# STATISTICS FOR MACHINE LEARNING

An Internship Project Report

Submitted in Partial fulfillment of the requirements for the award of the

Degree of Bachelor of Science in Mathematics

Submitted by

**I.SAFRIN FATHIMA**

Register Number: 18BM7468

Under the guidance of

**Mrs.M.Kalarani M.Sc., M.Phil.,**

Assistant professor

**Department of Mathematics (Self Finance)**


**Sri G.V.G Visalakshi College for Women (Autonomous)**

**(Affiliated to Bharathiyar University, Coimbatore)**

**Accredited at 'A+' Grade by NAAC (CGPA 3.27) By NAAC**

**An ISO 9001:2015 Certified Institution**

**Udumalpet - 642 128**

**March 2021**

**E-CERTIFICATE**

**CERTIFICATE**

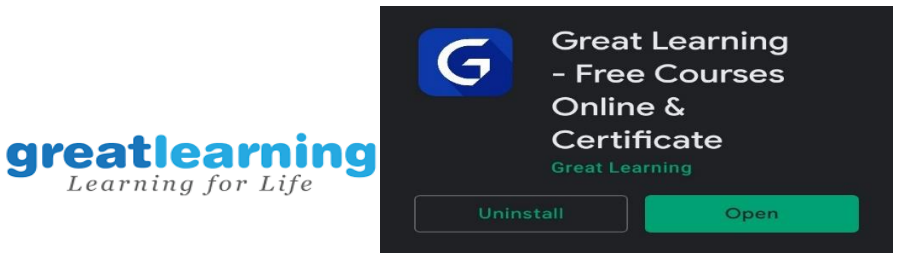
This is to certify that the project reported entitled **"STATISTICS FOR MACHINE LEARNING"** is a bonafied record work done by **I.SAFRIN FATHIMA (18BM7468)** Submitted in partial fulfillment of the requirement for the award Degree of Bachelor of science in Mathematics at Sri G.V.G Visalakshi College for Women (Autonomous), Udumalpet. Affiliated to Bharathiar University during the academic year 2020-2021.

**Head of the Department** **Signature of the Staff In-charge**

## Introduction

The Internship training program was organized by **"GREAT LEARNING"** application launched by Mohan lakhmraju in 2013. **Great Learning** is one of India's leading ed-tech companies for professional and higher education. Dr.Abhinanda Sarkar teaches about **"STATISTICS FOR MACHINE LEARNING"** which is foundation concepts like Descriptive Statitics,Data and Histogram,Standard Deviation and also explains the Empirical Rule,Chebyshev Rule and Correlation Analysis.

**Digital Tool**




**Dr. Abhinanda Sarkar Professor of Data Science gave us training under the topic "STATISTICS FOR MACHINE LEARNING".**

1. Introducion for Statistics

2. Raw Data

3. Descriptive Statistics

4. Frequency Distribution-Histograms

5. Central Tendency

6. Mean,Median,Mode

7. Measures of Dispersion

8. Range, IQR, Standard Deviation,coefficient of variation

9. The Empirical Rule,Chebyshev Rule

10. Five number summary,boxplots,QQ plots,Quantile plot,scatter plot.

11. Visualization:scatter plot matrix.

12. Correlation analysis

## 1. Introduction For Statistics

**"**By Statistics, we mean methods specially adopted to the elucidation of quantitative data affected to a marked extent by multiplicity of causes".

It is interesting to see what Thomas Davenport means by Business Analytics and note the similarities and dissimilarities between the two.

"Business Analytics (BA) can be defined as the broad use of data and quantitative analysis for decision making within organizations".

## 2. Raw Data

Raw Data represent numbers and facts in the orginal format in which the data have been collected. We need to convert the raw data into information for decision making.Raw data or primary data are collected directly related to their object of study. When people are the subject of an investigation,we may choose the form of a survey, an observation,or an experiment.

## Data Versus Information

When analysts are bewildered by plethora of data, which do not make any sense on the surface of it, they are looking for methods to classify data that would convey meaning. The idea here is to help them draw the right conclusion. Data needs to be arranged into information.

## 3. Descriptive statistics

Descriptive statistics are brief descriptive coefficients that summarize a given data set, which can be either a representation of the entire or a sample of a population. Descriptive statistics are broken down into measure of central tendency and measures of variability.

There are four major types of descriptive Statistics:

* Measures of Frequency: Count,percent,Frequency.

* Measures of Central Tendency: Mean,Median,and Mode.

* Measures of dispersion or variation: Range,Variance,Standard Deviation.

*Measures of Position: Percentile Ranks,Quartile Ranks.

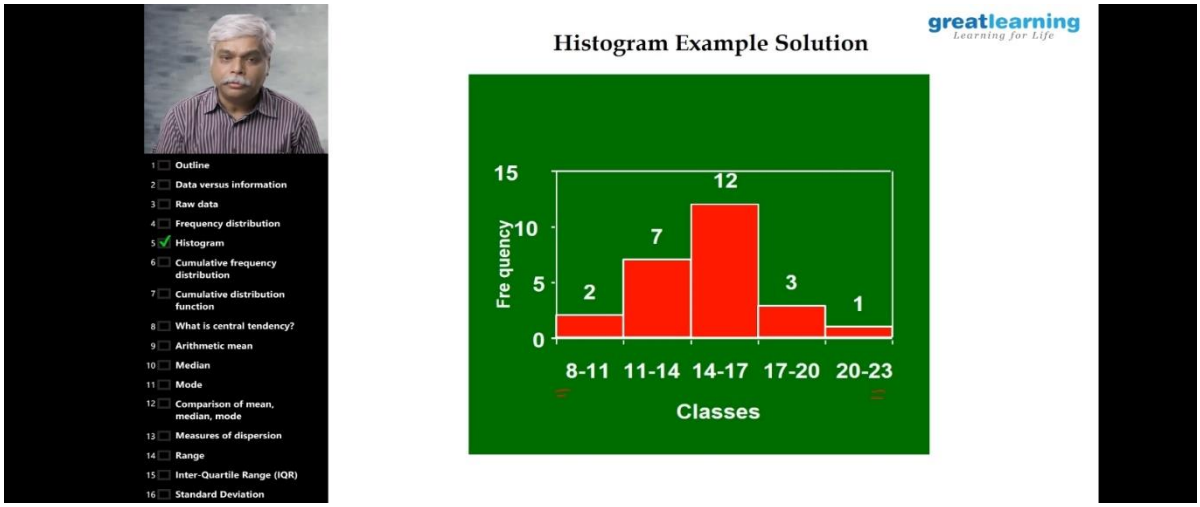
Descriptive Statistics are used to descriptive the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis they from the basis of virtually every quantitative analysis of data.

