INTRODUCTION TO DATA ANALYTICS

Class # 15

Relation Analysis

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

QUOTE OF THE DAY..

Nothing great was ever achieved without enthusiasm.

- RALPH WALDO EMERSON, American philosopher

THIS TOPIC INCLUDES…

- Introduction
- Measures of Relationship
- Correlation Analysis
  - χ^2 - Test
  - Spearman's Correlation Analysis
  - Pearson's Correlation Analysis
- Regression Analysis
- Auto-Regression Analysis

RELATIONSHIP ANALYSIS

- **Example: Wage Data**

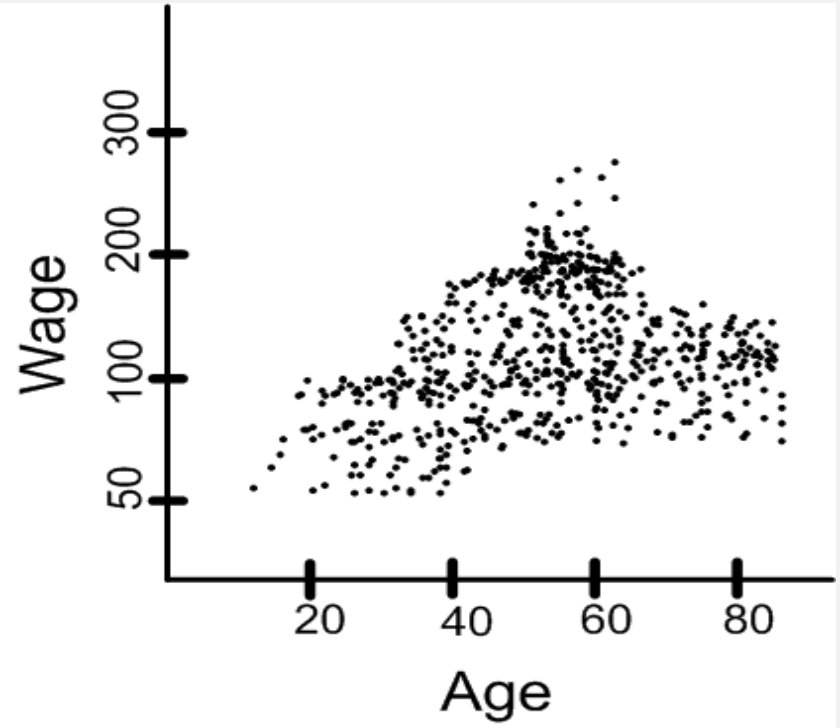
  A large data regarding the wages for a group of employees from the eastern region of India is given.

  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:

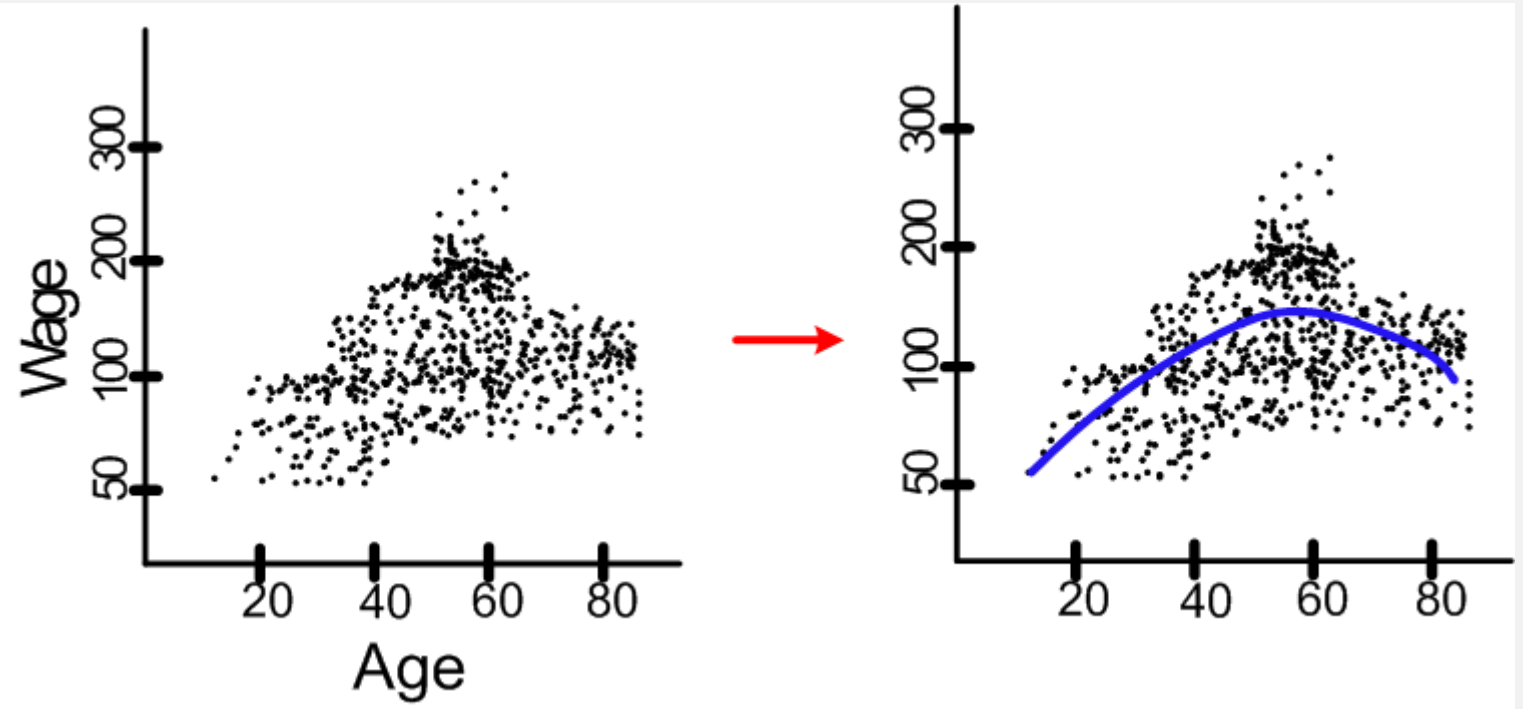


?

