# Data Analytics

Course Taught at IIFT

***Session 5: Correlation Analysis***

**Dr. Tanujit Chakraborty**

www.ctanujit.org

# Today's Topics……

- Introduction
- Measures of Relationship
- Correlation Analysis
  - $\chi^2$ - Test
  - Spearman's Correlation Analysis
  - Pearson's Correlation Analysis

# Relationship Analysis

- **Example: Wage Data**

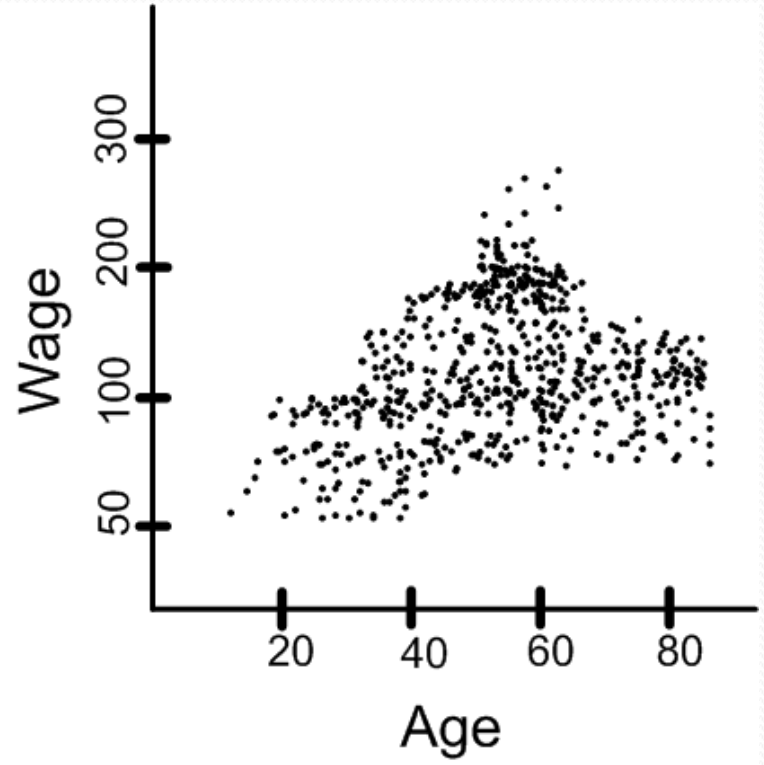
  A large data regarding the wages for a group of employees from the eastern region of India is given.

  In particular, we wish to understand the following relationships:

  - *Employee's age and wage:* How wages vary with ages?
  - *Calendar year and wage:* How wages vary with time?
  - *Employee's age and education:* Whether wages are anyway related with employees' education levels?

# Relationship Analysis

- Example: Wage Data
  - Case I. Wage versus Age
    - From the data set, we have a graphical representations, which is as follows:

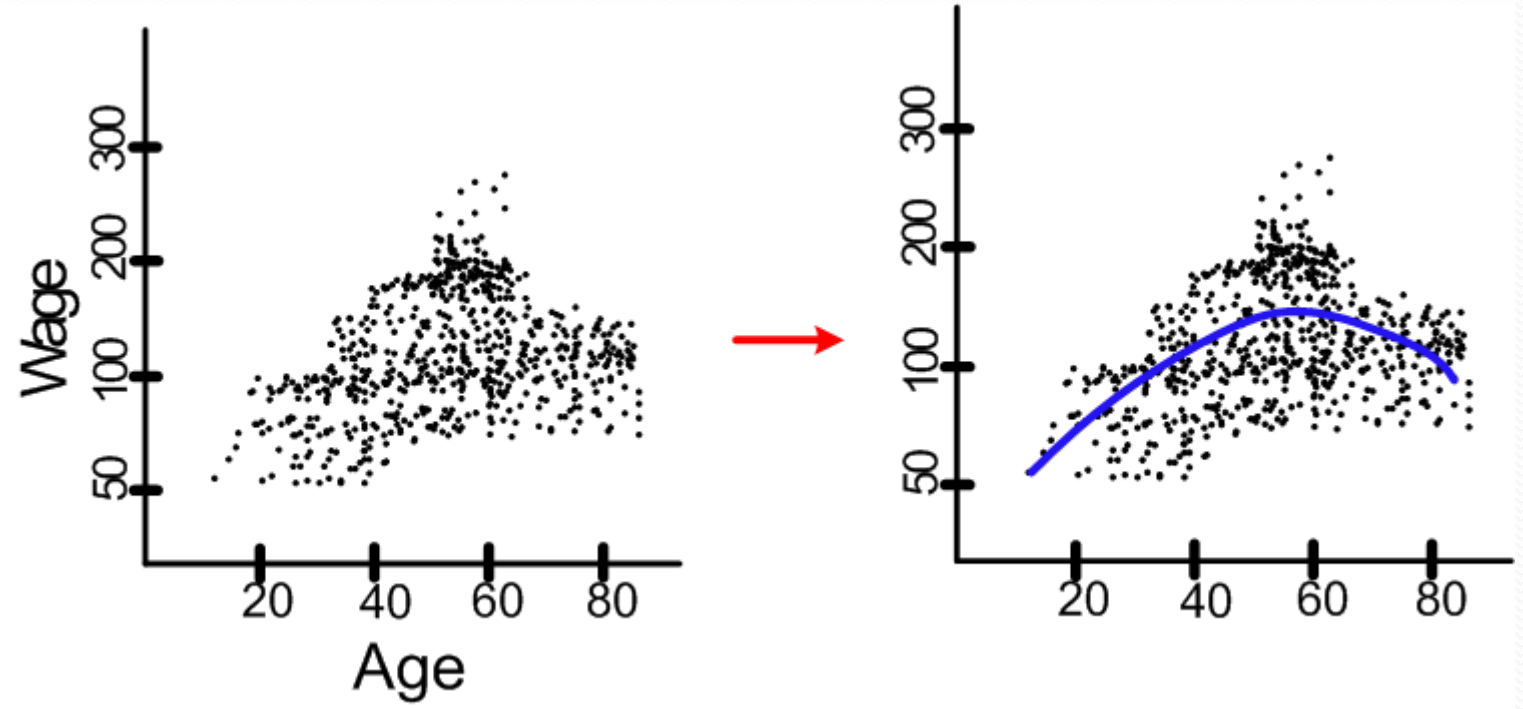



?

How wages vary with ages?

# Relationship Analysis

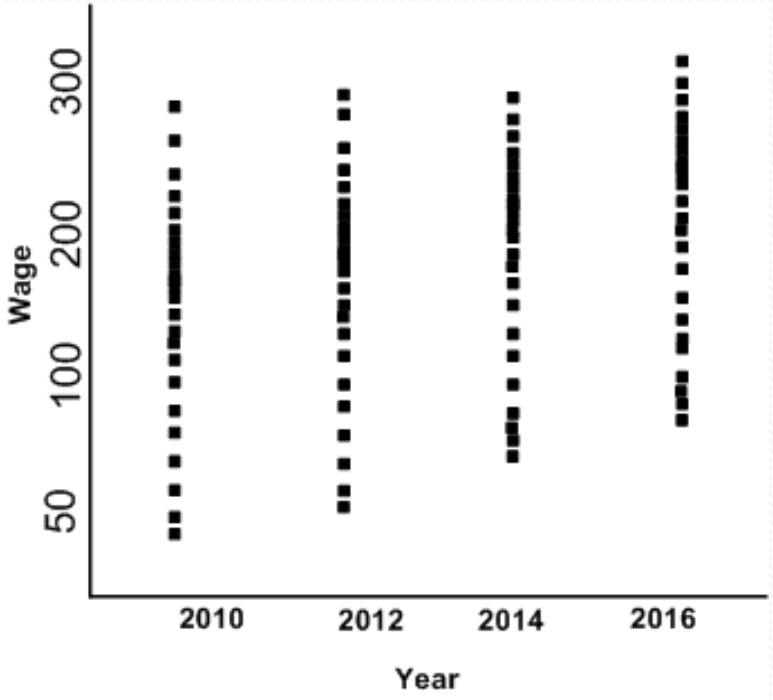
- Example: Wage Data
  - *Employee's age and wage:* How wages vary with ages?




Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

# Relationship Analysis

- Example: Wage Data
  - Case II. Wage versus Year
    - From the data set, we have a graphical representations, which is as follows:

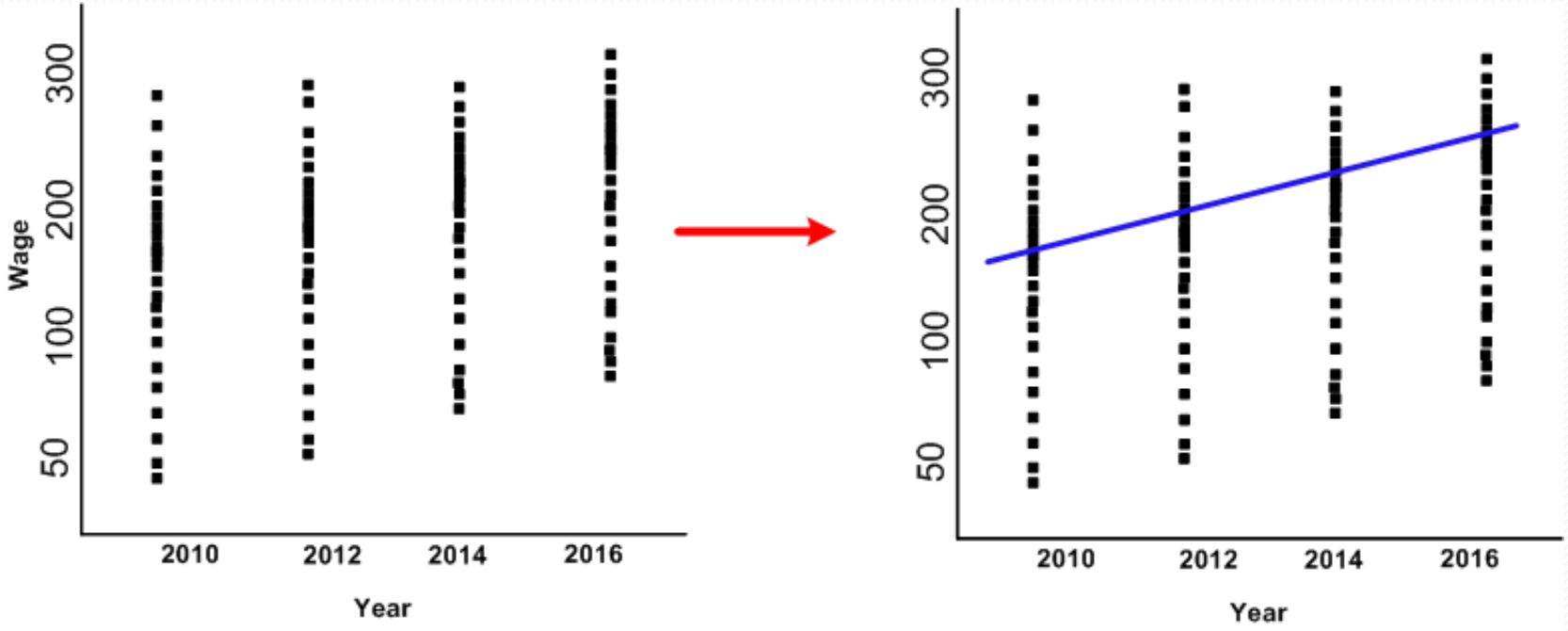



?

How wages vary with time?

# Relationship Analysis

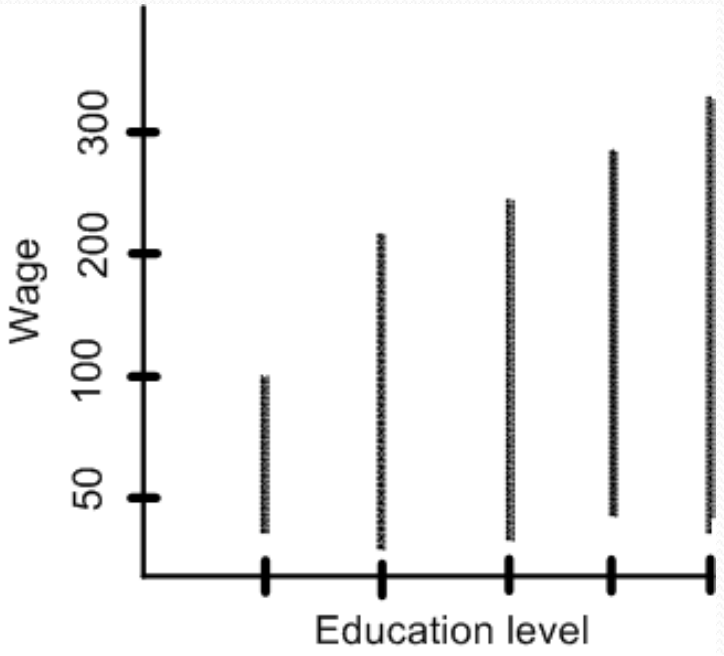
- Example: Wage Data
  - *Wage and calendar year:* How wages vary with years?




Interpretation: There is a slow but steady increase in the average wage between 2010 and 2016.

# Relationship Analysis

- Example: Wage Data
  - Case III. Wage versus Education
    - From the data set, we have a graphical representations, which is as follows:

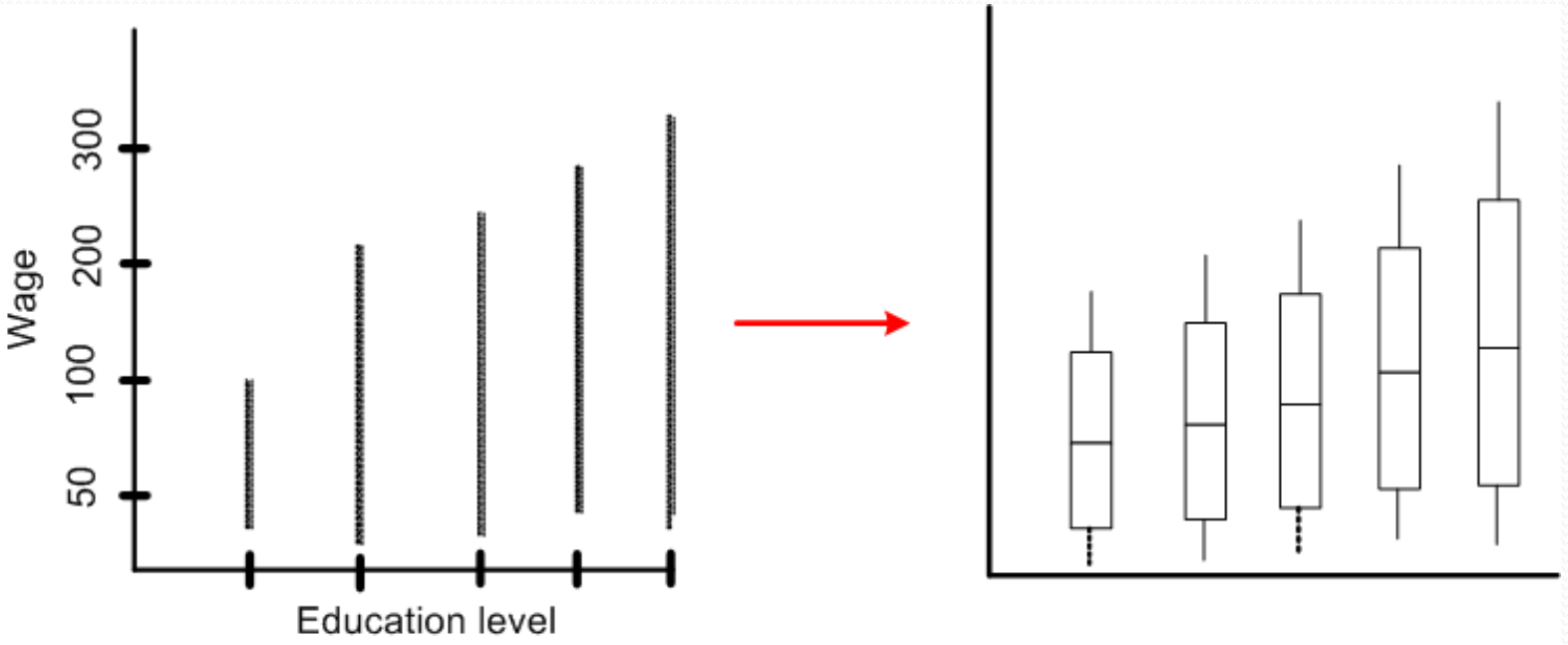



?

Whether wages are related with education?

# Relationship Analysis

- Example: Wage Data
  - *Wage and education level:* Whether wages vary with employees' education levels?




Interpretation: On the average, wage increases with the level of education.

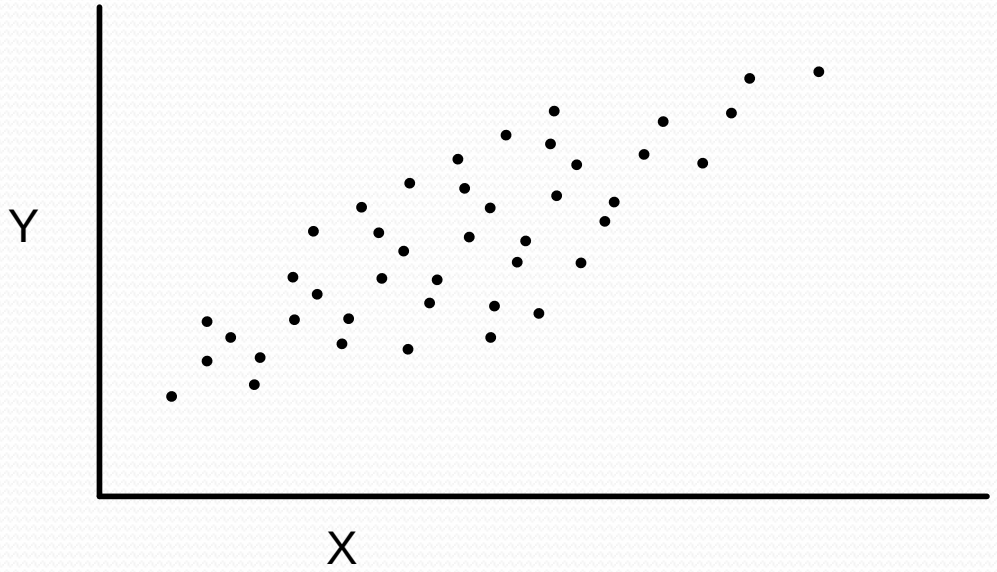
# Relationship Analysis

Given an employee's wage can we predict his age?

Whether wage has any association with both year and education level?

etc….