**4. Histogram**

Histogram (also known as frequency histogram) is a snap shot of the frequency distribution. Histogram is a graphical representation of the frequency distribution in which the X-axis represents the classes and the Y-axis represents the frequencies in bars Histogram depicts the pattern of the distribution emerging from the characteristic being measured.

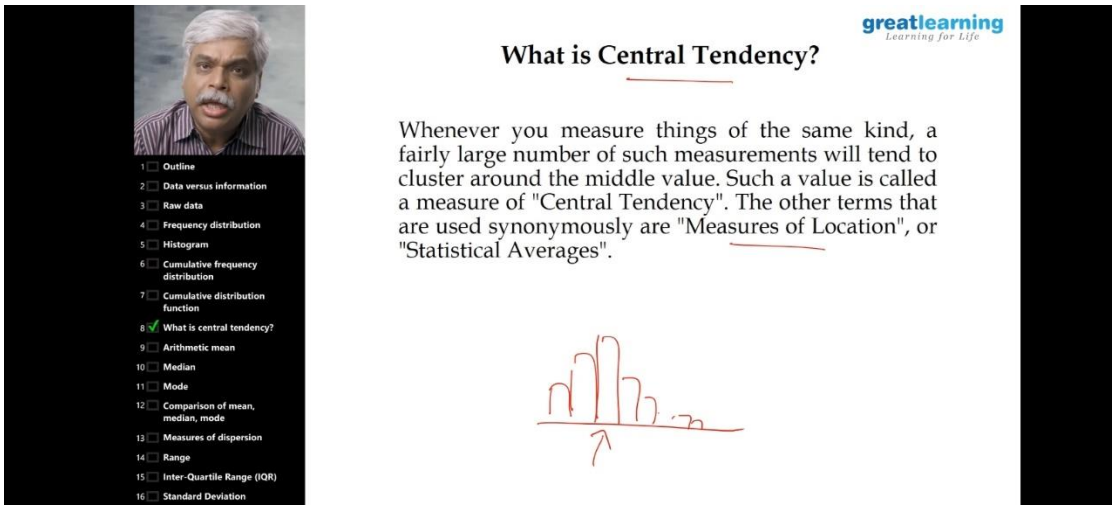
The inspection records of a hose assembly operation revealed a high level of rejection. An analysis of the records showed that the "leaks" were a major contributing factor to the problem. It was decided to investigate the hose clamping operation. The hose clamping force (torque) was measured on twenty five assemblies. (Figures in foot-pounds). The data are given below: Draw the frequency histogram and comment.

| | | | | |
|---|---|---|---|---|
| 8 | 13 | 15 | 10 | 16 |
| 11 | 14 | 11 | 14 | 20 |
| 15 | 16 | 12 | 15 | 13 |
| 12 | 13 | 16 | 17 | 17 |
| 14 | 14 | 14 | 18 | 15 |




## 5. Central Tendency:

Whenever you measure things of the same kind, a fairly large number of such measurements will tend to cluster around the middle value. Such a value is called a measure of "Central Tendency". The other terms that are used synonymously are "Measures of Location", or "Statistical Averages"




## 6. Arithmetic Mean:

Arithmetic Mean (called mean) is defined as the sum of all observations in a data set divided by the total number of observations. For example, consider a data set containing the following observations:

In symbolic form mean is given by

$$\bar{x} = \frac{\sum X}{n}$$

$\overline{X}$ = Arithmetic Mean

$\sum X$ =Indicates sum all X values in the data set

n= Total number of observations(Sample Size)

**Median:**

Median is the middle most observation when you arrange data in ascending order of magnitude. Median is such that 50% of the observations are above the median and 50% of the observations are below the median.Median is a very useful measure for ranked data in the context of consumer preferences and rating. It is not affected by extreme values (greater resistance to outliers)

$$\text{Median} = \frac{n+1}{2}$$

n= Number of observations in sample

**Median - Example**

Marks obtained by 7 students in Computer Science Exam are given below: Compute the median.

45 40 60 80 90 65 55

Arranging the data after ranking gives

90 80 65 60 55 45 40

Median = (n+1)/2 th value in this set = (7+1)/2 th

observation= 4th observation=60

Hence Median = 60 for this problem.

**Mode**

Mode is that value which occurs most often. It has the maximum frequency of occurrence. Mode also has resistance to outliers. Mode is a very useful measure when you want to keep in the inventory, the most popular shirt in terms of collar size during festive season.

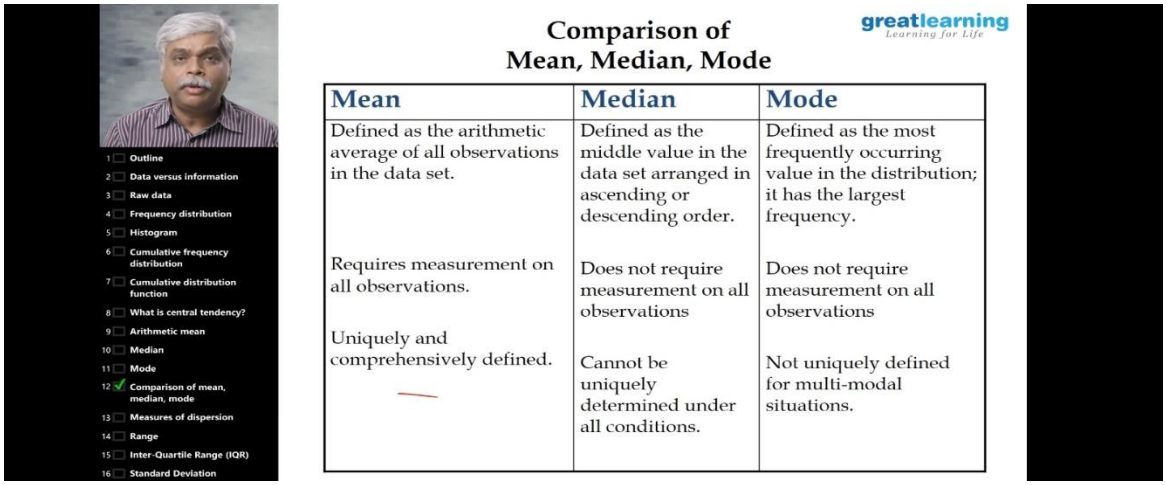
**Mode -Example**

The life in number of hours of 10 flashlight batteries are as follows: Find the mode.

340 350 340 340 320 340 330 330 340 350

340 occurs five times. Hence, mode=340.

**Comparison of Mean,Median,Mode**




**7. Measures of Dispersion**

In simple terms, measures of dispersion indicate how large the spread of the distribution is around the central tendency. It answers unambiguously the question " What is the magnitude of departure from the average value for different groups having identical averages?".