How wages vary with ages?

RELATIONSHIP ANALYSIS

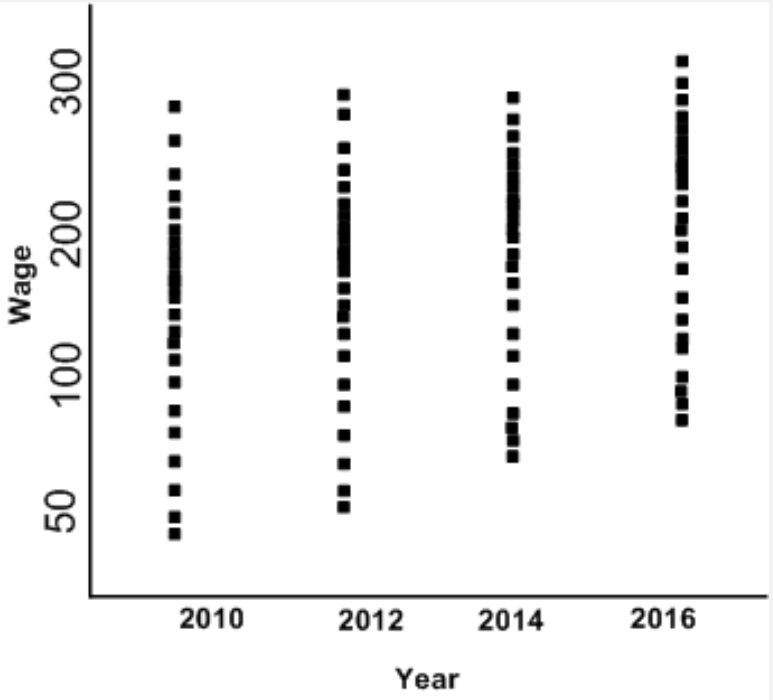
- Example: Wage Data
  - *Employee's age and wage:* How wages vary with ages?



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:

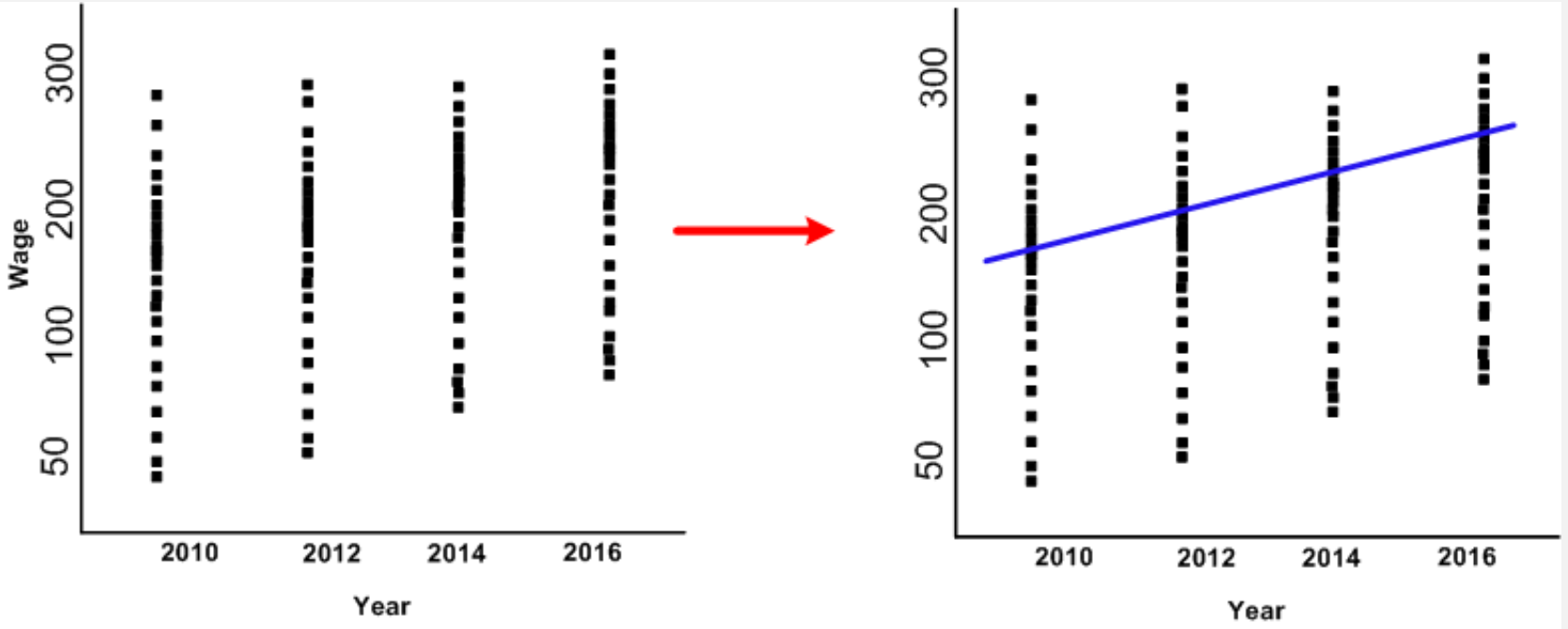


?

How wages vary with time?