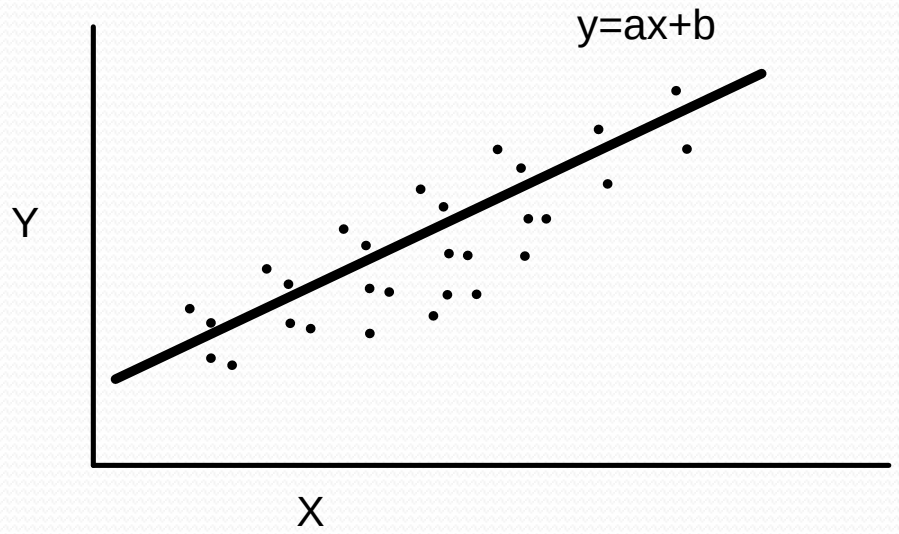
# Question for You!




Suppose there are countably infinite points in the *XY plane*. We need a huge memory to store all such points.

**Is there any way out to store this information with a least amount of memory?**

# Solution:




Just decide the values of **a** and **b**

(as if storing one point's data only!)

Note: Here, the trick was to find a relationship among all the points.

# Measures of Relationship

*Univariate population:* The population consisting of only one variable.

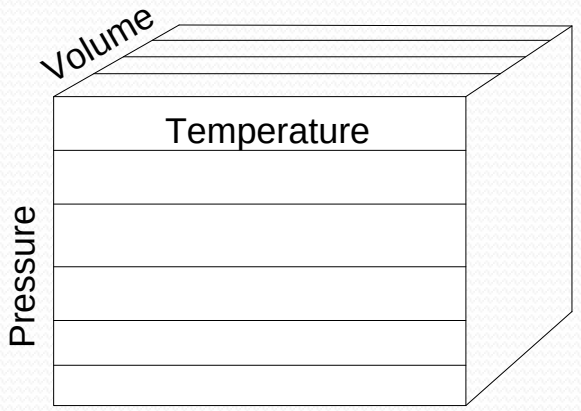
| *Temperature* | 20 | 30 | 21 | 18 | 23 | 45 | 52 |
|---|---|---|---|---|---|---|---|

Here, statistical measures are suffice to find a relationship.

*Bivariate population:* Here, the data happen to be on two variables.

| *Pressure* | 1 | 1.1 | ... ... | 0.8 |
|---|---|---|---|---|
| *Temperature* | 35 | 41 | | 29 |

*Multivariate population:* If the data happen to be one more than two variable.



# Measures of Relationship

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist correlation (i.e., association) between two (or more) variables?

If yes, of what degree?

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or one variable in one side and two or more variables on the other side (in case of multivariate population)?

If yes, of what degree and in which direction?

To find solutions to the above questions, two approaches are known.

- **Correlation Analysis**
- **Regression Analysis**

# Correlation Analysis

# Correlation Analysis

- In statistics, the word correlation is used to denote some form of association between two variables.
  - Example: **Weight** is correlated with **height**

| $A$ | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | … … |
|---|---|---|---|---|---|---|---|
| $B$ | $b_1$ | $b_2$ | $b_3$ | $b_4$ | $b_5$ | $b_6$ | … … |

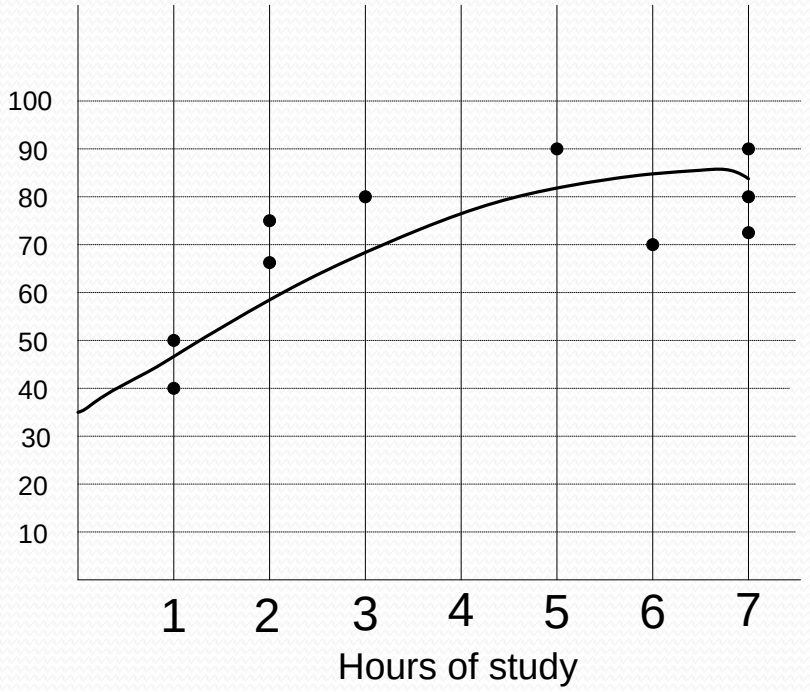
The correlation may be positive, negative or zero.

- **Positive correlation:** If the value of the attribute *A* increases with the increase in the value of the attribute *B* and vice-versa.
- **Negative correlation:** If the value of the attribute *A* decreases with the increase in the value of the attribute *B* and vice-versa.
- **Zero correlation:** When the values of attribute *A* varies at random with *B* and vice-versa.

# Correlation Analysis

- In order to measure the degree of correlation between two attributes.

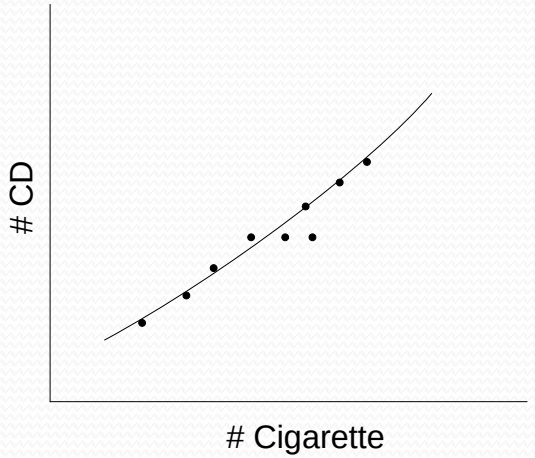
| *Hours Study* | *Exam Score* |
|---|---|
| 3 | 80 |
| 5 | 90 |
| 2 | 75 |
| 6 | 80 |
| 7 | 90 |
| 1 | 50 |
| 2 | 65 |
| 7 | 85 |
| 1 | 40 |
| 7 | 100 |



# Correlation Analysis

- Do you find any correlation between *X* and *Y* as shown in the table?.

| *No. of CD's sold in shop X* | **25** | **30** | **35** | **42** | **48** | **52** | **56** |
|---|---|---|---|---|---|---|---|
| *No. of cigarette sold in Y* | 5 | 7 | 9 | 10 | 11 | 11 | 12 |

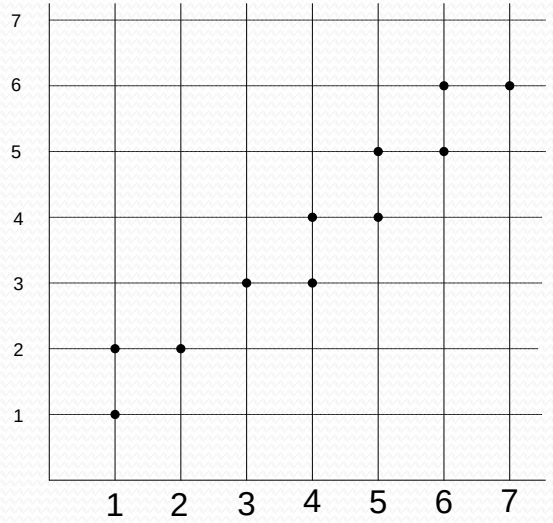



**Note:**

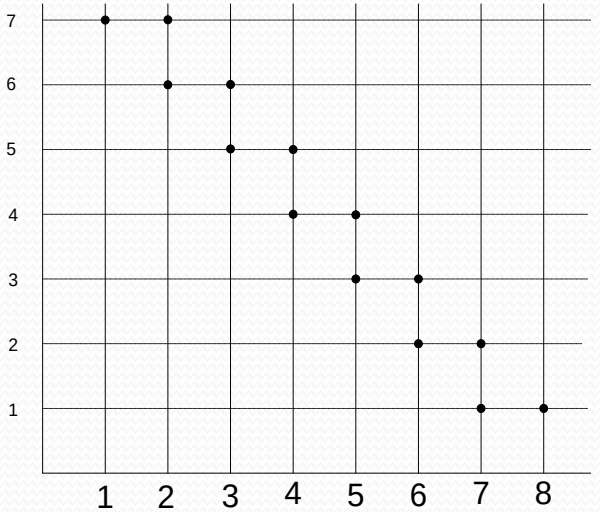
- Correlation does not always imply Causation.
- In data analytics, correlation analysis make sense only when relationship make sense. There should be a cause-effect relationship.