## 8. Range

Range is the simplest of all measures of dispersion. It is calculated as the difference between maximum and minimum value in the data set. Range = X Maximum − X Minimum

### Range-Example

Example for Computing Range

The following data represent the percentage return on investment for 10 mutual funds per annum. Calculate Range.

12, 14, 11, 18, 10.5, 11.3, 12, 14, 11, 9

Range = $X_{Maximum} - X_{Minimum}$ = 18-9=9

### Inter-Quartile Range(IQR)

IQR= Range computed on middle 50% of the observations after eliminating the highest and lowest 25% of observations in a data set that is arranged in ascending order. IQR is less affected by outliers.

IQR =Q3 -Q1

### Interquartile Range-Example

The following data represent the annual percentage returns of 9 mutual funds.

Data Set: 12, 14, 11, 18, 10.5, 12, 14, 11, 9

Arranging in ascending order, the data set becomes

9, 10.5, 11, 11, 12, 12, 14, 14, 18

IQR=Q3 -Q1=14-10.75=3.25

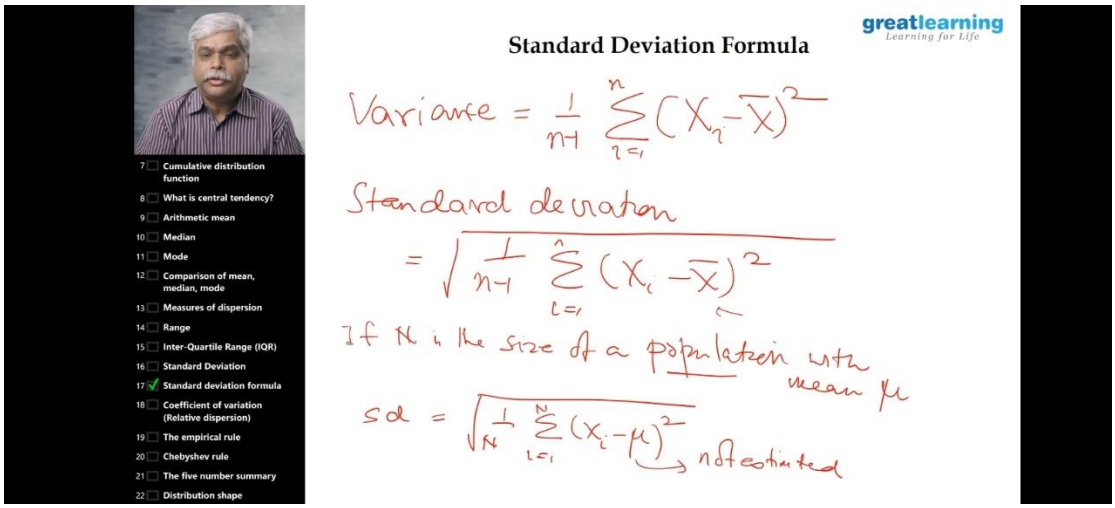
### Standard Deviation

To define standard deviation, you need to define another term called variance. In simple terms, standard deviation is the square root of variance.

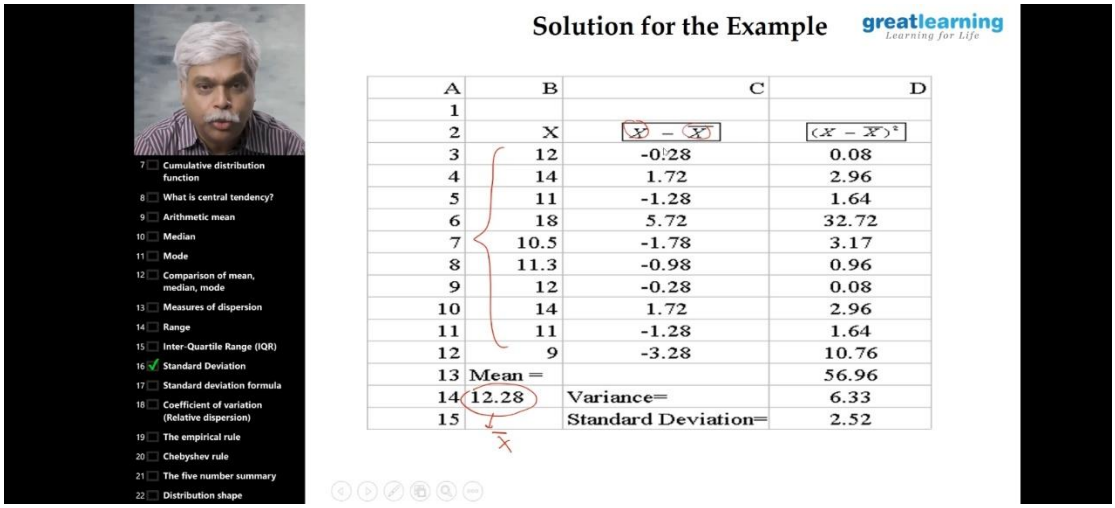
**Example of standard deviation**

The following data represent the percentage return on investment for 10 mutual fund sperannum .Calculate the sample standard deviation.

12,14,11,18,10.5,11.3,12,14,11,9




**Solution for the Example:**




| A | B | C | D |
|---|---|---|---|
| 1 | | | |
| 2 | X | $(x-\bar{x})$ | $(x-\bar{x})^2$ |
| 3 | 12 | -0.28 | 0.08 |
| 4 | 14 | 1.72 | 2.96 |
| 5 | 11 | -1.28 | 1.64 |
| 6 | 18 | 5.72 | 32.72 |
| 7 | 10.5 | -1.78 | 3.17 |
| 8 | 11.3 | -0.98 | 0.96 |
| 9 | 12 | -0.28 | 0.08 |
| 10 | 14 | 1.72 | 2.96 |
| 11 | 11 | -1.28 | 1.64 |
| 12 | 9 | -3.28 | 10.76 |
| 13 | Mean = | | 56.96 |
| 14 | 12.28 | Variance= | 6.33 |
| 15 | | Standard Deviation= | 2.52 |

($\bar{X}$)

**Coefficient of Variation**

**(Relative Dispersion)**

Coefficient Variation (CV) is defined as the ratio of Standard Deviation to Mean.

