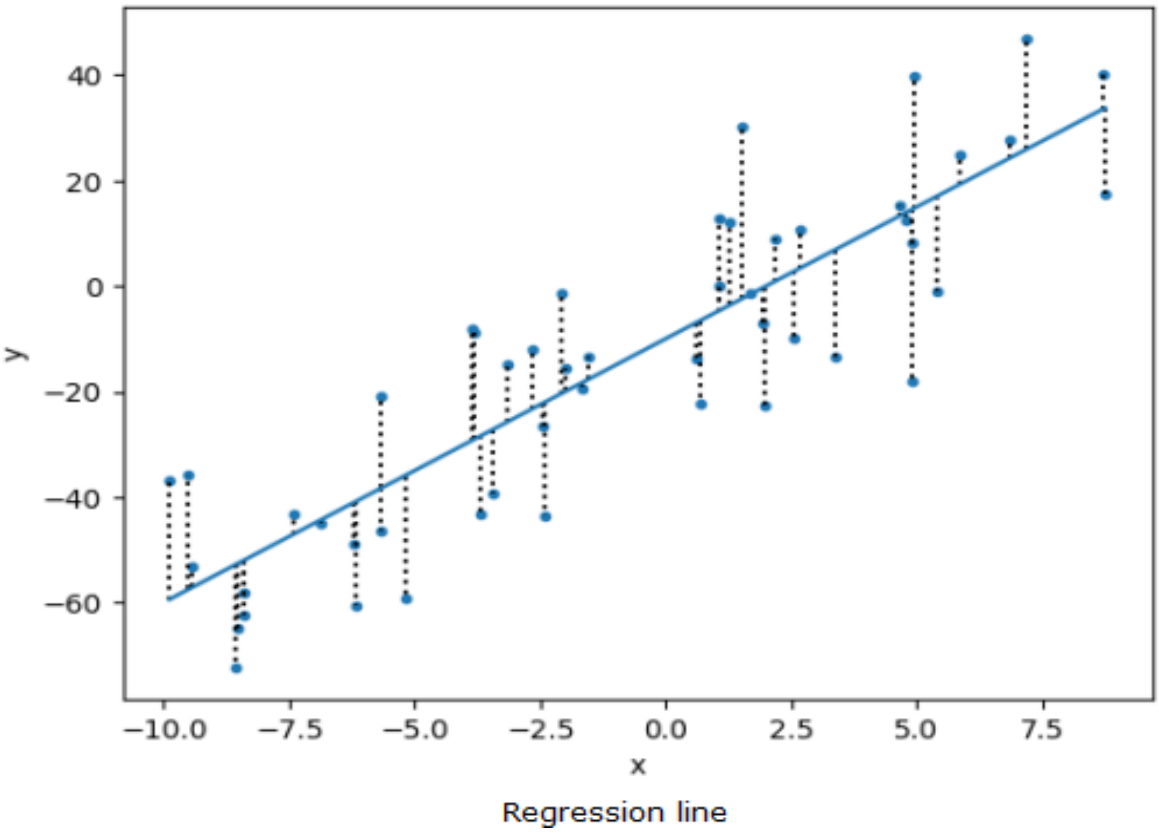



**Regression Analysis:** Regression analysis involves identifying the relationship between a dependent variable and one or more variables. The outcome variable is known as the dependent or response variable and the risk elements, and cofounders are known as predictors or independent variables. The dependent variable is shown by $\boldsymbol{y}$ and independent variables are shown by $\boldsymbol{x}$ in regression analysis.