# Correlation Analysis

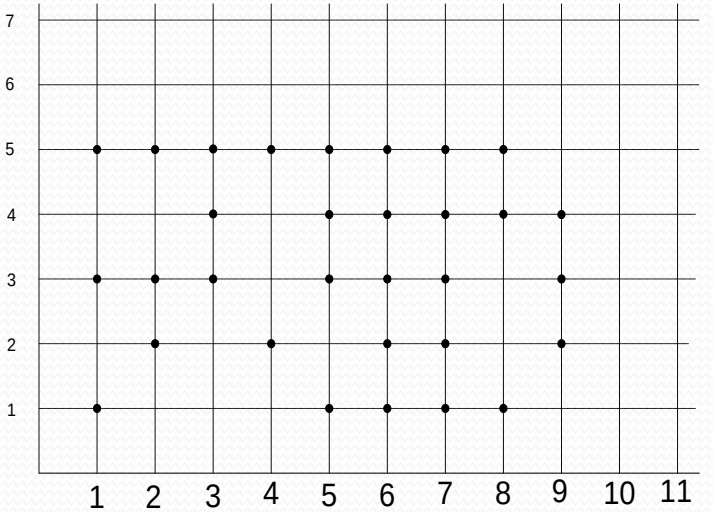
Positive correlation




Negative correlation




Zero correlation



# Correlation Coefficient

- **Correlation coefficient is used to measure the degree of association.**
- It is usually denoted by *r*.
- The value of *r* lies between +1 and -1.
- Positive values of *r* indicates positive correlation between two variables, whereas, negative values of *r* indicate negative correlation.
- $r = +1$ implies perfect positive correlation, and otherwise.
- The value of *r* nearer to +1 or -1 indicates high degree of correlation between the two variables.
- *r* = 0 implies, there is no correlation

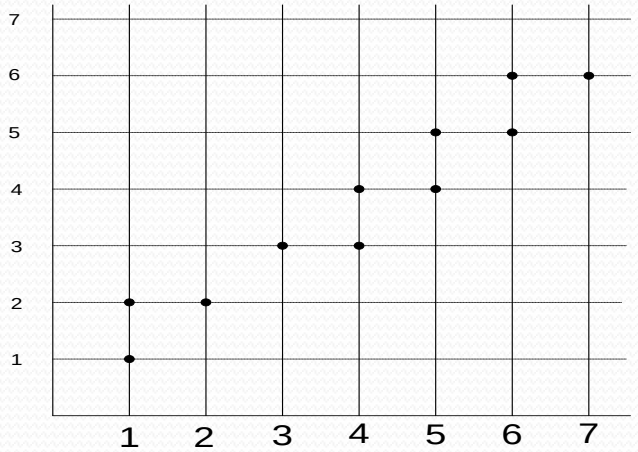
# Correlation Coefficient

**Note:** **The Correlation Coefficient** is a measure of the degree of linear relation between variables.

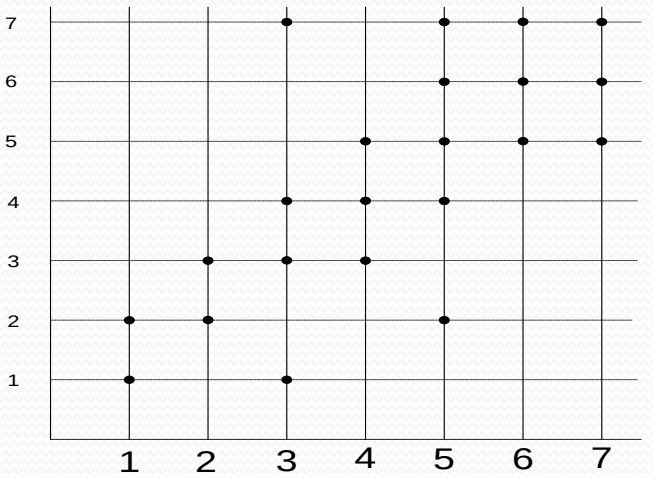
- **A simple correlation coefficient** shows the degree of linear relation between two variables.
- **A multiple correlation coefficient** shows the degree of linear relation between more than two variables. It is a generalization of the simple correlation coefficient. Thus a simple correlation coefficient is a special case of multiple correlation coefficient.