In symbolic form

CV = $\frac{S}{\bar{X}}$ for the sample data and = $\frac{\sigma}{\mu}$ for the population

**Coefficient of Variation Example**

Consider two Sales Persons working in the same territory

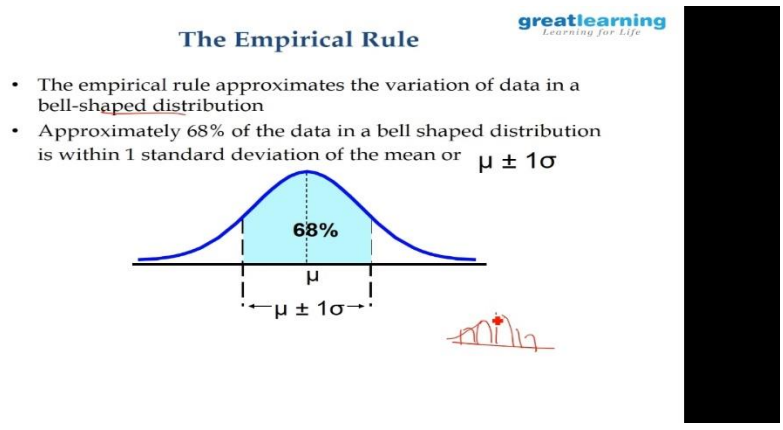
The sales performance of these two in the context of selling PC sare given below.

Comment on the results.

| **Sales Person 1** | **Sales Person 2** |
|---|---|
| Mean Sales (One year average) | Mean Sales(One year average) |
| 50 units | 75 units |
| Standard deviation | Standard deviation |
| 5 units | 25 units |

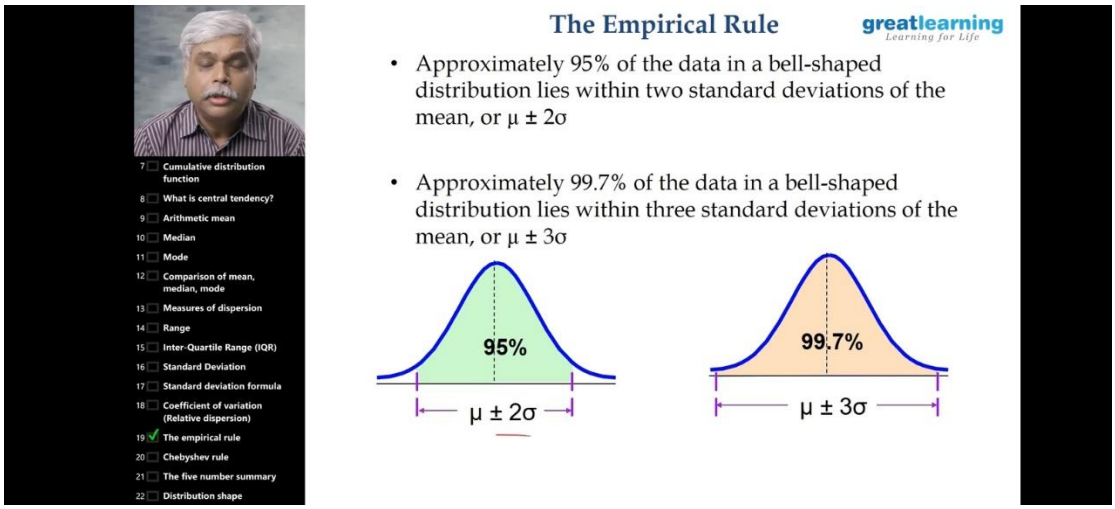
**9. The Empirical Rule**

- The empirical rule approximation the variation of data in a bell-shaped distribution.
- Approximately 68% of the data in a bell shaped distribution is within 1 standard deviation of the mean or $\boldsymbol{\mu} \pm 1\boldsymbol{\sigma}$


- Approximately 95% of the data in a bell-shaped distribution lies within two standard deviations of the mean, or $\mu \pm 2\sigma$

- Approximately 99.7% of the data in a bell-shaped distribution lies within three standard deviations of the mean, or $\mu \pm 3\sigma$




**Chebyshev Rule**

- Regardless of how the data are distributed, at least $(1 - 1/k^2) \times 100\%$ of the values will fall within k standard deviations of the mean (for $k > 1$)

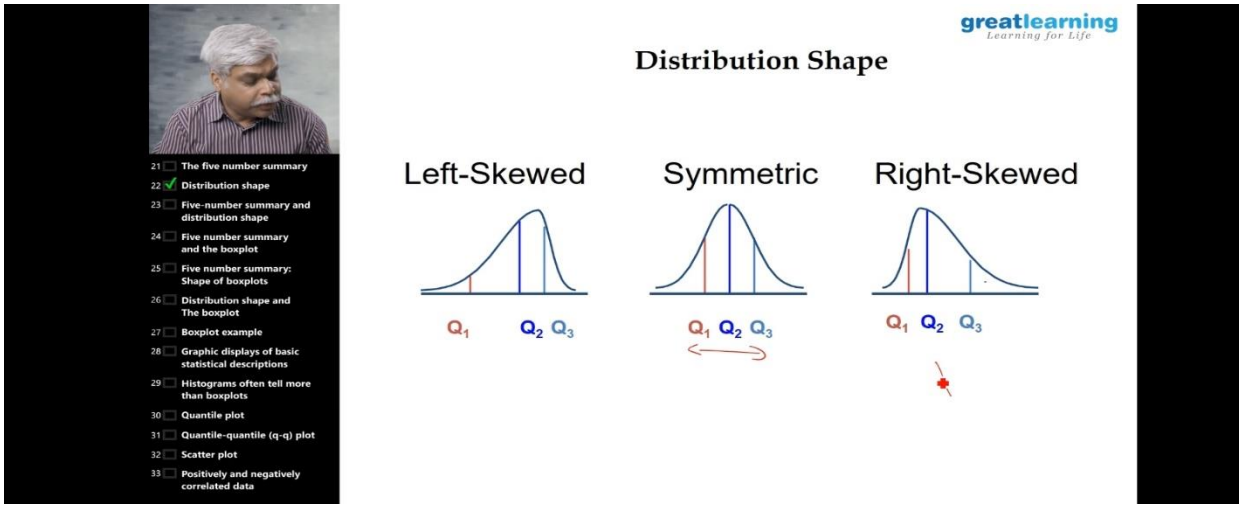
- For Example, when k=2, at least 75% of the values of any data set will be within $\mu \pm 2\sigma$

**The Five number Summary**

The five numbers that help describe the center, spread and shape of data are:

- $X_{smallest}$
- First Quartile ($Q_1$)
- Median ($Q_2$)
- Third Quartile ($Q_3$)
- $X_{largest}$

