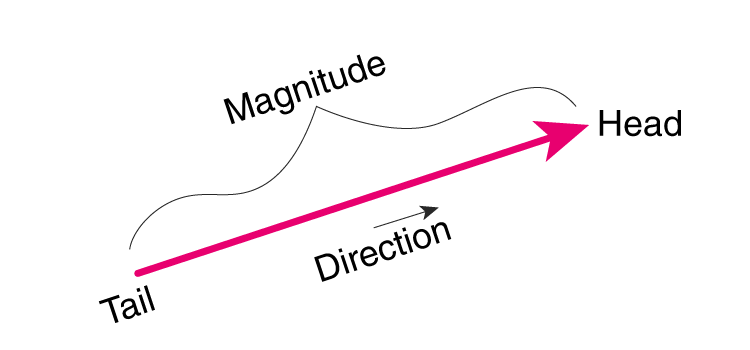
**Problem Statement: Write the Solutions to the Top 50 Interview Questions and Explain any 5 Questions in a Video**

### ****1. What is a vector in mathematics?****

Vector is a mathematical quantity . It has both magnitude and direction . Its representation is ‘a line segment with an arrow’. Vector has head and tail . Tail is the initial point and the head is terminal point. Initial point is where it starts and terminal point is that which tells the final position of the point.



The direction of vector is from tail to head. Vectors are represented in bold lowercase such as a or using an arrow over the letter  a⃗ or b⃗.

Vectors can also be denoted by their initial and terminal points with an arrow above them, for example, vector AB can be [denoted](https://www.cuemath.com/geometry/representation-of-vector/) as **AB⃗**

Magnitude of the vector a is represented by **∥a∥**

Vectors have length that is magnitude and and has direction. In layman terms, if a car is moving at speed of 50km/hr and its going towards home , then it is called as vector quantity.

### ****2. How is a vector different from a scalar?****

Scalar quantity has only magnitude but vector quantity has both magnitude and direction.

Scalar quantities can be mass, temperature, speed and Vector quantities can be force, acceleration, velocity

|  |  |
| --- | --- |
| Scalar | Vector |
| It has only magnitude | It has both magnitude and direction |
| It is one dimensional only | It can be in multidimensional |
| Withchange in quantity has change in magnitude | With change in quantity has change in either magnitude or direction or both. |
| Mathematical scalar operations result in scalar only | Vector operations can produce both scalar and vector result. Product of two vectors is a scalar. Subtraction,cross product and summation of two vectors is a also a vector |
| Example -Speed,Distance | Example - Velocity, force |

Let’s conclude the whole explanation and definition with some examples for better understanding.

Suppose a car is moving at a speed of 100 mph, then it is a scalar quantity because speed is given there that is magnitude but we don't know in which direction the car is going.

A person is walking towards his home with a speed of 1mph, here we are given with both direction that is home and also magnitude that is speed so it's a vector quantity.

### ****3. What is the magnitude of a vector?****

Magnitude of a vector is length of the vector. It is a scalar quantity and it is always positive. Example – in vector A, magnitude is represented as |A|.

Example - a vector is defined as xi + yj then its magnitude is the square root of the sum of squares of the individual terms. It is calculated using Pythagorean theorem in eucliedean space.

### ****4. What is the difference between a square matrix and a rectangular matrix?****

Main difference between square matrix and rectangular matrix is-

Square matrix has equal number of rows and columns and rectangular matrix has unequal number of rows and columns.

|  |  |
| --- | --- |
| Square Matrix | Rectangular Matrix |
| Its representations is as n×n, where n is number of rows (or columns). | Its representation is m×n, where m is number of rows and n is number of columns. |
| A square matrix represents linear transformations that preserve the dimension of the vector space | Rectangular matrix represents linear transformations that change the dimension of the vector space. |
| Matrix A has 3 rows and 3 columns(3x3), Matrix B has 4 rows and 4 columns (4x4), so both are square matrix | Matrix A has 4 rows and 3 columns, denoted as (4x3)and in Matrix B it has 3 rows and 5columns denoted as (3x5), so both have unequal number of rows and columns and are rectangular matrix. |
|  |  |

It can be concluded in the end that shape and size is the basic difference between square matrix and rectangular matrix.