# Correlation

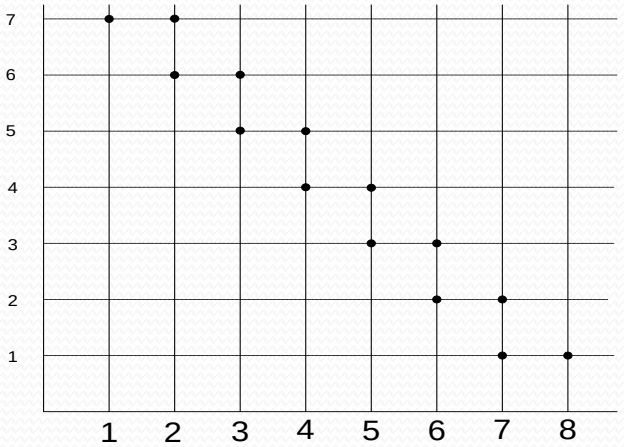
High Positive Correlation



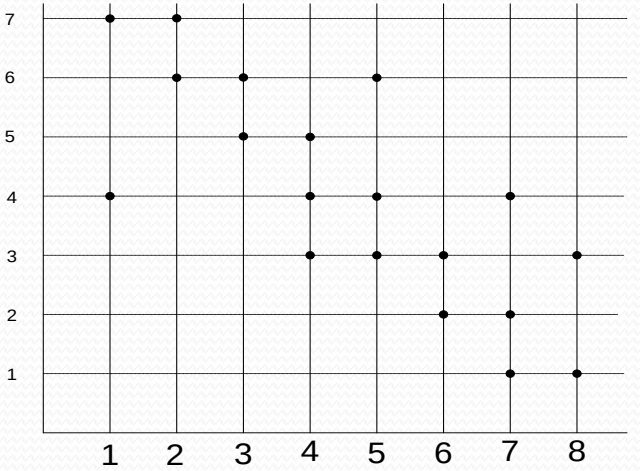

Low Positive Correlation



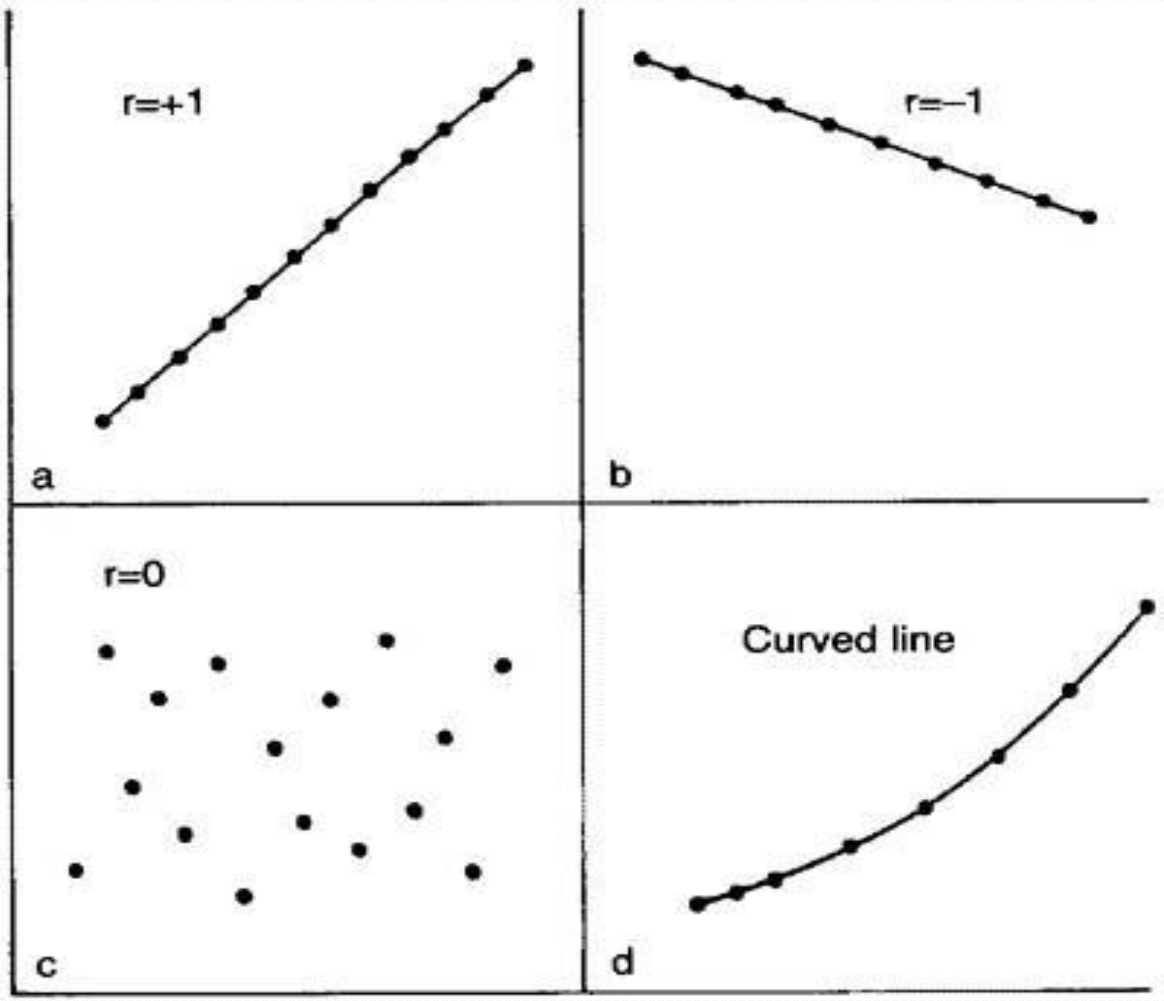

High Negative Correlation




Low Negative Correlation



# Correlation Coefficient

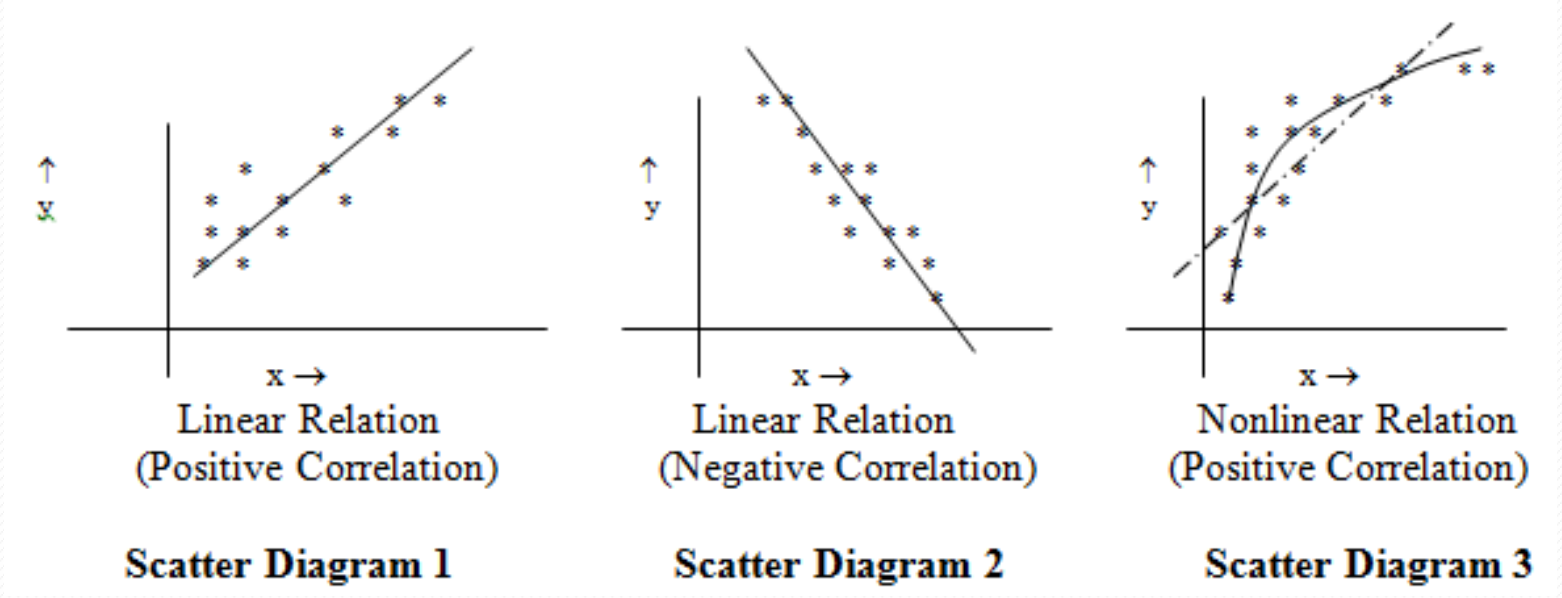


# Data

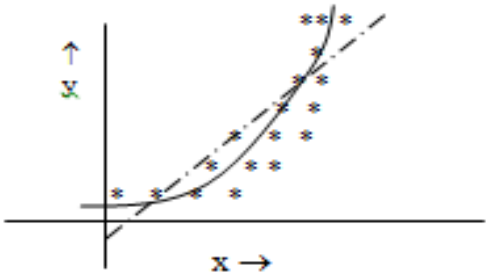
- Let us assume that we have a random sample of size n on two variables namely x and y. The sampled values are $\{(x_1,y_1), (x_2,y_2), \ldots, (x_n,y_n)\}$.
- It is necessary to have ordered pairs if we want to find relations between x and y; i.e., each observation consists of two values – one on x and the other on y, i-th observation being $(x_i,y_i)$, $i = 1,2, \ldots, n$.
- It may sometime be necessary to assume that the sample observations are coming from continuous distribution.

# Scatter or dot diagram

- The first step to get an idea about whether there exists any relation between x and y and if exists, the degree of relation, is to draw **scatter** or **dot diagram**.
- The scatter diagram is nothing but the set of n points of (x,y) pairs shown on a graph paper. The following are some examples of scatter diagrams:

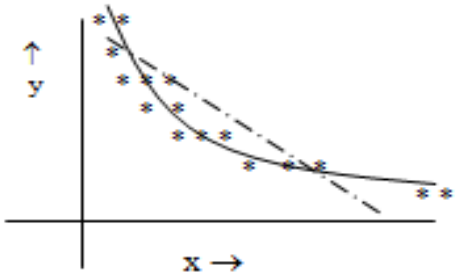


# Scatter or dot diagram



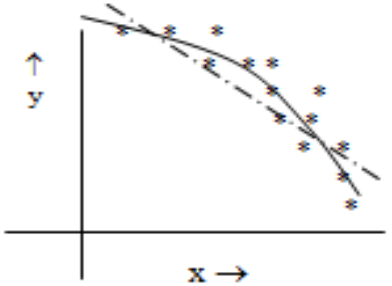

Nonlinear Relation
(Positive Correlation)

**Scatter Diagram 4**



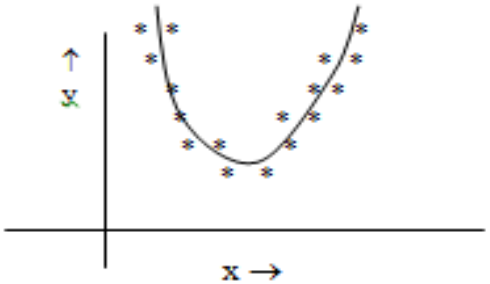

Nonlinear Relation
(Negative Correlation)

**Scatter Diagram 5**



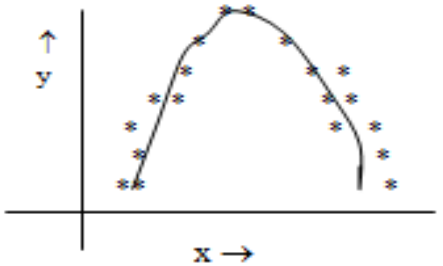

Nonlinear Relation
(Negative Correlation)

**Scatter Diagram 6**



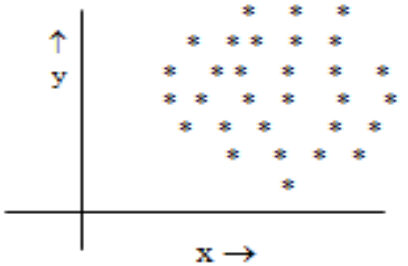

Nonlinear Relation
(Zero Correlation)

**Scatter Diagram 7**




Nonlinear Relation
(Zero Correlation)

**Scatter Diagram 8**




No Relation
(Zero Correlation)
*(Independent)*

**Scatter Diagram 9**

# Explaining the Scatter Diagrams

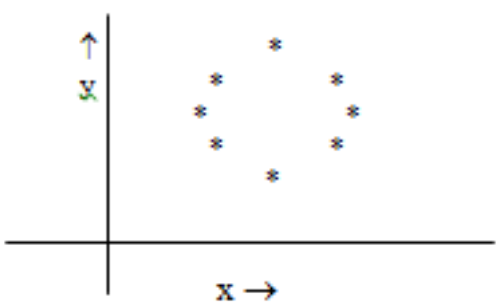
- The scatter diagrams were drawn assuming that x-values are in the direction of x-axis and y-values are in the direction of y-axis. It can be seen from the above diagrams that the first two diagrams depict the linear relationships between x and y. Diagram 3 through 8 depict nonlinear relationships. The last diagram shows "no relationship" between x and y.

- If we want to find out the degree of linear relationship between x and y, the first two cases will be the ideal ones. Though the Diagrams 3 through 6 show nonlinearity, it is possible to approximate the relationship by a straight line, though it may not be appropriate to do so.

- However, Diagrams 7 and 8 do not allow us to approximate it by a straight line. If we want to draw the best straight line going through the dots, it will either become a horizontal line (i.e., the line which is horizontal to x-axis) or vertical line (i.e., the line which is vertical to x-axis).

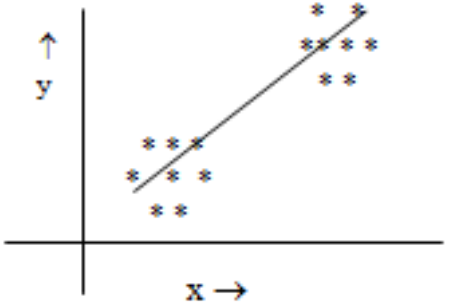
- The diagram 9 shows neither linear nor nonlinear relation between x and y. In this case, we say that x and y are independent. Any straight line going through the Center of Gravity of the dots is the best line.