**Distribution Shape:**



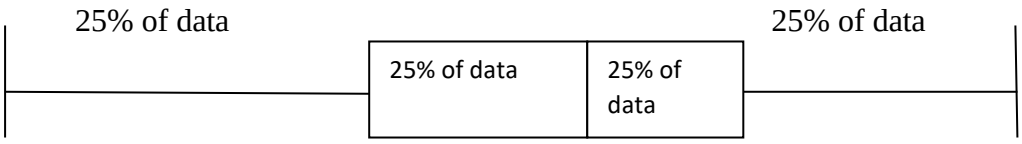

**Relationships amoung the five-number summary and distribution shape**

| **Left-Skewed** | **Symmetric** | **Right-Skewed** |
|---|---|---|
| Median – X $_{smallest}$ > X $_{largest}$ –Median | Median – X $_{smallest}$ ≈ X $_{largest}$ –Median | Median – X $_{smallest}$ < X $_{largest}$ –Median |
| **$Q_1$ –** X $_{smallest}$ > X $_{largest}$ _- $Q_3$ | **$Q_1$ –** X $_{smallest}$ ≈ X $_{largest}$ _- $Q_3$ | **$Q_1$ –** X $_{smallest}$ < X $_{largest}$ _- $Q_3$ |
| Median – $Q_1$ > $Q_3$ - Median | Median – $Q_1$ ≈ $Q_3$ - Median | Median – $Q_1$ < $Q_3$ - Median |

**10. Five Number Summary and The Boxplot**

The Boxplot: A Graphical display of the data based on the Five –number summary:

**Example:**




$X_{smallest}$ $Q_1$ Median $Q_3$ $X_{largest}$

**Five Number Summary:**

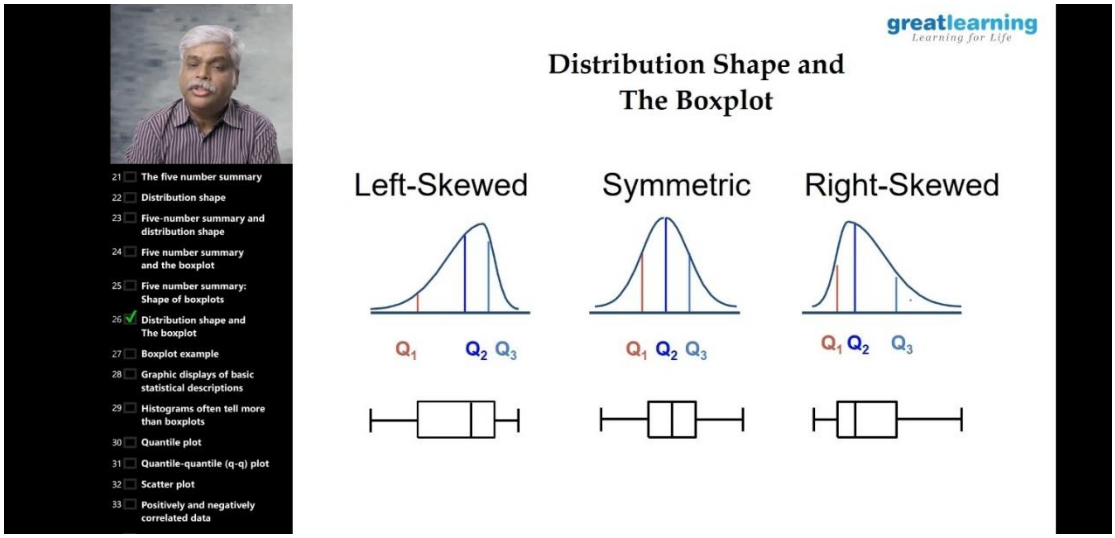
**Shape of Boxplot**

- If data are symmetric around the median then the box and central line are centered between the endpoints.

- A Boxplot can be shown in either a vertical or horizontal orientation.

**Distribution Shape and**

**The Boxplot**




**Graphic Displays of Basic Statistical Descriptions**

**Boxplot:** graphic display of five-number summary

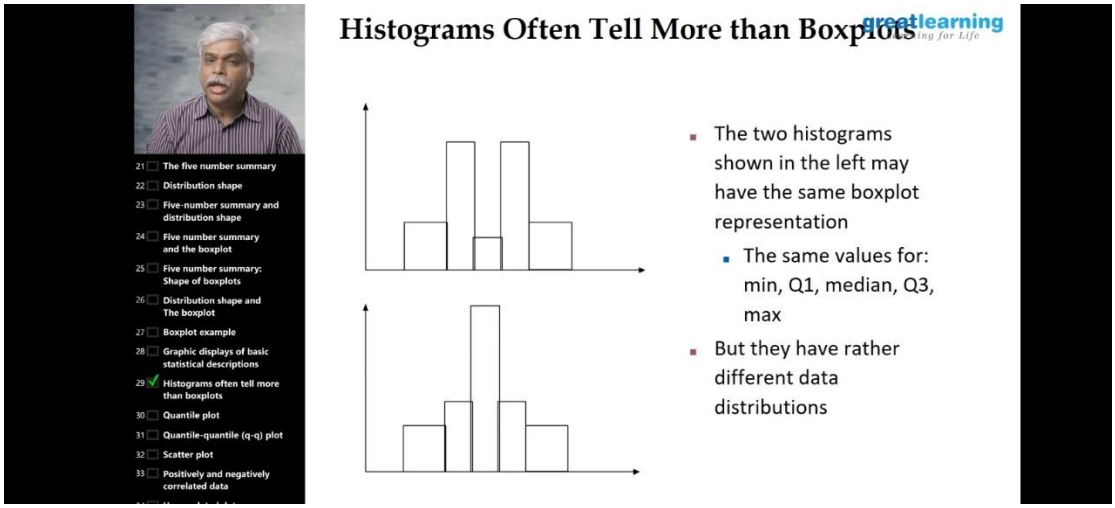
**Histogram:** x-axis are values, y-axis repres. Frequencies

**Quantile plot:** each value $x_i$is paired with $f_i$ indicating that approximately 100 $f_i$ % of data are $\leq x_i$

**Quantile-quantile (q-q) plot:** graphs the quantiles of one univariant distribution against the corresponding quantiles of another**.**

**Scatter plot:** each pair of values is a pair of coordinates and plotted as points in the plane.

## Histograms Often Tell More than Boxplots

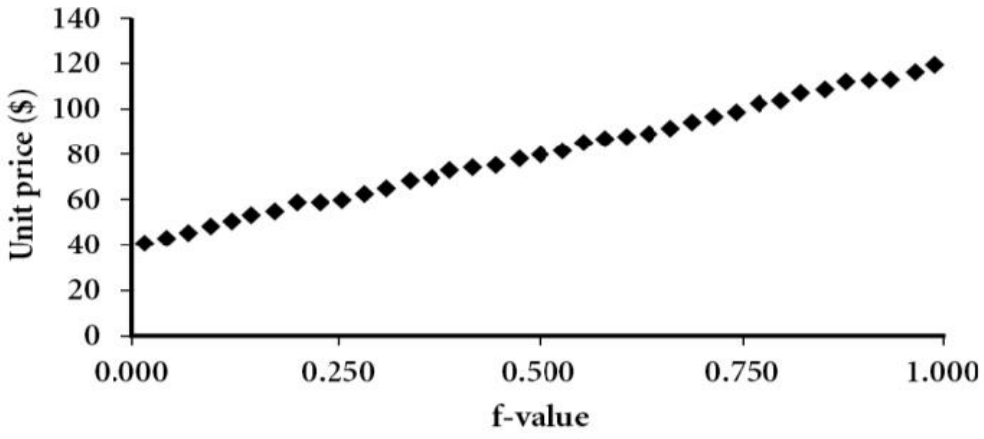



## Quantile Plot

Displays all of the data (allowing the user to access both the overall behavior and unusual occurances)