RELATIONSHIP ANALYSIS

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

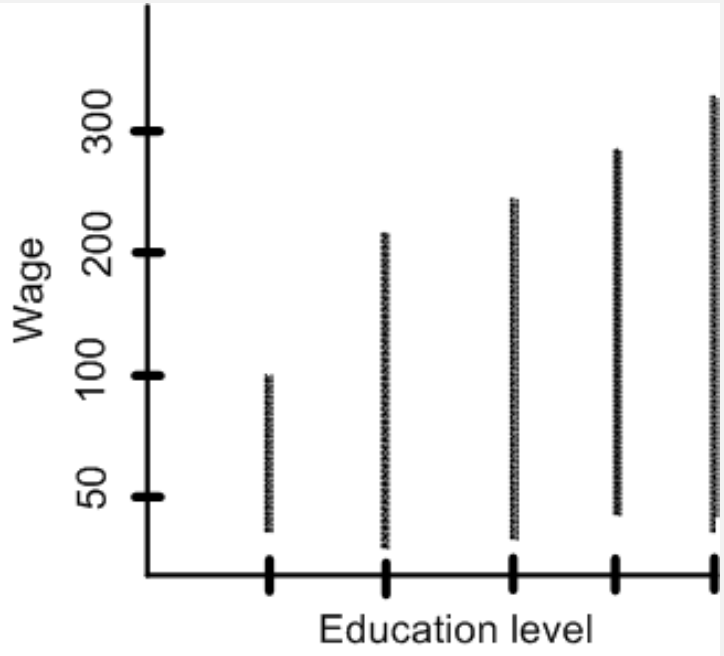


Interpretation: There is a slow but steady increase in the average wage between 2010 and 2016.

.

RELATIONSHIP ANALYSIS

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

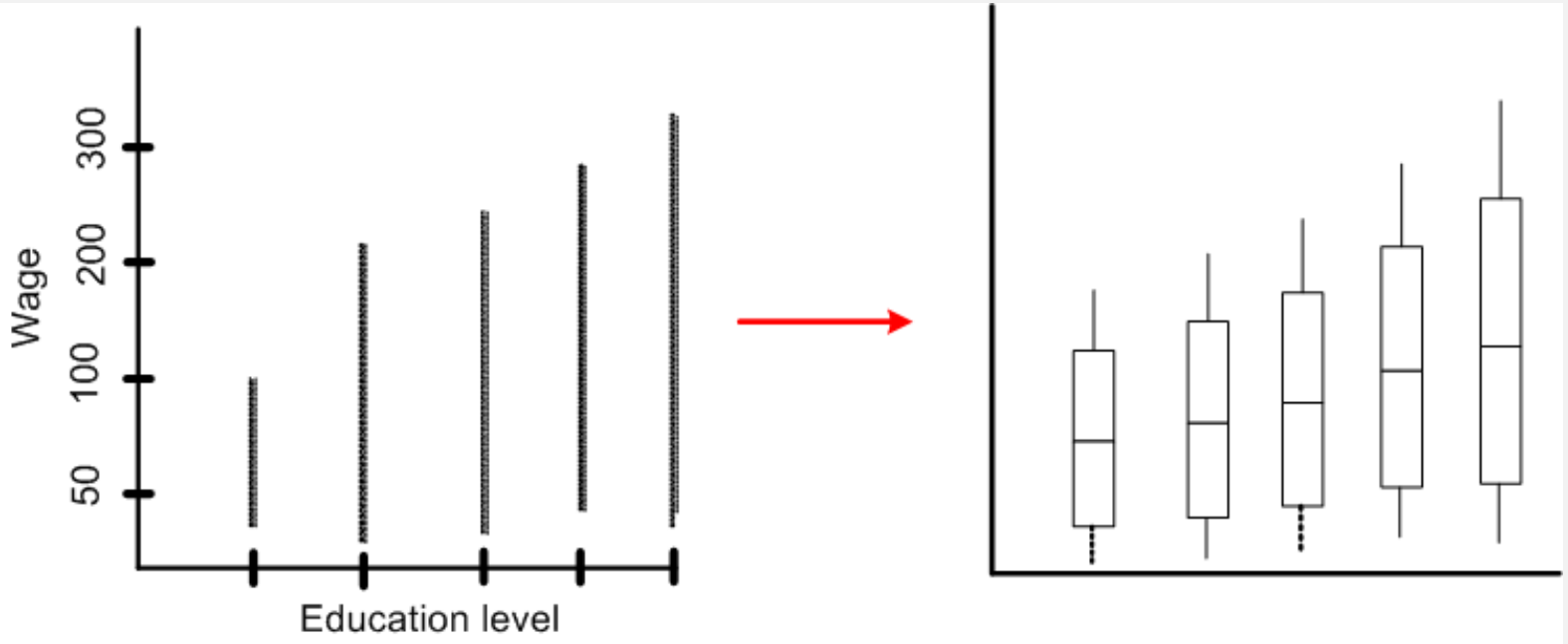


?

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.

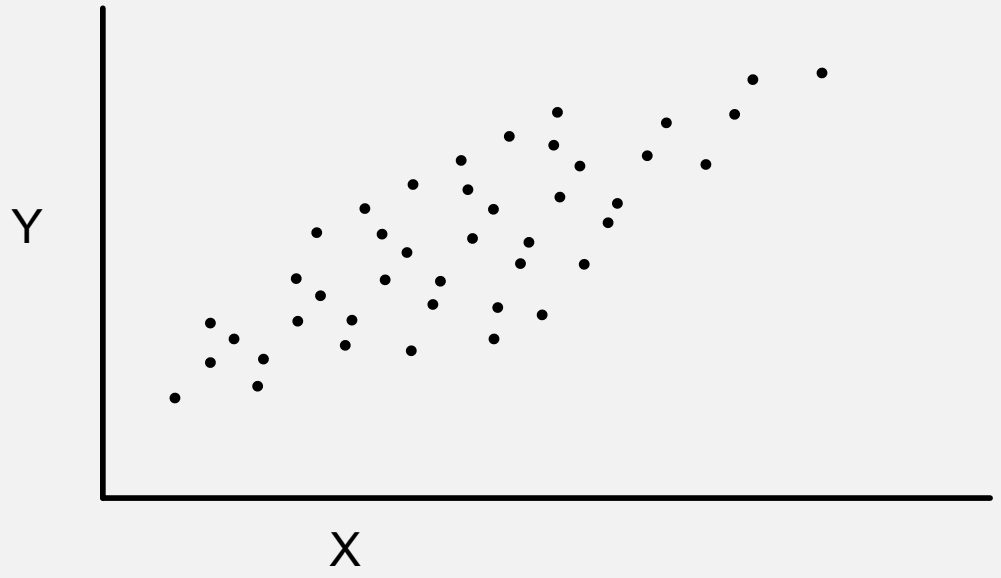
RELATIONSHIP ANALYSIS

Given an employee's wage can we predict his education level?