# Some special cases of the scatter diagrams



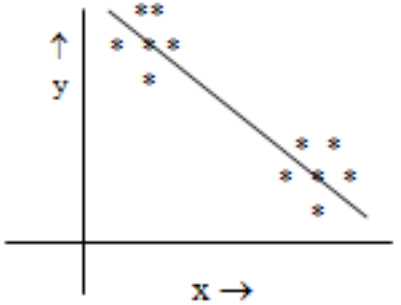

Nonlinear Relation
(Zero Correlation)

**Scatter Diagram 10**



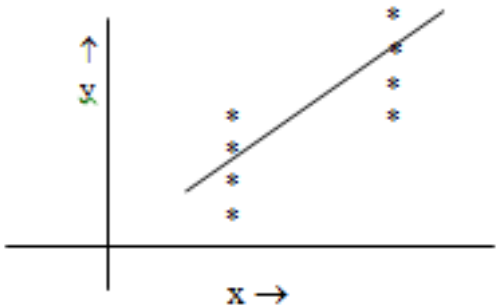

(Positive Correlation)

**Scatter Diagram 11**



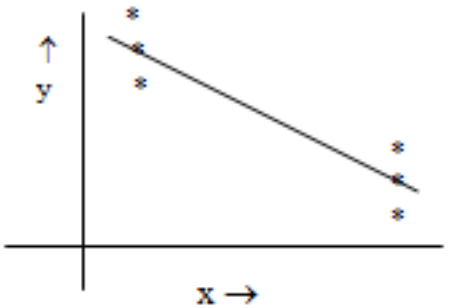

(Negative Correlation)

**Scatter Diagram 12**




(Positive Correlation)

**Scatter Diagram 13**




(Negative Correlation)

**Scatter Diagram 14**

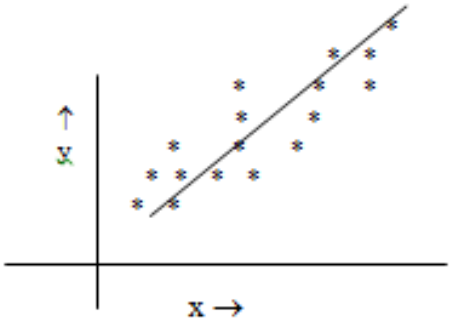
# Explanation of the Special Cases

- Diagram 10 shows a scatter plot with dots situated on the border of a circle. This is another case where there is a nonlinear relationship between x and y which can not be approximated by a straight line. Any straight line going through the center may be considered to be the 'best' approximating straight line.
- The diagrams 11 and 12 have clusters of points at two places. Thus they can be approximated by straight lines which go through the centers of two clusters of points.
- Diagrams 13 and 14 are similar to the diagrams 11 and 12 except that instead of clusters of points the x-values are concentrated at two points. Again the straight line approximations are possible.
- Observe that in the last four diagrams the basic assumptions that x and y values were drawn from continuous distributions are clearly violated.
- In the cases of diagrams 11 and 12, the random variable x (and the random variable y) can take values in one of two possible intervals whereas in the cases of diagrams 13 and 14, the random variables x can take only one of two possible values. Thus the straight line approximations to these cases are not meaningful enough.

# Positive, negative and zero correlations

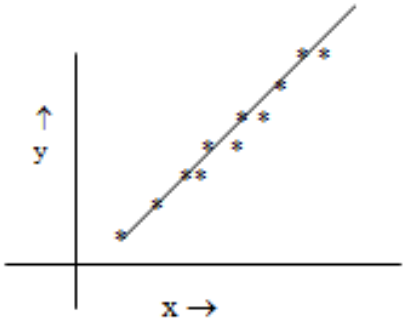
- Two random variables are positively related if x and y more or less move in the same direction, i.e., if the value of x increases then the value of y also tends to increase and if the value of x decreases then the value of y also tend to decrease.
- Similarly x and y are negatively related if x and y more or less move in the opposite directions.
- In other words if it is possible to approximate the relation by a straight line, it will either have a positive slope or have a negative slope depending on whether x and y are respectively positively or negatively related.
- The positive and negative relations are defined with respect to straight line relation or approximate straight line relation. The straight line relation between x and y is known as "**correlation**".
- The scatter diagrams 1, 3, 4, 11, and 13 show positive correlation and the scatter diagrams 2, 5, 6, 12 and 14 show negative correlation. All other scatter diagrams show no or zero correlation.
- **Zero correlation** does not mean that x and y are not related. It only means that x and y are not linearly related or can not be approximated by a linear relation. There may be nonlinear relation between x and y. Diagram 9 shows that there can neither be linear nor be nonlinear relations between x and y and thus the two random variables x and y are called independent of each other. Independence between random variables implies zero correlation but zero correlation may not imply independence.

# The degree of linear relation



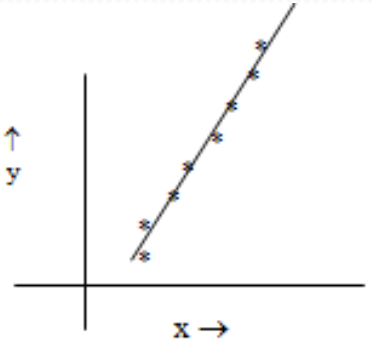

Linear Relation
(Positive Correlation)

**Scatter Diagram 15**



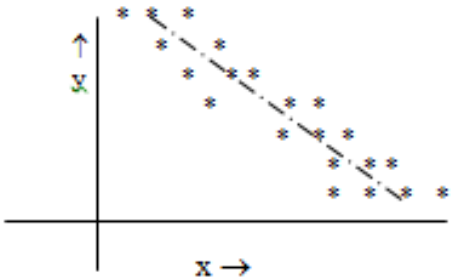

Linear Relation
(Positive Correlation)

**Scatter Diagram 16**



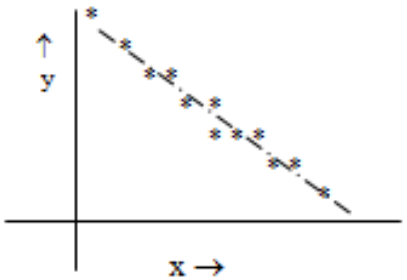

Linear Relation
(Positive Correlation)

**Scatter Diagram 17**



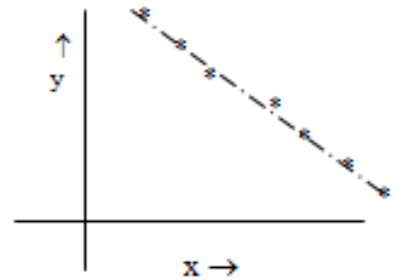

Linear Relation
(Negative Correlation)

**Scatter Diagram 18**




Linear Relation
(Negative Correlation)

**Scatter Diagram 19**




Linear Relation
(Negative Correlation)

**Scatter Diagram 20**

# Explaining the degree of linear relation

- All the diagrams 15, 16 and 17 show positive correlation between x and y.
- If we compare diagrams 15 and 16, we see that the scatter points are more close to the straight line in diagram 16.
- Thus we say that the degree of correlation (in this case positive correlation) is more in diagram 16 than in diagram 15.
- The extreme situation is diagram 17. In this case all dots are on the straight line and the degree of correlation is highest.
- Similar is the case for diagrams 18, 19 and 20. But here we have negative correlation between x and y.

# Measuring Correlation Coefficients

- There are three methods known to measure the correlation coefficients
  - Karl Pearson's coefficient of correlation
    - This method is applicable to find correlation coefficient between two **numerical** attributes
  - Charles Spearman's coefficient of correlation
    - This method is applicable to find correlation coefficient between two **ordinal** attributes
  - Chi-square coefficient of correlation
    - This method is applicable to find correlation coefficient between two **categorical** attributes

# Karl Pearson's Correlation Coefficient

- This is also called **Pearson's Product Moment Correlation**

Definition : **Karl Pearson's correlation coefficient**

Let us consider two attributes are $X$ and $Y$.

The Karl Pearson's coefficient of correlation is denoted by $r^*$ and is defined as

$$r^* = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sigma_X . \sigma_Y}$$

where $X_i = \text{i} - \text{th value of } X - \text{variable}$

$\bar{X} = \text{mean of } X$

$Y_i = \text{i} - \text{th value of } Y - \text{variable}$

$\bar{Y} = \text{mean of } Y$

$n = \text{number of pairs of observation of } X \text{ and } Y$

$\sigma_X = \text{standard deviations of } X$

$\sigma_Y = \text{standard deviation of } Y$

# Karl Pearson's coefficient of Correlation

**Example :** Correlation of Gestational Age and Birth Weight

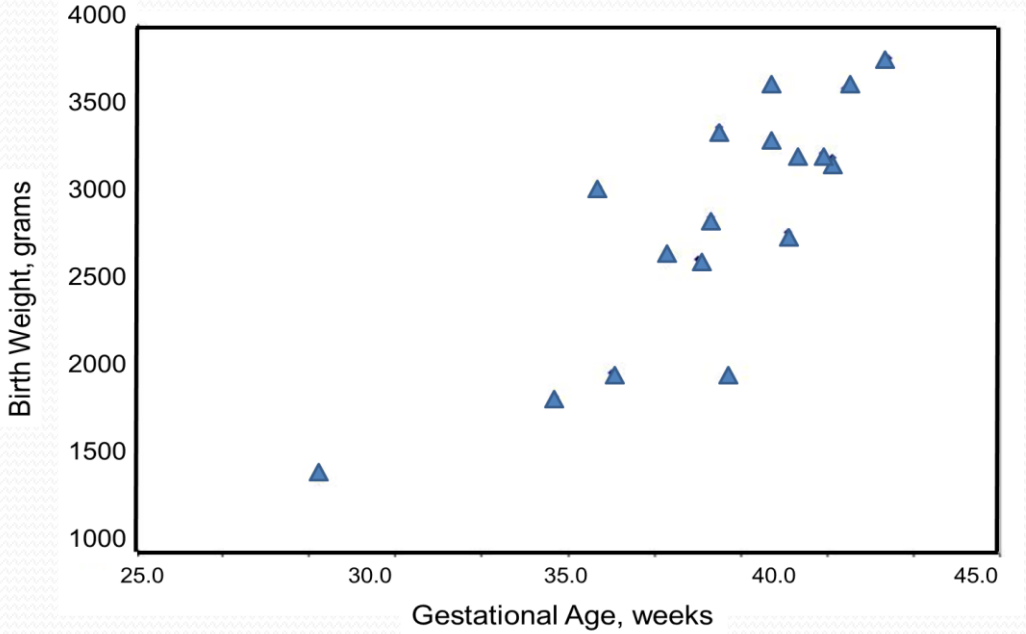
- A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