Plots quantile information

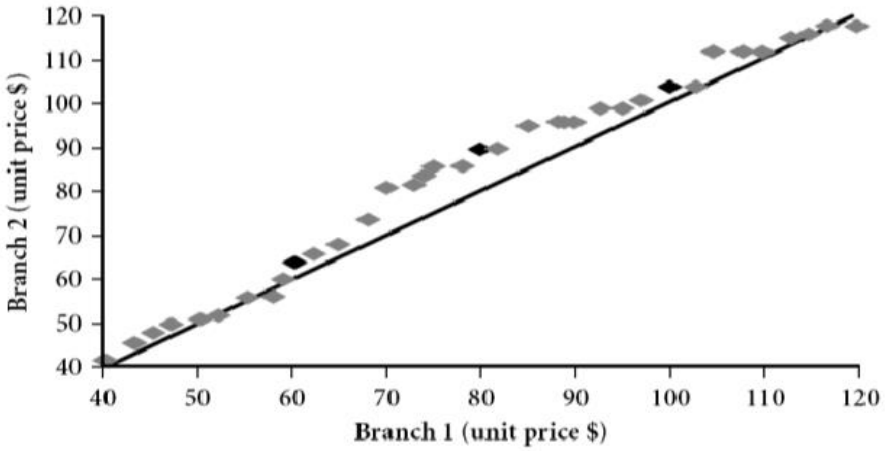
For a data x, data sorted in increasing order, $f_i$ indicates that approximately 100 $f_i$% of the data are below or equal to the value $x_i$




**Quantile-Quantile(Q-Q) Plot**

Graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

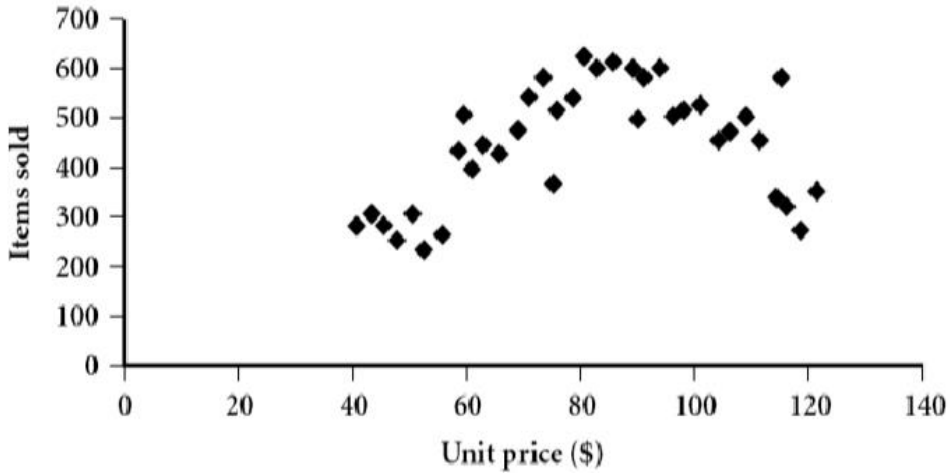
**View:** Is there is a shift in going from one distribution to another? Example shows unit price of items sold at Branch 1 vs. Branch 2 for each quantile. Unit prices of items sold at Branch 1 tend to be lower than those at Branch 2.



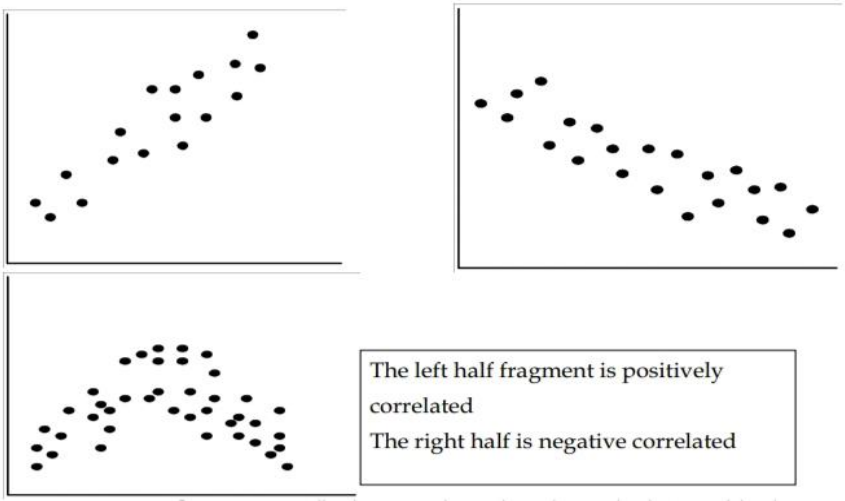

**11. Scatter plot**

Provides a first look at bivariate data to see clusters of points, outliers, etc

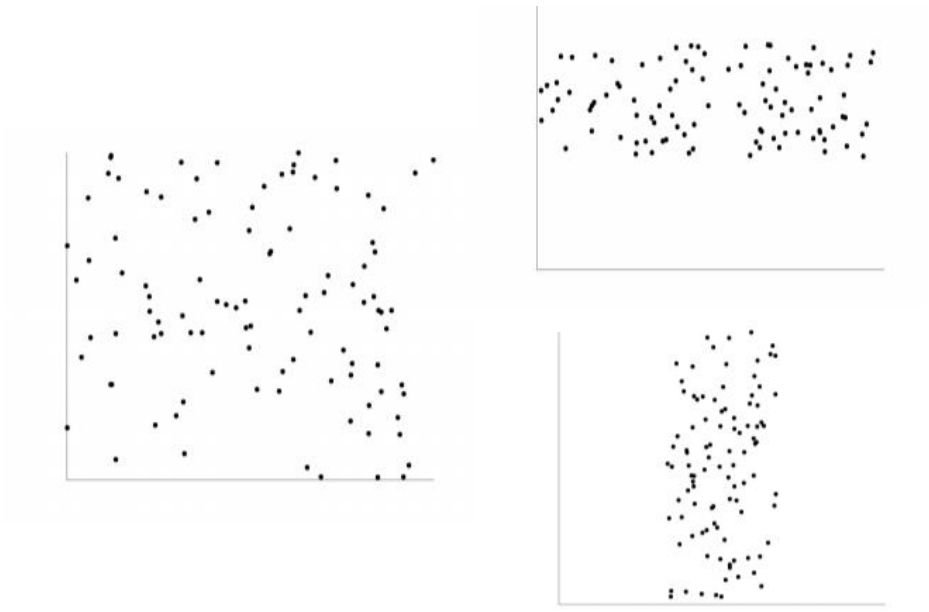
Each pair of values is treated as a pair of coordinates and plotted as points in the plane.