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

etc….

AN OPEN CHALLENGE!

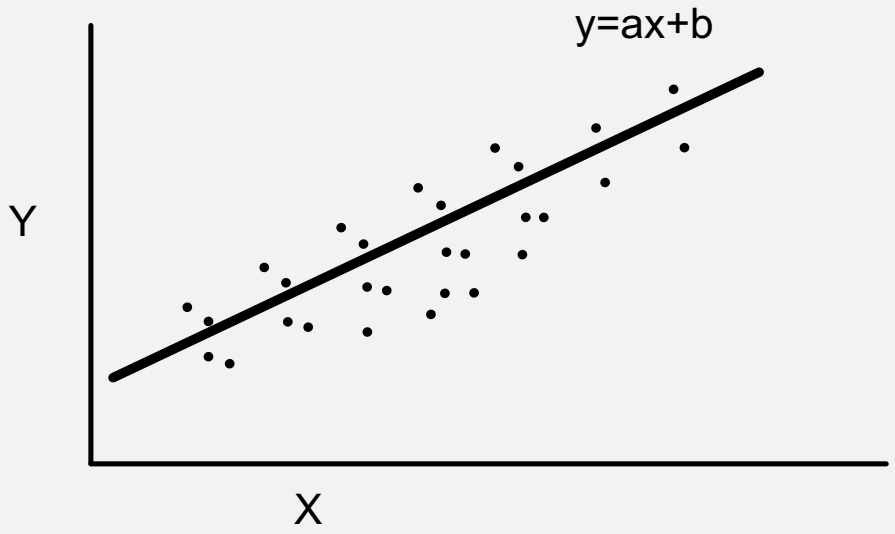


Suppose there are countably infinite points in the *XY plane*. We need a huge memory to store all such points.

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

Say, with two values only.

YAHOO!



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

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

Note: Here, tricks 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

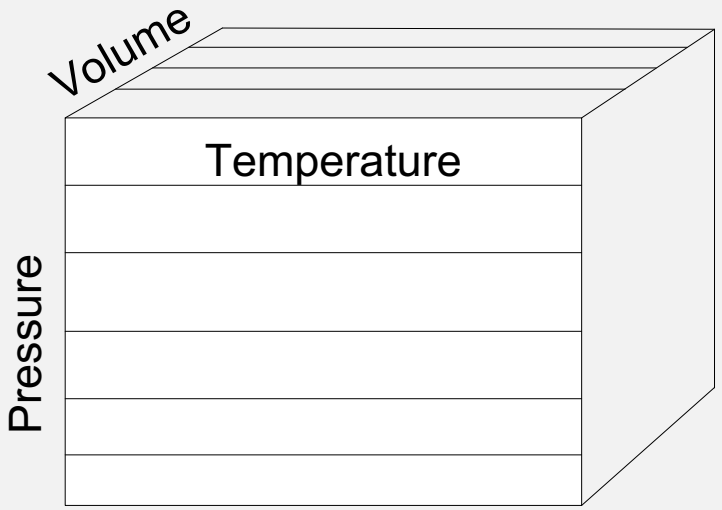
Here, statistical measures are suffice to find a relationship.

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

Pressure	1	1.1	… …	0.8
Temperature	35	41		29

MEASURES OF RELATIONSHIP

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



If we add another variable say viscosity in addition to Pressure, Volume or Temperature?

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**

Example:

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	……

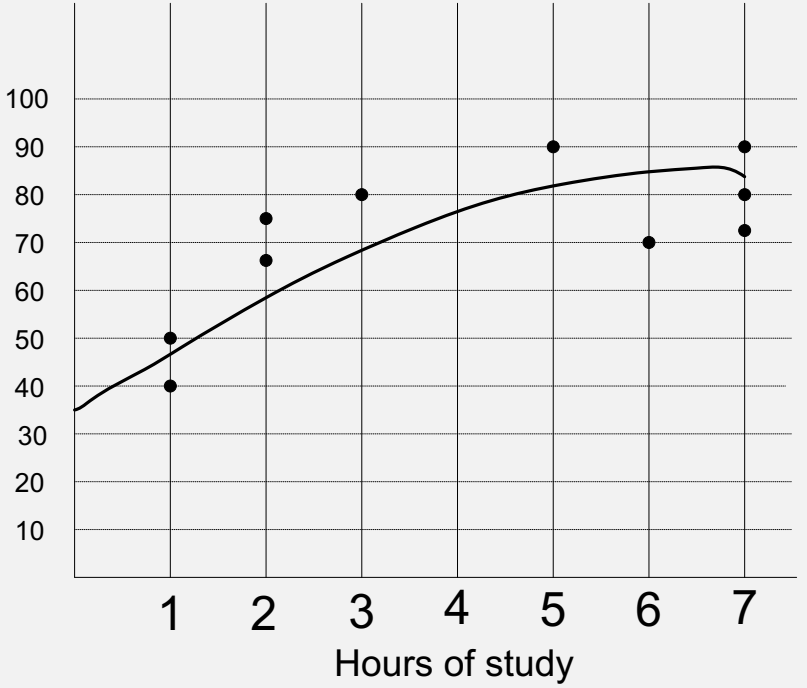
The correlation may be positive, negative or zero.

- **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.