| Infant ID # | Gestational Age (wks) | Birth Weight (gm) |
|---|---|---|
| 1 | 34.7 | 1895 |
| 2 | 36.0 | 2030 |
| 3 | 29.3 | 1440 |
| 4 | 40.1 | 2835 |
| 5 | 35.7 | 3090 |
| 6 | 42.4 | 3827 |
| 7 | 40.3 | 3260 |
| 8 | 37.3 | 2690 |
| 9 | 40.9 | 3285 |
| 10 | 38.3 | 2920 |
| 11 | 38.5 | 3430 |
| 12 | 41.4 | 3657 |
| 13 | 39.7 | 3685 |
| 14 | 39.7 | 3345 |
| 15 | 41.1 | 3260 |
| 16 | 38.0 | 2680 |
| 17 | 38.7 | 2005 |

# Karl Pearson's coefficient of Correlation

**Example :** Correlation of Gestational Age and Birth Weight

- We wish to estimate the association between gestational age and infant birth weight.
- In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus $Y$ = birth weight and $X$ = gestational age.
- The data are displayed in a scatter diagram in the figure below.



# Karl Pearson's coefficient of Correlation

**Example :** Correlation of Gestational Age and Birth Weight

- For the given data, it can be shown the following

$$\overline{X} = \frac{\Sigma X}{n} = \frac{652.1}{17} = 38.4.$$

$$\overline{Y} = \frac{\Sigma Y}{n} = \frac{49{,}334}{17} = 2902.$$

$$s_x^2 = \frac{\Sigma(X - \overline{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

$$s_y^2 = \frac{\Sigma(Y - \overline{Y})^2}{n-1} = \frac{7{,}767{,}660}{16} = 485{,}578.8.$$

$$r^* = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{s_x . s_y} = 0.82$$

**Conclusion:** The sample's correlation coefficient indicates a strong positive correlation between Gestational Age and Birth Weight.

# Karl Pearson's coefficient of Correlation

**Example :** Correlation of Gestational Age and Birth Weight

- Significance Test
  - To test whether the association is merely apparent, and might have arisen by chance use the *t* test in the following calculation

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

  - Number of pair of observation is 17. Hence,

$$t = 0.82\sqrt{\frac{17-2}{1-0.82^2}} = 1.44$$

  - Consulting the t-test table, at degrees of freedom 15 and for $\alpha = 0.05$, we find that t = 1.753. Thus, the value of Pearson's correlation coefficient in this case may be regarded as highly significant.

# Properties of Correlation coefficient

- The correlation coefficient is not only invariant under changes of unit of measurements but also unaffected by changes of origin for both variables. I.e., if all x-values or y-values are added or subtracted by the same constant then the value of the correlation coefficient remains unchanged.
- The above properties (invariance under changes of origin and scale) can be summarized by saying that the correlation coefficient is invariant under linear transformation of x and y (except for the sign).
- We have already seen that if x and y are positively/negatively related then the value of r will be positive/negative. r has other good properties. The value of r always lies in-between –1 and +1. r takes the value +1 when all the values are on the positively sloped straight line and the value –1 when all the points are on the negatively sloped straight line.
- As the scatter points move closer to the (hypothetical) straight line, the value of |r| moves to 1. As the points move away from the straight line, the value of r approaches zero. Thus the value of r of diagram 16 is higher than that of diagram 15 and the value of ρ of diagram 19 is higher in absolute value than that of diagram 18.

# Spearman's Rank Correlation Coefficient

- This correlation measurement is also called **Rank correlation**.
- This technique is applicable to determine the degree of correlation between two variables in case of ordinal data.
- We can assign rank to the different values of a variable with ordinal data type.

Example:

Height: [VS S L T VT]
1 2 3 4 5

T – shirt: [XS S L XL XXL]
11 12 13 14 15

Rank assigned

# Spearman's Rank Correlation Coefficient

Definition : **Charles Spearman's correlation coefficient**

The rank correlation can be defined as

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

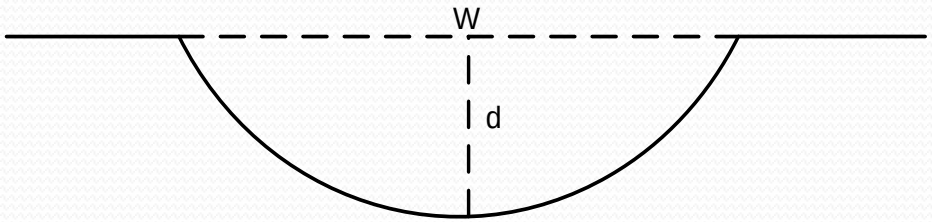
where $d_i$ = Difference between ranks of $i^{\text{th}}$ pair of the two variables

$n$ = Number of pairs of observations

- The Spearman's coefficient is often used as a statistical methods to aid either proving or disproving a hypothesis.

# Spearman's Rank Correlation Coefficient

**Example :** The hypothesis that the depth of a river **does not progressively increase** with the width of the river.




A sample of size 10 is collected to test the hypothesis, using Spearman's correlation coefficient.

| *Sample#* | *Width in m* | *Depth in m* |
|---|---|---|
| 1 | 0 | 0 |
| 2 | 50 | 10 |
| 3 | 150 | 28 |
| 4 | 200 | 42 |
| 5 | 250 | 59 |
| 6 | 300 | 51 |
| 7 | 350 | 73 |
| 8 | 400 | 85 |
| 9 | 450 | 104 |
| 10 | 500 | 96 |

# Spearman's Rank Correlation Coefficient

**Step 1:** Assign rank to each data. It is customary to assign rank 1 to the largest data, and 2 to next largest and so on.

Note: If there are two or more samples with the same value, the mean rank should be used.

| *Data* | 20 | 25 | 25 | 25 | 30 |
|---|---|---|---|---|---|
| *Assign rank* | 5 | 4 | 3 | 2 | 1 |
| *Final rank* | 5 | 3 | 3 | 3 | 1 |

# Spearman's Rank Correlation Coefficient

**Step 2:** The contingency table will look like

| *Sample#* | *Width* | *Width rank* | *Depth* | *Depth rank* | $d$ | $d^2$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 10 | 0 | 10 | 0 | 0 |
| 2 | 50 | 9 | 10 | 9 | 0 | 0 |
| 3 | 150 | 8 | 28 | 8 | 0 | 0 |
| 4 | 200 | 7 | 42 | 7 | 0 | 0 |
| 5 | 250 | 6 | 59 | 5 | 1 | 1 |
| 6 | 300 | 5 | 51 | 6 | −1 | 1 |
| 7 | 350 | 4 | 73 | 4 | 0 | 0 |
| 8 | 400 | 3 | 85 | 3 | 0 | 0 |
| 9 | 450 | 2 | 104 | 1 | 1 | 1 |
| 10 | 500 | 1 | 96 | 2 | −1 | 1 |

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 4}{10 \times 99}$$

$$r_s = 0.9757$$

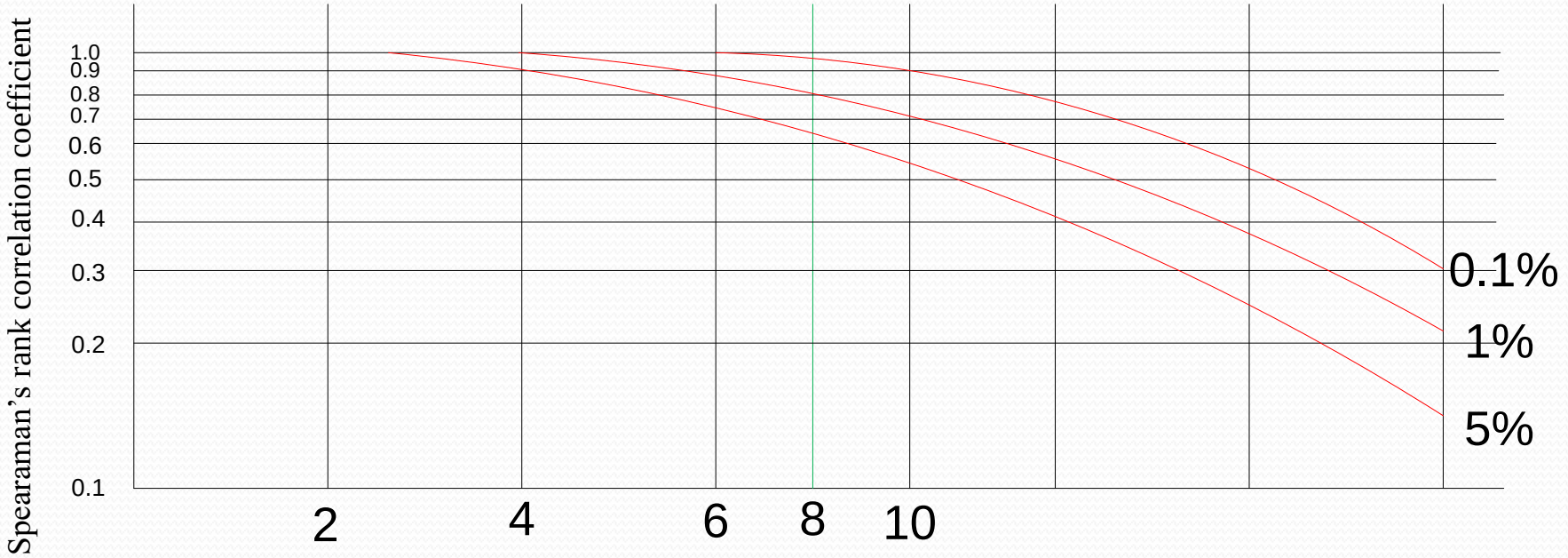
$$\sum d^2 = 4$$

# Spearman's Rank Correlation Coefficient

**Step 3:** To see, if this $r_s$ value is significant, the Spearman's rank significance table (or graph) must be consulted.