### ****5. What is an eigenvector in linear algebra?****

In linear algebra ,eigen vector is a non zero vector when multiplied by square matrix or any linear transformation is applied then it results in scalar vector , it means the direction is not changed but magnitude may be changed Eigenvectors represent directions in space that remain unchanged when transformed by the matrix.

Eigenvector equation A**v** = λ**v**

**A is square matrix , v is non zero eigen vector ,** λ **is scalar and eigen value of matrix A**

**Eigen vector and eigen values are used in many applications in linear algebra ,**solving linear differential equations,PCA

### ****6. What is probability theory?****

Probability theory is a branch of mathematics that deals with probabilities associated with a random phenomenon which can have several outcomes.

It uses some fundamentals such as sample space, probability distributions, random variables, etc. for finding the likelihood of occurrence of an event. It also uses random variables and probability distributions to assess uncertain situation and it assigns a numerical description to the likelihood of occurrence of an event.

We can calculate probability by dividing the number of favorable outcomes to the total number of possible outcomes .

### Lets understand with example : We roll a fair dice and wants to know the probability of getting 5. The total number of possible outcome is [1,2,3,4,5,6]. So there are total 6 outcomes and count of favorable outcome is 1.

### So probability of getting 5 on dice is 1/6 =0.16

### 

### ****7. What is conditional probability, and how is it calculated?****

### **The conditional probability states that probability of event given that another event has already occurred.**

### **If there are two events A and B , to calculate conditional probability formula is :**

### P(A|B)=P(A∩B)/P(B)

### **P(A|B) the conditional probability of event A given that B event has already occurred,**

### **P(B) Probability of event B**

### **P(A∩B) -Probability of both events occurring together**

### 

### **Let’s understand with example :**

### **Tossing a fair coin 3 times.**

### **Sample space HHH ,HHT,HTT,TTT,TTH,THH,HTH,THT**

### **Sample space is 8**

### **First Event (E) is atleast 2 tail occurs ,the probability P(E) is 4/8 that is 1/2**

### **Another event (F) is First coin shows head ,Probabilty is P(F) 4/8 that is 1/2**

### **Find probability of event E given that F event has taken place.**

### **P(E/F) = P(E** ∩ **F) / P(F)**

### **So P(E** ∩ **F) will be 1/8 as it is HTT (both event occurring together)**

### **P(E/F) = (1/8)/1/2 =1/4**

### ****8. What is Bayes theorem, and how is it used?****

### **Bayes theorem is called as formula for probability of causes.It helps in determining probability of event that is based on some event that has already occurred.**

### **It finds the conditional probability of event A given that event B has already occurred. It calculates probability based on hypothesis.**

### 

### Uses of Bayes theorem

### Medical diagnosis : It is used to help doctors to make accurate diagnosis. With certain symptoms it finds the probability of diseases

### Weather forecasting : It is used to get the probability of future weather condition on basis of current data

### Spam filter : It is used to get probability that incoming email is spam or not based on its content

### ****9. What is a random variable, and how is it different from a regular variable?****

### Random variable :Random variables are used to describe the possible outcomes of a random process. A random variable is a real valued function that maps each outcome of a random process to a numerical value.

### Regular variable : A regular variable has a fixed value and can be determined from given information. It is used in algebra and other branches of mathematics.

### Lets understand with example.

### When we roll a dice , each time the outcome can be any number from 1 to 6 . Suppose a random variable is X , possible values of X can be 1to 6 , it is random because outcome of rolling dice is uncertain.