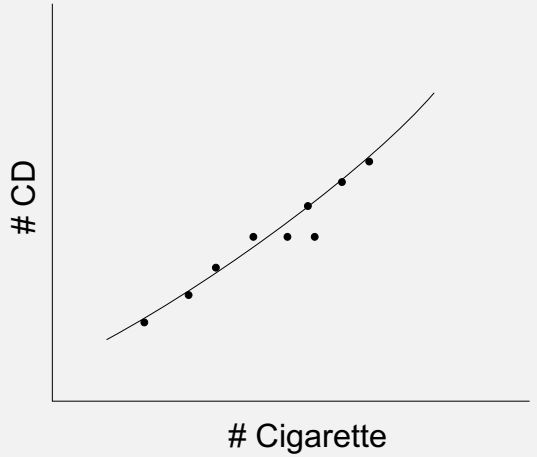
Hours Study	*Exam Score*
3	80
5	90
2	75
6	80
7	90
1	50
2	65
7	85
1	40
7	100



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



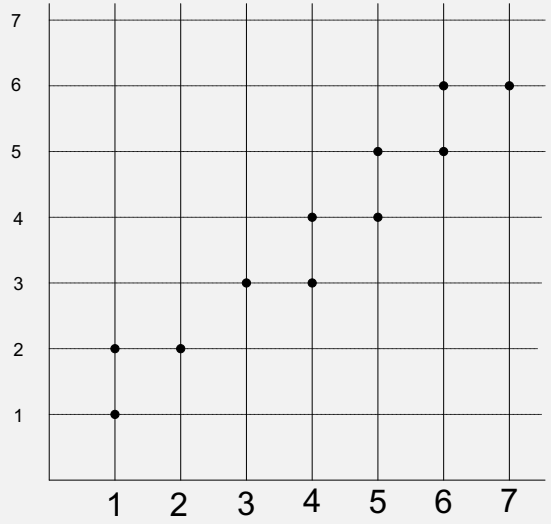
Note:

In data analytics, correlation analysis make sense only when relationship make sense.

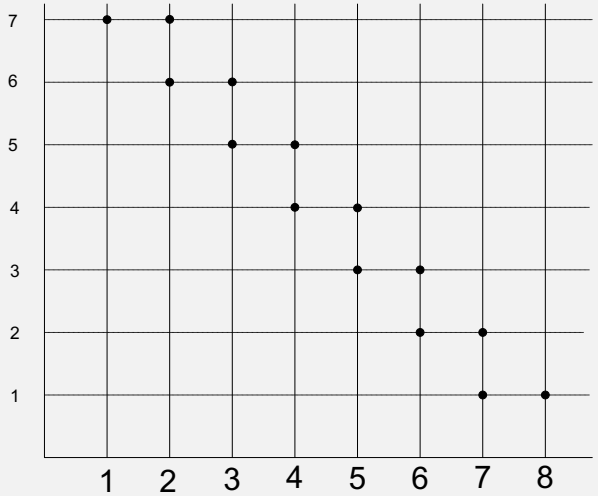
There should be a cause-effect relationship.

CORRELATION ANALYSIS

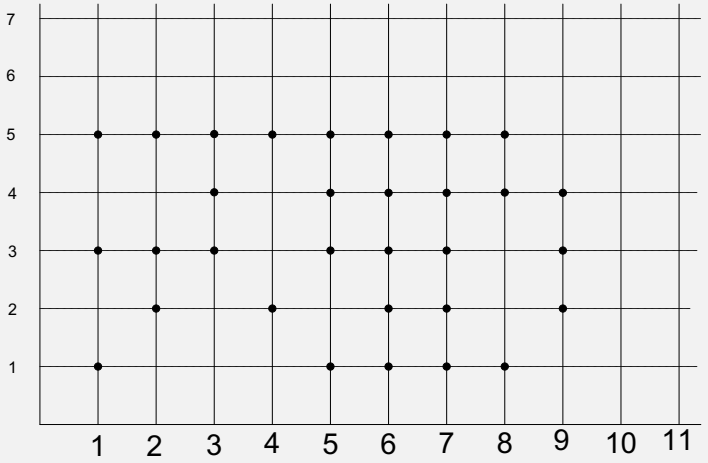
Positive correlation



Negative correlation



Zero correlation

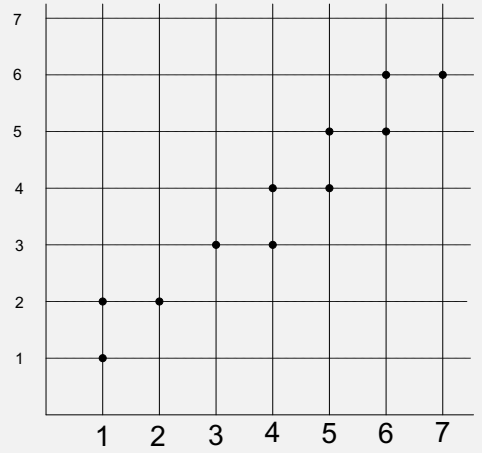


CORRELATION COEFFICIENT

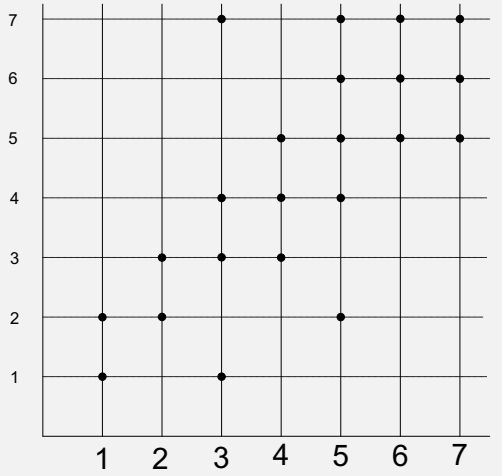
- **Correlation coefficient is used to measure the degree of association.**
- It is usually denoted by *r*.
- The value of *r* lies between +1 and -1.
- Positive values of *r* indicates positive correlation between two variables, whereas, negative values of *r* indicate negative correlation.
- $r = +1$ 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