**Positively and Negatively correlated Data**




**Uncorrelated Data**



**12. Correlation Analysis (Nominal Data):Chi-Square Test**

| | Play chess | Not play chess | Sum(row) |
|---|---|---|---|
| Like science fiction | 250(90) | 200(360) | 450 |
| Not like science fiction | 50(210) | 1000(840) | 1050 |
| Sum(col.) | 300 | 1200 | 1500 |

$$e_{ij} = \frac{count(male) x count(fiction)}{n}$$

= 300x450/1500

=9

For this 2 x 2 table the degree of freedom are(2-1)(2-1) = 1. For 1 degree of freedom, the 2 values needed to reject the hypothesis at the the 0.001 significant level is 10. 828

$X^2$(chi-square) calculation(number in parenthesis are expected counts calculated based on the data distribution in the two categories)

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

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

=507.93

Shows that like_science_fiction and play_chess are correlated in the group.

**Correlation Analysis(Numeric Data)**

Correlation coefficient (also called pearson's product moment coefficient)

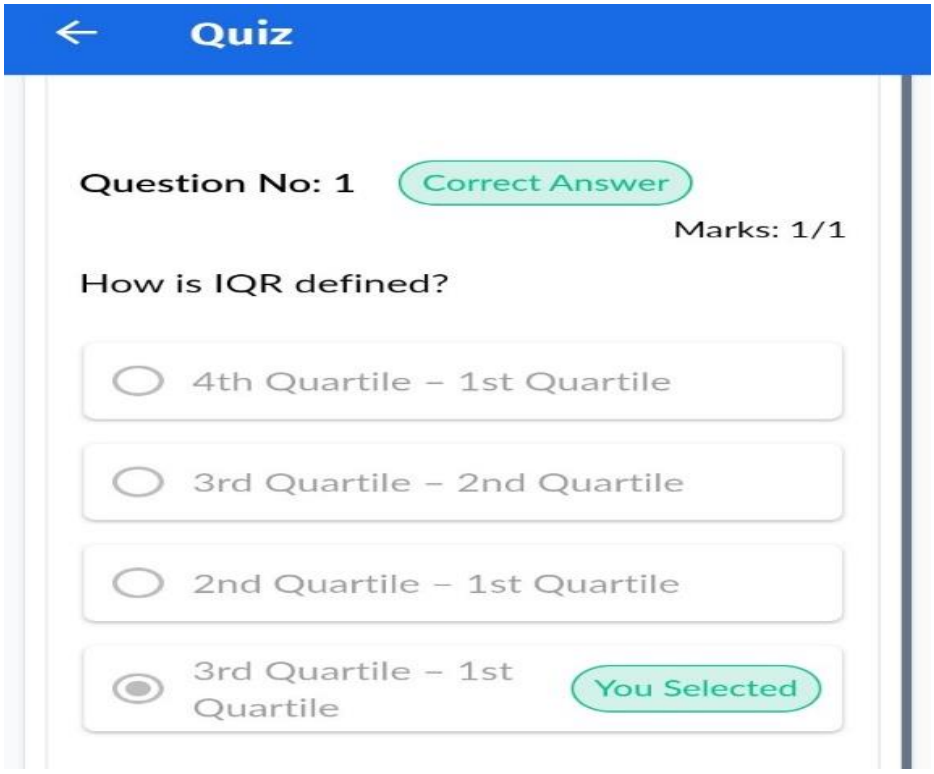
$$c\, r_{A,B} = \frac{\sum_{i=1}^{n}(a_i - \overline{A}(b_i - \overline{B})}{(n-1)\sigma_A \sigma_B} = \frac{\sum_{i=1}^{n}(a_i b_i) - n\overline{A}\overline{B}}{(n-1)\sigma_A \sigma_B}$$

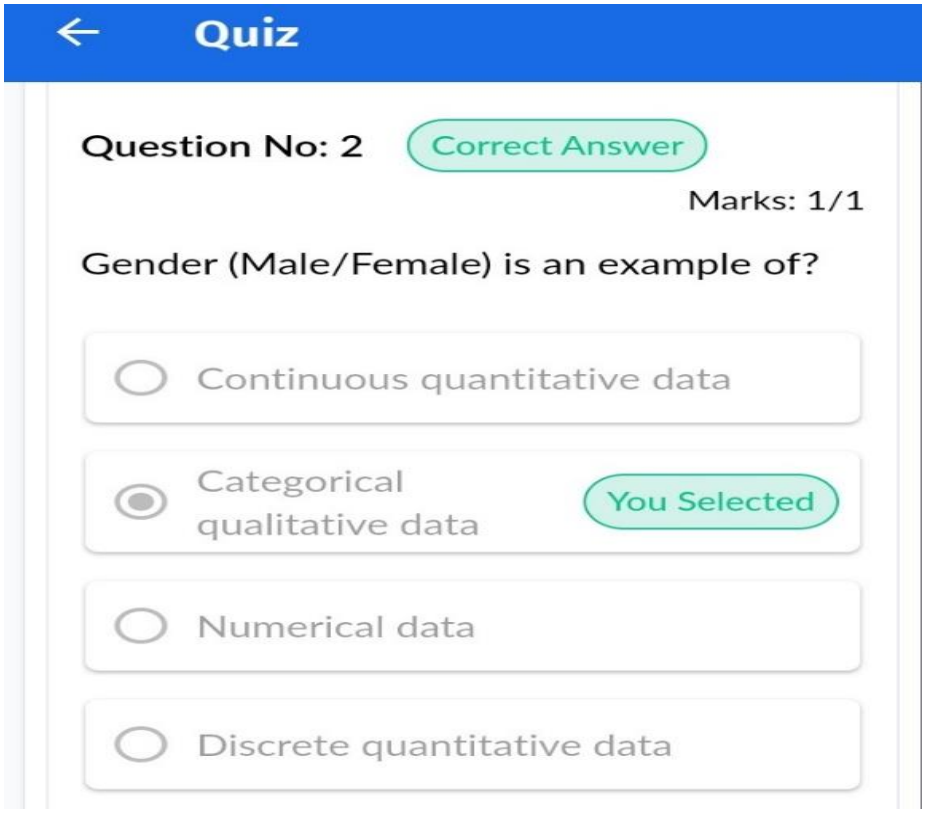
Where n is the number of tuples, $\bar{A}$ and $\bar{B}$ are the respective means of A and B, $\sigma A$ and $\sigma B$ are the respective standard deviation of A and B,and $\sum(a,b)$ is the sum of the AB cross product.