### In regular variable lets suppose x=x+5 then with this information we find x that is 5 which is fixed value.

### Two types of random variable: Discrete and Continuous

### Discrete random variable : Discrete random variable can take on a countable number of values. Examples : the number of heads in multiple coin flips .The probability distribution of a discrete random variable can be described using a probability mass function (PMF), which assigns probabilities to each possible outcome.

### Continuous random variable :Continuous random variable can take on any value in a continuous range . Example : weight of any randomly selected person. The probability distribution of a continuous random variable can be described using a probability density function (PDF), which specifies the likelihood of observing values within different intervals.

### ****10. What is the law of large numbers, and how does it relate to probability theory?****

### **Law of large numbers( LLN) in probability states that as the sample size increases , sample mean tends to close to population mean . When we repeat a experiment multiple number of times and average the result we get the value that is close to expected value.**

### **Lets understand it with help of example.**

### **If we flip a coin there is 50% chances of getting head and tail. So 50% is expected value (population mean)**

### **If we toss a coin 4 times we get 1 times head so probability of head is 25%**

### **Now we toss a coin 100 times we get 47 times head so probability of head is 47%**

### **Now again tossing 1000 times we get 543 head so probability is 54.3%**

### **Here we are increasing our sample size (number of trials) that is tossing of coin , we are tend to approach to 50% that is population mean.**

We can conclude that it forms the basis for the statistical inference that allows us to make predictions and estimate parameters based on sample data. The LLN illustrates the principle that larger samples provide more accurate estimates of population parameters.

### ****11. What is the central limit theorem, and how is it used?****

### **The Central Limit Theorem (CLT) states that the distribution of the sample mean will always be normally distributed , as long as the sample size is large enough.The sampling distribution of mean will always be normal, regardless of whether population has normal, binomial or any other distribution.**

### **The conditions required for the CLT to hold are:**

### **1. Sample size should be large enough, greater than or equal to 30.**

### **2. Sample size should be drawn from a finite population or an infinite population with a finite variance.**

### **3. The random variables in the sample are independent and identically distributed.**

### 

### **Uses of CLT :**

### **It is used to make probabilistic inferences about a population based on a sample of data.**

### **We can use the CLT to construct confidence intervals, perform hypothesis tests, and make predictions about the population mean based on thesample data.**

### **It provides a theoretical justification for many statistical techniques, such as t-tests, ANOVA, and linear regression.**

### ****12. What is the difference between discrete and continuous probability distributions?****

### **Discrete probability distribution :The likelihood of occurrence of each possible value of discrete random variable is given by discrete probability distribution . Probabilty mass function and probability distribution function are associated with it.**

### **Example :If we roll a dice then possible outcome can be {1,2,3,4,5,6} .All numbers have equal chance , so probability of getting any one number is 1/6.**

### **The continuous probability distribution given the probability of continuous random variable .In continuous probability distribution continuous random variable can take values which can be inifinte and uncountable.**

### **Example : time is infinite it can be from 0 to trillion seconds , weight of person**

### 

### Discrete distributions are suitable for scenarios where outcomes are distinct and countable, such as counting the number of heads in a series of coin flips.

### Continuous distributions are more appropriate for situations where outcomes are measurements or observations that can take any value within a range, such as heights or weights of individual.

### ****13. What are some common measures of central tendency, and how are they calculated?****

### Common measures of central tendency are mean , median and mode.

### Mean : It is also called as average. It is calculated by adding all the values and dividing it by the count of values present.

### Mean = sum of values / count of values

### For example : Marks of 8 students in class are 24,26,50,78,90,45,50,88,60,65

### Mean =(24+26+50+78+90+45+50+88+60+65) / 8 =576/8 = 72

### Median : It is the middle value of dataset in which data set is arranged in ascending order. If the dataset has even number of values then median is mean of the middle two values.