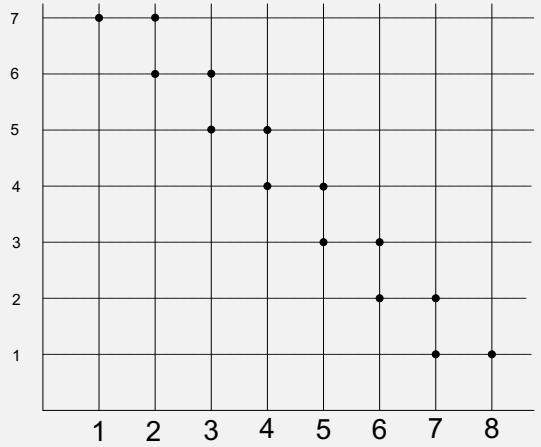
High Positive Correlation



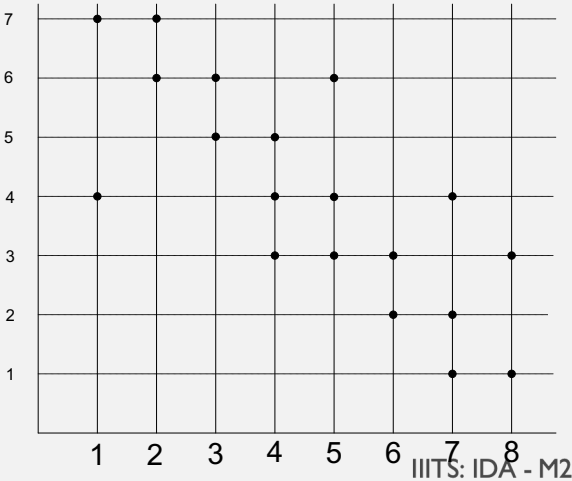
Low Positive Correlation



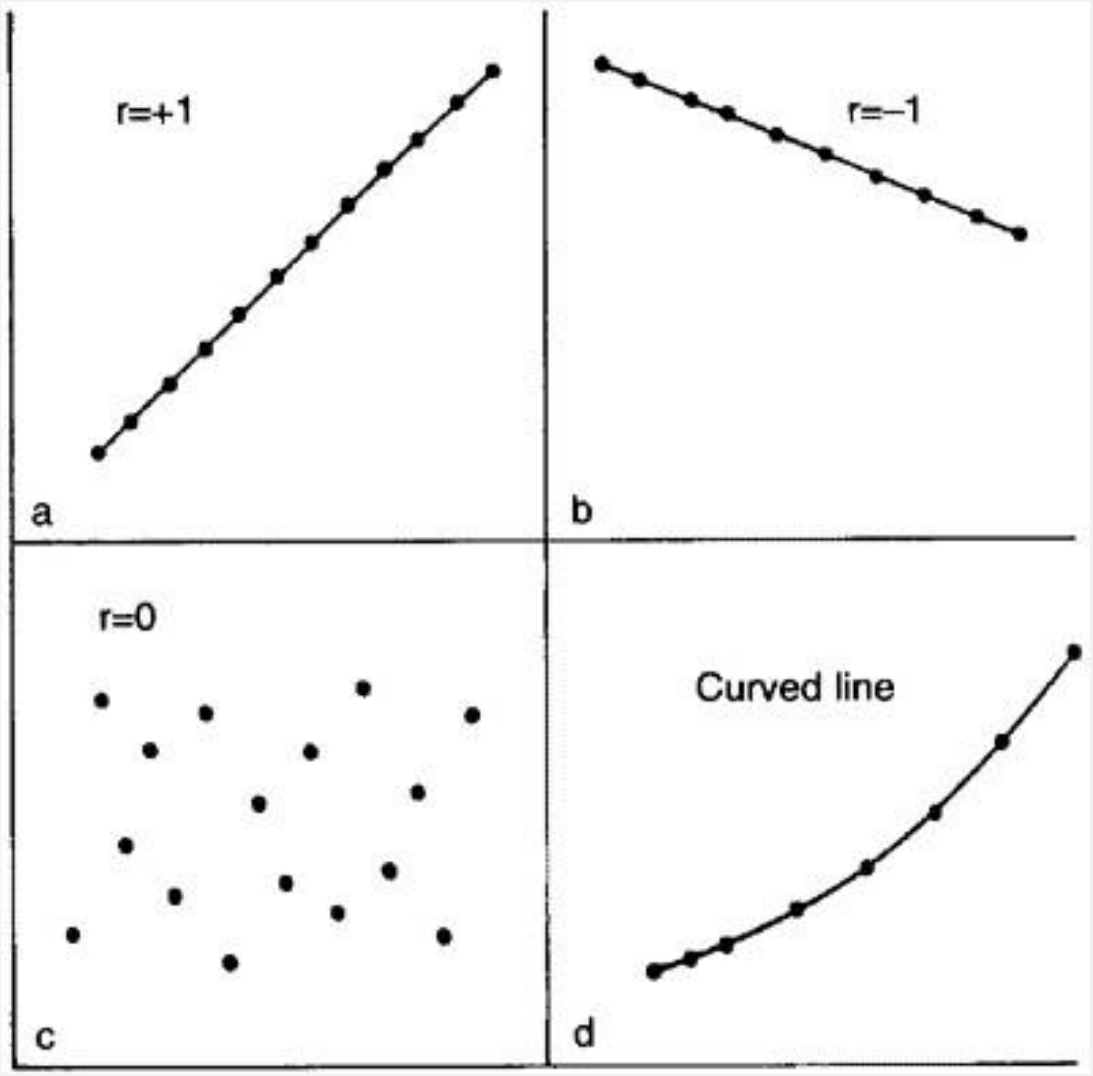
High Negative Correlation



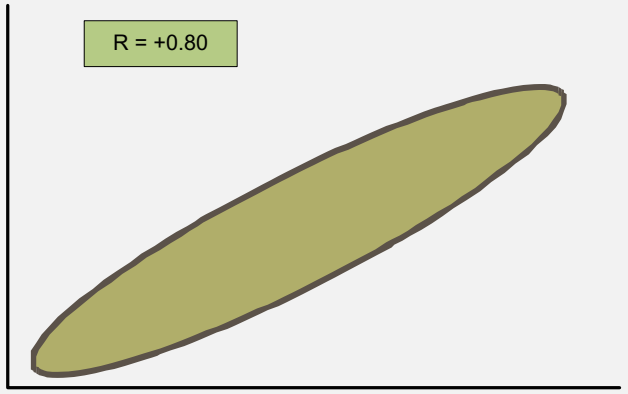
Low Negative Correlation

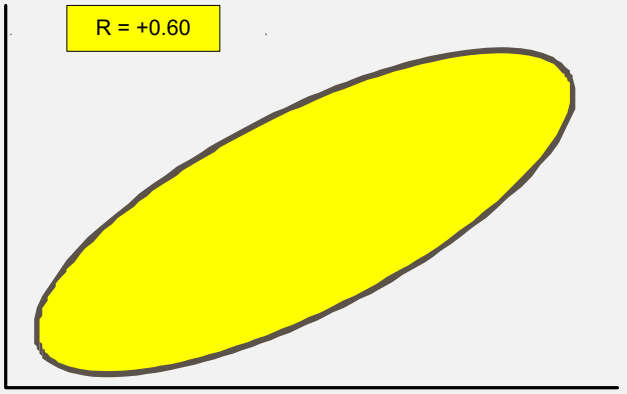


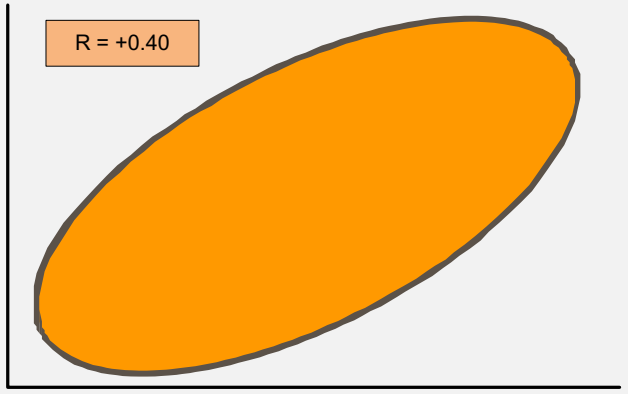
CORRELATION COEFFICIENT

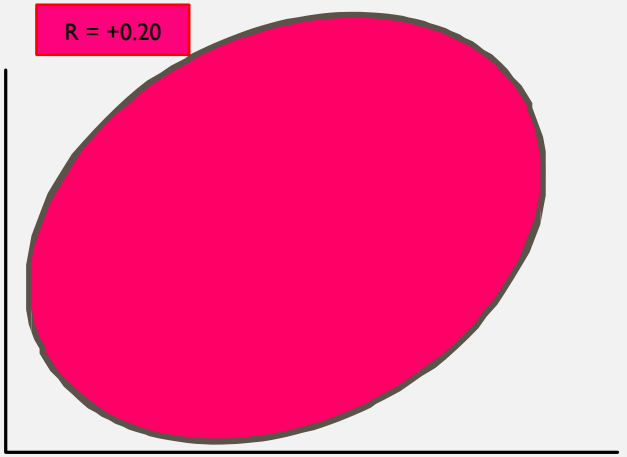


CORRELATION COEFFICIENT









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

PEARSON'S CORRELATION COEFFICIENT

KARL PEARSON'S CORRELATION COEFFICIENT

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

Definition 7.1: **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 7.1: Correlation of Gestational Age and Birth Weight