# STATISTICAL BASICS

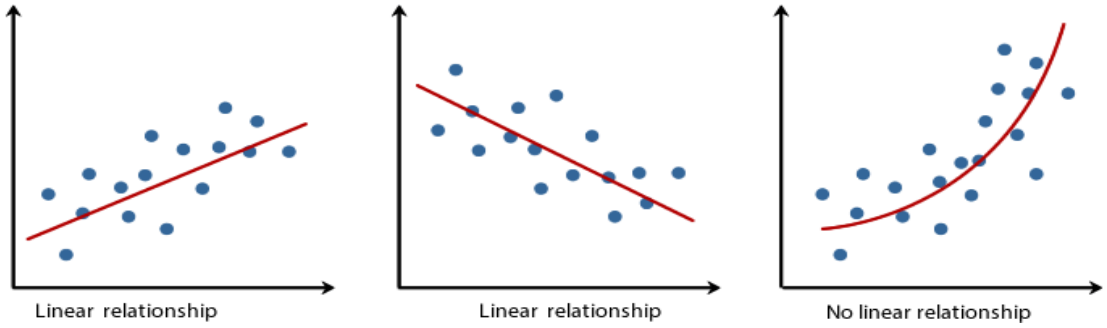
A model of the relationship is hypothesized, and estimates of the parameter values are used to develop an estimated regression equation. Various tests are then employed to determine if the model is satisfactory. If the model is deemed satisfactory, the estimated regression equation can be used to predict the value of the dependent variable given values for the independent variables.




**Linear Regression:** This is a linear approach to modeling the relationship between the scalar components and one or more independent variables. If the regression has one independent variable, then it is known as a simple linear regression. If it has more than one independent variable, then it is known as multiple linear regression.

Linear regression only focuses on the conditional probability distribution of the given values rather than the joint probability distribution. In general, all the real world regressions models involve multiple predictors. So, the term linear regression often describes multivariate linear regression.

# STATISTICAL BASICS




**Comparison between Correlation and Regression:**

| **Basis** | **Correlation** | **Regression** |
|---|---|---|
| Meaning | A statistical measure that defines co-relationship or association of two variables. | Describes how an independent variable is associated with the dependent variable. |
| Dependent and Independent variables | No difference | Both variables are different. |
| Usage | To describe a linear relationship between two variables. | To fit the best line and estimate one variable based on another variable. |
| Objective | To find a value expressing the relationship between variables. | To estimate the values of a random variable based on the values of a fixed variable. |

Correlation coefficient

$$r_{xy} = r_{yx} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}} = \frac{N\sum xy - \sum x \sum y}{\sqrt{(N\sum x^2 - (\sum x)^2)(N\sum y^2 - (\sum y)^2)}}$$

Here, $N$ is the number of inputs and $x$ & $y$ are two variables.

# STATISTICAL BASICS

Assume $\boldsymbol{u} = \frac{\boldsymbol{x}-\boldsymbol{a}}{\boldsymbol{h}}$ & $\boldsymbol{v} = \frac{\boldsymbol{y}-\boldsymbol{b}}{\boldsymbol{k}}$, where $\boldsymbol{a}$ & $\boldsymbol{b}$ are the shifts and $\boldsymbol{h}$ & $\boldsymbol{k}$ are the scales. Then it can be shown that

$$\boldsymbol{r_{xy} = r_{uv} = \frac{\sum(u-\bar{u})(v-\bar{v})}{\sqrt{\sum(u-\bar{u})^2 \sum(v-\bar{v})^2}} = \frac{N\sum uv - \sum u \sum v}{\sqrt{(N\sum u^2 - (\sum u)^2)(N\sum v^2 - (\sum v)^2)}}}$$

So, the Correlation coefficient is independent of the shift and scale.