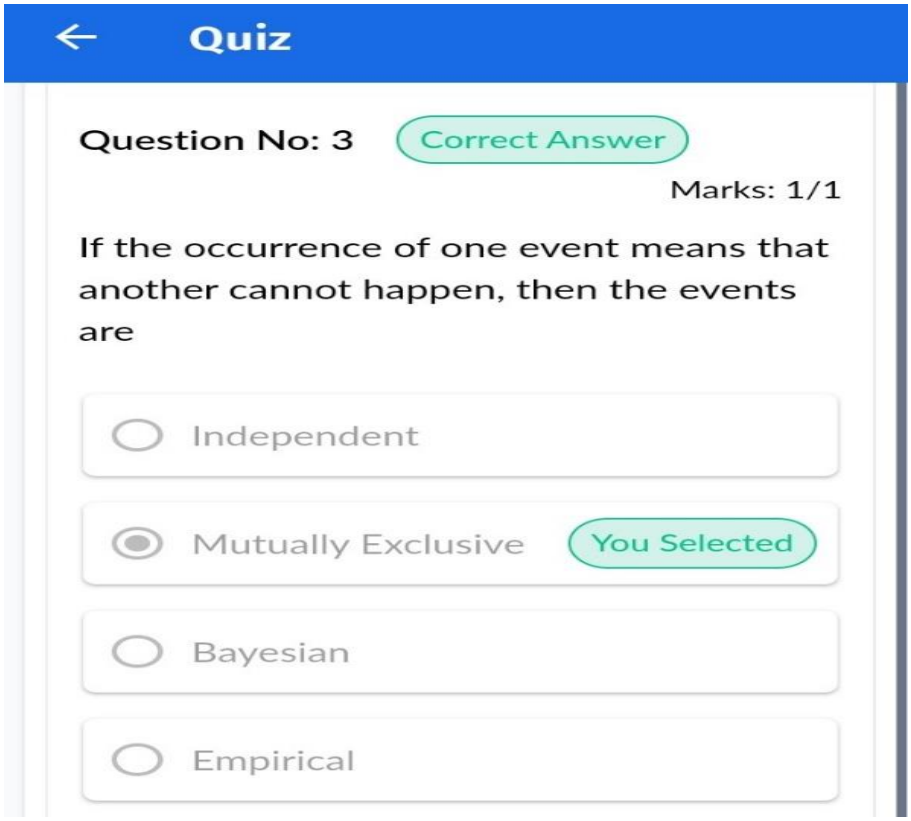
If $r_{A,B} \rangle 0$ A and B are positively correlated(A's values increases as B's). The higher the stronger correlation.

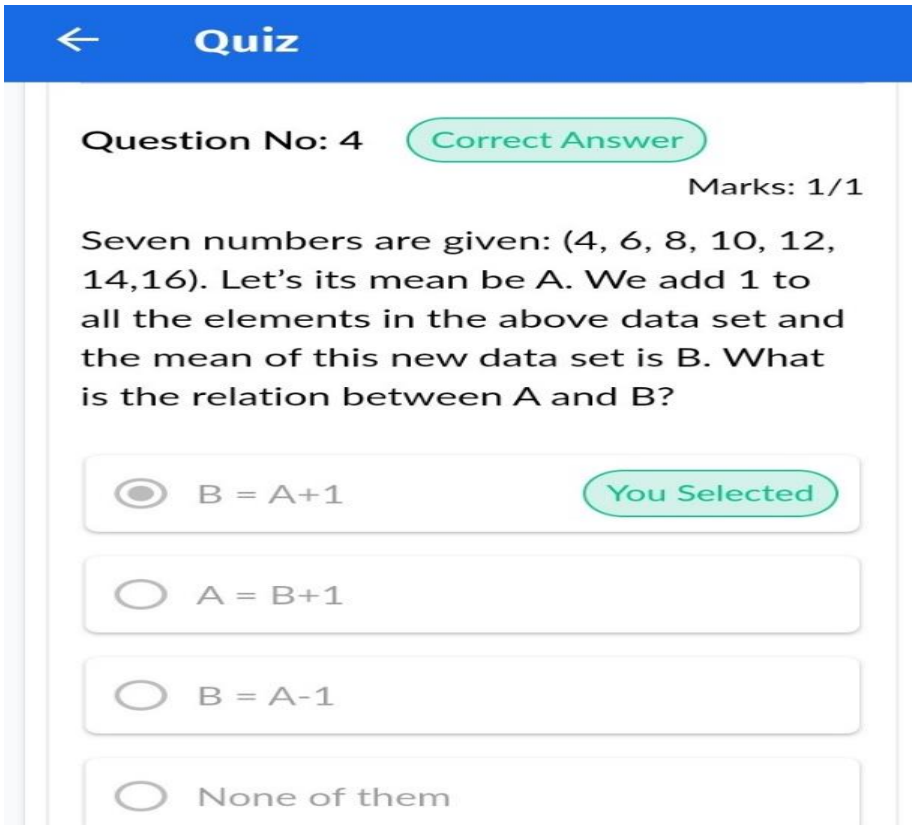
$r_{A,B} = 0$; independent; $r_{AB} \langle 0$ :negatively correlated.

**Quiz**


Quiz
Question No: 1
Correct Answer
Marks: 1/1
How is IQR defined?
4th Quartile – 1st Quartile
3rd Quartile – 2nd Quartile
2nd Quartile – 1st Quartile
3rd Quartile – 1st Quartile
You Selected


Quiz
Question No: 2
Correct Answer
Marks: 1/1
Gender (Male/Female) is an example of?
Continuous quantitative data
Categorical qualitative data
You Selected
Numerical data
Discrete quantitative data


Quiz
Question No: 3
Correct Answer
Marks: 1/1
If the occurrence of one event means that another cannot happen, then the events are
Independent
Mutually Exclusive
You Selected
Bayesian
Empirical


Quiz
Question No: 4
Correct Answer
Marks: 1/1
Seven numbers are given: (4, 6, 8, 10, 12, 14,16). Let's its mean be A. We add 1 to all the elements in the above data set and the mean of this new data set is B. What is the relation between A and B?
B = A+1
You Selected
A = B+1
B = A-1
None of them

## CONCLUSION

The internship was a useful experience . It helped to gain new knowledge and skills. It provides a path to achieve several of our learning goals.This also helpful to learn and to calculate some of the statistical methods. I have learned more about statistics in Machine learning.