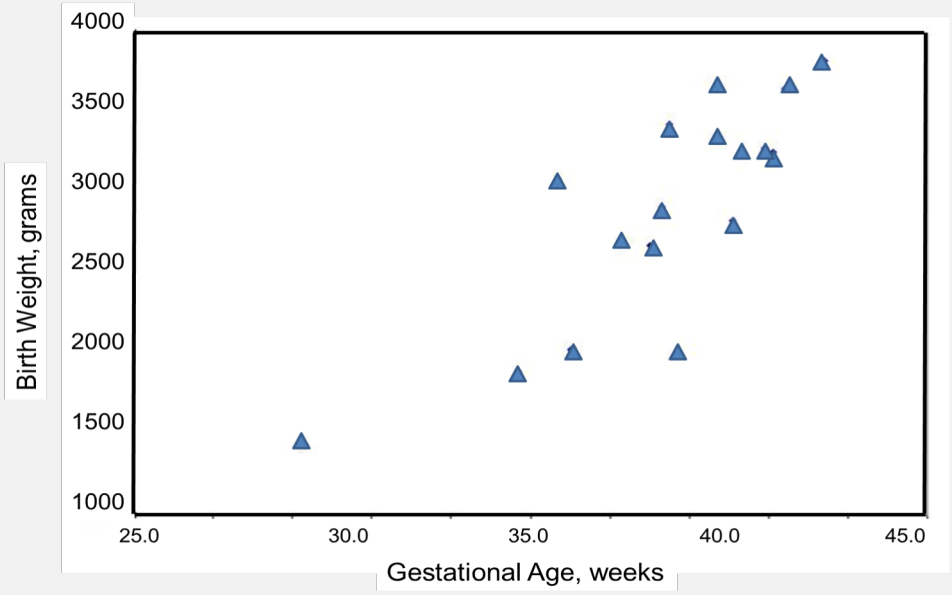
- A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

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 7.1: 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 7.1: 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 7.1: 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.

RANK CORRELATION COEFFICIENT

CHARLES SPEARMAN'S 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

CHARLES SPEARMAN'S CORRELATION COEFFICIENT

Definition 7.2: Charles Spearman's correlation coefficient

The rank correlation can be defined as

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

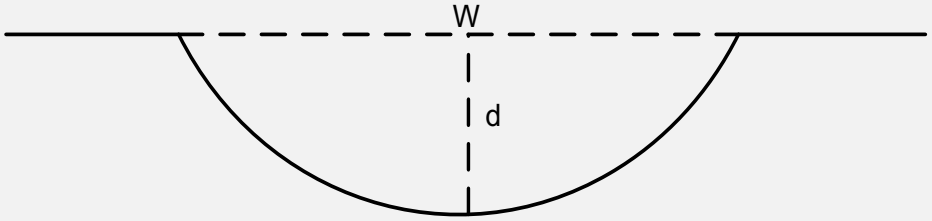
where d_i = Difference between ranks of i^{th} pair of the two variables

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.

CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

Example 7.2: 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

CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

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

CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

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

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

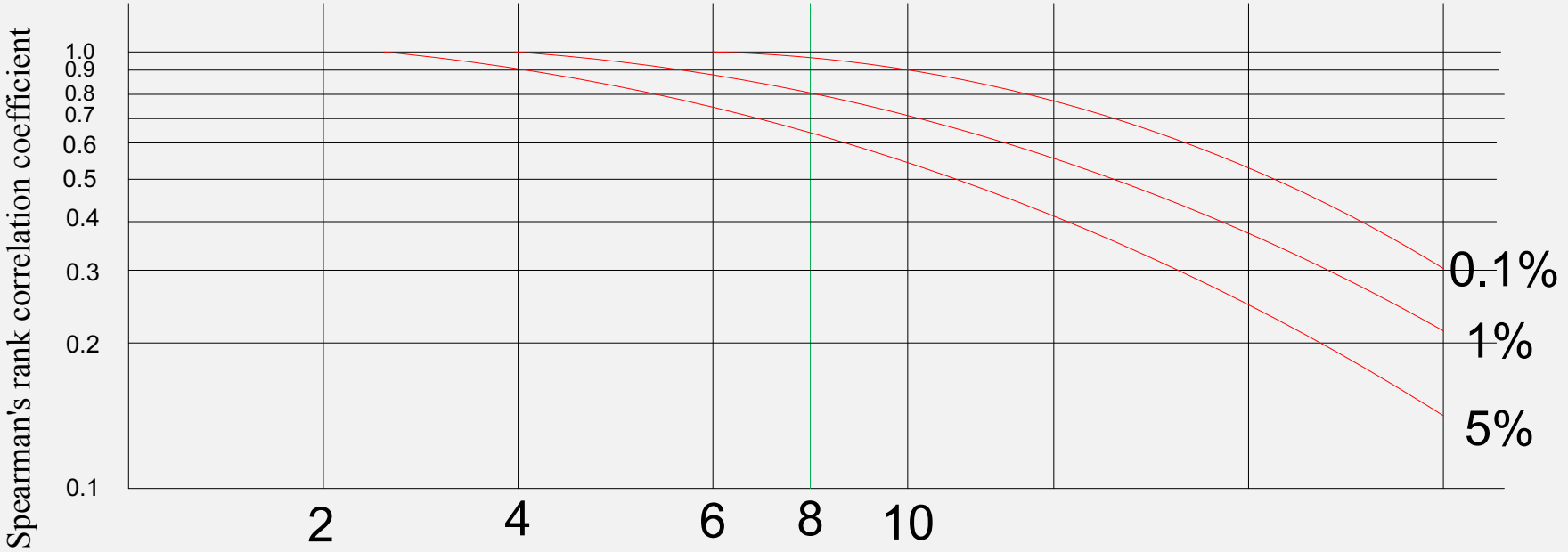
$$r_s = 0.9757$$

$$\sum d^2 = 4$$

CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

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 $= 0.01$



CHARLES SPEARMAN'S COEFFICIENT OF CORRELATION

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.

X2-CORRELATION ANALYSIS

CHI-SQUARED TEST OF CORRELATION

- This method is also alternatively termed as Pearson's χ^2–test or simply χ^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.

X^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**

X^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**

X^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**

X^2 – TEST

Definition 7.3: χ^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

X^2 – TEST

- The cell that contribute the most to the χ^2 value are those whose actual count is very different from the expected.

- The χ^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.

X^2 – TEST

Example 7.3: 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.

X^2 – TEST

Example 7.3: 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**

X^2 – TEST

Example 7.3: 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		
HOBBY		Male	Female	Total
	Book	90	360	450
	Computer	210	840	1050
	Total	300	1200	**1500**

X^2 – TEST

- Using equation for χ^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 χ^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 χ^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.

X^2 – TEST

Example 7.4: Hypothesis on "accident proneness" versus "driver's handedness".

- Consider the following contingency table on car accidents among left and right-handed drivers' of sample size 175.
- Hypothesis is that *"fatality of accidents is independent of driver's handedness"*

FATALITY	HANDEDNESS		
	Left-Handed	Right-Handed	Total
Non-Fatal	8	141	149
Fatal	3	23	26
Total	11	164	175

- Find the correlation between Fatality and Handedness and test the significance of the correlation with significance level 0.1%.

REFERENCE

- The detail material related to this lecture can be found in

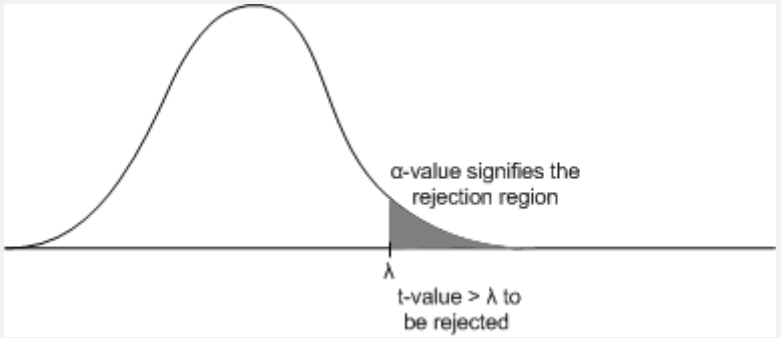
The Elements of Statistical Learning, Data Mining, Inference, and Prediction (2nd Edn.), Trevor Hastie, Robert Tibshirani, Jerome Friedman, Springer, 2014.

Any question?

QUESTIONS OF THE DAY…

1. For a given sample data the correlation coefficient according to the Karl Pearson's correlation analysis is found to be r = 0.79 with degree of freedom 69. Further, with significant test , the t-value is calculated as t = 2.36. From the t-test table, it is found that with degree of freedom 69, the t-value at 5% confidence level is 3.61. What is the inference that you can have in this case?

2. For a given degree of freedom, if α, the value of confidence level increases, then t-value increases. Is the statement correct? If not, what is the correct statement? Justify your answer. You can refer the following figure in your explanation.



QUESTIONS OF THE DAY…

3. Whether the Spearman's correlation coefficient analysis is applicable to the numeric data? If so, how?

4. Can χ^2–analysis be applied to ordinal data or numeric data? Justify your answer.

5. Briefly explain the following with reference to the χ^2 correlation analysis.
   a) Contingency table
   b) Observed frequency
   c) Expected frequency
   d) Expression for –vale calculation
   e) Hypothesis to be tested
   f) Degree of freedom of sample data