There are two Regression co-efficient

$$\boldsymbol{b_{\frac{y}{x}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}}$$

$$\boldsymbol{b_{\frac{x}{y}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2} = \frac{N\sum xy - \sum x \sum y}{N\sum y^2 - (\sum y)^2}}$$

Again, for shifting and scaling data set

$$\boldsymbol{b_{\frac{y}{x}} = \frac{k}{h} \times b_{\frac{v}{u}} = \frac{k}{h} \times \frac{\sum(u-\bar{u})(v-\bar{v})}{\sum(u-\bar{u})^2} = \frac{k}{h} \times \left(\frac{N\sum uv - \sum u \sum v}{N\sum u^2 - (\sum u)^2}\right)}$$

$$\boldsymbol{b_{\frac{x}{y}} = \frac{h}{k} \times b_{\frac{u}{v}} = \frac{h}{k} \times \frac{\sum(u-\bar{u})(v-\bar{v})}{\sum(v-\bar{v})^2} = \frac{h}{k} \times \left(\frac{N\sum uv - \sum u \sum v}{N\sum v^2 - (\sum v)^2}\right)}$$

So, the Regression coefficients are independent of the shifts but dependent on scales.

Now, the Regression line of $\boldsymbol{y}$ on $\boldsymbol{x}$ is

$$\boldsymbol{y - \bar{y} = b_{\frac{y}{x}}(x-\bar{x})}$$
$$\boldsymbol{\Rightarrow\ y - \frac{\sum y}{N} = b_{\frac{y}{x}}\left(x - \frac{\sum x}{N}\right)}$$
$$\boldsymbol{\Rightarrow\ y = b_{\frac{y}{x}}\left(x - \frac{\sum x}{N}\right) + \frac{\sum y}{N}}$$

Similarly, the Regression line of $\boldsymbol{x}$ on $\boldsymbol{y}$ is

$$\boldsymbol{x - \bar{x} = b_{\frac{x}{y}}(y-\bar{y})}$$
$$\boldsymbol{\Rightarrow x - \frac{\sum x}{N} = b_{\frac{x}{y}}\left(y - \frac{\sum y}{N}\right)}$$

# STATISTICAL BASICS

$$\Rightarrow\ x = b_{\frac{x}{y}}\left(y - \frac{\sum y}{N}\right) + \frac{\sum x}{N}$$

Now, we have

$$b_{\frac{y}{x}} \times b_{\frac{x}{y}} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \times \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(y-\bar{y})^2} = \frac{(\sum(x-\bar{x})(y-\bar{y}))^2}{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2} = r_{xy}^2$$

$$\Rightarrow r_{xy} = \pm\sqrt{b_{\frac{y}{x}} \times b_{\frac{x}{y}}}$$

There are some well-known notations

$$\sigma_x^2 = \frac{\sum(x-\bar{x})^2}{N} = S_{xx}$$

$$\sigma_y^2 = \frac{\sum(y-\bar{y})^2}{N} = S_{yy}$$

$$\sigma_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{N} = S_{xy}$$

Here, $S_{xy}$ is called the covariance between $x$ & $y$ and denoted by $Cov(x, y)$.

Then, the above notations can be applied to find the following.

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$b_{\frac{y}{x}} = \frac{\sigma_{xy}}{\sigma_x^2}$$

$$b_{\frac{x}{y}} = \frac{\sigma_{xy}}{\sigma_y^2}$$

So, we can find

$$b_{\frac{y}{x}} = r_{xy} \times \frac{\sigma_y}{\sigma_x}$$

$$b_{\frac{x}{y}} = r_{xy} \times \frac{\sigma_x}{\sigma_y}$$

# STATISTICAL BASICS

**Rank Correlation:** Sometimes there doesn't exist a marked linear relationship between two random variables but a monotonic relation (if one increases, the other also increases or instead, decreases) is clearly noticed.

If instead of measuring the correlation between two sets of continuous random variables $\boldsymbol{x}$ & $\boldsymbol{y}$ we replace their numerical values by their rankings, then we obtain the Rank Correlation coefficient. The rank of the $\boldsymbol{i-th}$ element of a sample of size $\boldsymbol{N}$ is equal to the index of the order statistic.

The Rank Correlation coefficient

$$\boldsymbol{r_{rank} = 1 - \frac{6\sum d_i^2}{N(N^2-1)}\ ;\ d_i = x_i - y_i}$$