Note: The degrees of freedom for the sample $= n - 2 = 8$

Assume, the significance level $= 0.1\%$



# Spearman's Rank Correlation Coefficient

**Step 4:** Final conclusion

From the graph, we see that $r_s = 0.9757$ lies above the line at 8 and 0.01% significance level. Hence, there is a greater than 99% chance that the relationship is significant (i.e., not random) and hence the hypothesis should be rejected.

Thus, we can reject the hypothesis and conclude that in this case, depth of a river **progressively increases** the further with the width of the river.

# Chi-Squared Test of Correlation

- This method is also alternatively termed as Pearson's $\chi^2$–test or simply $\chi^2$ -test
- This method is applicable to categorical (discrete) data only.

  - Suppose, two attributes *A* and *B* with categorical values

    $A = a_1, a_2, a_3, \ldots\ldots, a_m$ and

    $B = b_1, b_2, b_3, \ldots\ldots, b_n$

    having *m* and *n* distinct values.

| $A$ | $a_1$ | $a_2$ | $a_3$ | $a_1$ | $a_5$ | $a_1$ | … … |
|---|---|---|---|---|---|---|---|
| $B$ | $b_1$ | $b_2$ | $b_3$ | $b_1$ | $b_5$ | $b_1$ | … … |

Between whom we are to find the correlation relationship.

# $\chi^2$ –Test Methodology

**Contingency Table**

Given a data set, it is customary to draw a contingency table, whose structure is given below.

| | $b_1$ | $b_2$ | ----- | $b_j$ | ------ | $b_n$ | **Row Total** |
|---|---|---|---|---|---|---|---|
| $a_1$ | | | | | | | |
| $a_2$ | | | | | | | |
| ⋮ | | | | | | | |
| $a_i$ | | | | | | | |
| ⋮ | | | | | | | |
| $a_m$ | | | | | | | |
| **Column Total** | | | | | | | **Grand Total** |

# $\chi^2$ –Test Methodology

**Entry into Contingency Table: Observed Frequency**

In contingency table, an entry $O_{ij}$ denotes the event that attribute $A$ takes on value $a_i$ and attribute $B$ takes on value $b_j$ (i.e., $A = a_i$, $B = b_j$).

| $A$ | $a_i$ | $a_2$ | $a_3$ | $a_i$ | $a_5$ | $a_i$ | … … |
|---|---|---|---|---|---|---|---|
| $B$ | $b_j$ | $b_2$ | $b_3$ | $b_j$ | $b_5$ | $b_j$ | … … |

| | $b_1$ | $b_2$ | ----- | $b_j$ | ------ | $b_n$ | Row Total |
|---|---|---|---|---|---|---|---|
| $a_1$ | | | | | | | |
| $a_2$ | | | | | | | |
| ⋮ | | | | | | | |
| $a_i$ | | | | $O_{ij}$ | | | |
| ⋮ | | | | | | | |
| $a_m$ | | | | | | | |
| Column Total | | | | | | | Grand Total |

# $\chi^2$ –Test Methodology

**Entry into Contingency Table: Expected Frequency**

In contingency table, an entry $e_{ij}$ denotes the expected frequency, which can be calculated as

$$e_{ij} = \frac{Count(A = a_i) \times Count(B = b_j)}{Grand\ Total} = \frac{A_i \times B_j}{N}$$

| | $b_1$ | $b_2$ | ----- | $b_j$ | ------ | $b_n$ | Row Total |
|---|---|---|---|---|---|---|---|
| $a_1$ | | | | | | | |
| $a_2$ | | | | | | | |
| ⋮ | | | | | | | |
| $a_i$ | | | | $e_{ij}$ | | | $A_i$ |
| ⋮ | | | | | | | |
| $a_m$ | | | | | | | |
| Column Total | | | | $B_j$ | | | N |

# $\chi^2$ – Test

**Definition : $\chi^2$-Value**

The $\boldsymbol{\chi}^2$ value ( also known as the Pearson's $\boldsymbol{\chi}^2$ test) can be computes as

$$\boldsymbol{\chi}^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

where $o_{ij}$ is the **o**bserved frequency

$e_{ij}$ is the **e**xpected frequency

# $\chi^2$ – Test

- The cell that contribute the most to the $\chi^2$ value are those whose actual count is very different from the expected.
- The $\chi^2$ statistics tests the hypothesis that *A* and *B* are independent. The test is based on a significance level, with $(n-1) \times (m-1)$ degrees of freedom., with a contingency table of size $n \times m$
- If the hypothesis can be rejected, then we say that *A* and *B* are statistically related or associated.

# $\chi^2$ – Test

**Example : Survey on Gender versus Hobby.**

- Suppose, a survey was conducted among a population of size 1500. In this survey, gender of each person and their hobby as either “book” or “computer” was noted. The survey result obtained in a table like the following.

| GENDER | HOBBY |
|---|---|
| ..... | ..... |
| ..... | ..... |
| M | Book |
| F | Computer |
| ..... | ..... |
| ..... | ..... |
| ..... | ..... |

- We have to find if there is any association between Gender and Hobby of a people, that is, we are to test whether “gender” and “hobby” are correlated.

# $\chi^2$ – Test

**Example : Survey on Gender versus Hobby.**

- From the survey table, the observed frequency are counted and entered into the contingency table, which is shown below.

| | | GENDER | | |
|---|---|---|---|---|
| | | Male | Female | Total |
| HOBBY | Book | 250 | 200 | 450 |
| | Computer | 50 | 1000 | 1050 |
| | Total | 300 | 1200 | **1500** |

# $\chi^2$ – Test

**Example : Survey on Gender versus Hobby.**

- From the survey table, the expected frequency are counted and entered into the contingency table, which is shown below.

| | | GENDER | | |
|---|---|---|---|---|
| | | Male | Female | Total |
| HOBBY | Book | 90 | 360 | 450 |
| | Computer | 210 | 840 | 1050 |
| | Total | 300 | 1200 | **1500** |

# $\chi^2$ – Test

- Using equation for $\chi^2$ computation, we get

$$\chi^2 = \frac{(250-90)^2}{90} + \frac{(50-210)^2}{210} + \frac{(200-360)^2}{360} + \frac{(1000-840)^2}{840}$$
$$= 507.93$$

- This value needs to be compared with the tabulated value of $\chi^2$ (available in any standard book on statistics) with 1 degree of freedom (for a table of $m \times n$, the degrees of freedom is $(m-1)\times(n-1)$; here $m = 2$, $n = 2$).

- For 1 degree of freedom, the $\chi^2$ value needed to reject the hypothesis at the 0.01 significance level is 10.828. Since our computed value is above this, we reject the hypothesis that "Gender" and "Hobby" are independent and hence, conclude that the two attributes are *strongly correlated* for the given group of people.

# Partial Correlation Coefficients

- We begin with the following example.
- Suppose data on IQ ($x_1$), result in the final examination ($x_0$) and number of times visited cinema hall ($x_2$) were taken from a group of students and the simple correlations were calculated as

$$r_{01} = 0.8,\ r_{02} = 0.3 \text{ and } r_{12} = 0.6.$$

$r_{02} = 0.3$ has been found to be significantly different from zero. But it is an unexpected result. We do not expect $x_0$ and $x_2$ to have a positive correlation. It means that as the students increase there visit to cinema hall their results are likely to be better. There must be something wrong.

- After scrutiny the investigator discovered that the intelligent students mostly visited the cinema hall. To find the true correlation between $x_0$ and $x_2$ we should thus eliminate the effect of IQ. This can be done by separately regressing $x_0$ and $x_2$ on $x_1$ and finding the simple correlation of the two residuals.
- The correlation coefficient between two variables after eliminating the effect of a third variable is known as partial correlation coefficient. It is also possible to eliminate the effect of as many variables as we want.

# Partial Correlation Coefficients

The formula for partial correlation coefficient of $x_0$ and $x_2$ after eliminating the effect of $x_1$ is

$$r_{02.1} = (r_{02} - r_{01}\, r_{21})/(\sqrt{(1-r_{01}^2)}\sqrt{(1-r_{21}^2)})$$

In the above example the value of $r_{02.1}$ is

$$(0.3-0.8\times0.6)/(\sqrt{(1-.8^2)}\sqrt{(1-.6^2)}) = -\ 0.375.$$

This has a negative sign as expected.

# Reference


- Provost, Foster, and Tom Fawcett. (2013). Data Science for Business: What you need to know about data mining and data-analytic thinking. O'Reilly.

# Any question?

You may also send your question(s) at ctanujit@gmail.com