### For example : The dataset contains number 45,10,12,96,1,30

### First arrange in ascending order 1,10,12,30,45,96

### It contains even number of observations. The middle two values are 12,30 . So the median will be mean of these two numbers (12 + 30) / 2 = 21

### Mode :It is the most frequently occurring value in dataset . There can be one mode called as unimodal , two modes called as bimodal and also multiple modes called as multimodal and can be zero mode also.

### Example : 5,7,2,5,9,3,5 Mode is 5

### ****14. What is the purpose of using percentiles and quartiles in data summarization?****

Percentiles divide a dataset into 100 equal parts, each representing 1% of the data.

Quartiles divide a dataset into four equal parts, each representing 25% of the data. The first quartile i.e. Q1 is the 25th percentile, the second quartile i.e Q2 is the 50th percentile (also known as the median), and the third quartile i.e. Q3 is the 75th percentile.

Purpose of using percentile and quartiles :

* You can compare quartiles and percentiles to assess how data sets differ in terms of central tendency and spread.
* The position of quartiles can indicate the skewness of the data distribution.
* If Q1 and Q3 are closer to the median, the distribution is likely symmetric.
* If one quartile is significantly farther from the median, it suggests skewnes
* Percentiles and quartiles are less sensitive to outliers than the mean and standard deviation.
* By comparing the observation to the quartiles, we can determine whether the observation is in the bottom 25%, middle 50%, or top 25%

### ****15.What is the covariance of a joint probability distribution?****

### **Covariance of joint probability distribution is a measure of how much two random variables change together.**

### **It measures the linear relationship between variables but it doesn’t provide any information about strength of relationship.**

### **It quantifies the degree to which variables tend to move in relation to each other.**

### **If there are two random variables X and Y , the joint probability distribution is P(X,Y)**

### **Covariance is denoted by Cov (X , Y) =E [(X – μx)(Y- μy)]**

### **Here E is expected value**

### **μx is mean of variable X**

### **μy is mean of variable Y**

### **If covariance is positive it means that if one variable increases then other variables also increases.**

### **If covariance is negative it means that if one variable increases then other variables will decrease.**

### **If covariance is close to zero it means that there is little or no linear relation between variables.**

### ****16. What are the different sampling methods commonly used in statistical inference?****

### **Different sampling methods used in statistical inference are :**

### **1)Simple random sampling:**

### **Every member of population has equal chance of being selected .Selection of items completely depends on selection of chance and randomness.**

### 

### **Example :Population of different age, sex, occupation**

### **We select sample from population based on randomness and chance,it removes bias from selection process.**

### **2)Systematic sampling**

### **The first element is selected randomly from list then every nth element is selected.**

### **Example :Broad category of people , first sample is selected randomly then every 2nd member from population.**

### 

### **3)Cluster sampling**

### **Randomly selecting cluster of elements from population and then selecting every element in each selected cluster.**

### **The researcher will divide the population into separate groups called cluster which could be group of externally homogenous but internally heterogenous groups then simple random sample of cluster is selected from population .**

### 

### **Example :Broad category of population .Divide entire population into groups, which is externally heterogenous but internally homogenous . After identifying cluster we will pick all element of selected cluster**

### **4)Stratified sampling**

### **In this method first researchers divide population into homogeneous groups called strata. After that they draw a**[**random sample**](https://statisticsbyjim.com/glossary/sample/)**from each group i.e. stratum and combines it to form their complete representative sample.**

### 

### **Example :Divide population into group male ,female ,elderly people .Select sample from these sample. Selecting atleast one person from each group.**

### 5) Convenience sampling :

### It involves selecting sample based on convenience.

### It simply includes indiviuals who are most accessible to researchers . This is easy and inexpensive way to gather data.

### 

### 6)Snow ball sampling:

### It is used in qualitative search . Select samples and ask them to refer to you to others. It is also referred as network sampling

### 

### 7)Quota sampling

### It means to take a very tailored sample that’s in proportion to some charactersitc or trait of population

### 

### Example : divide by state they live in based on education on income level. It is often used by market researchers where interviewers are given quota of subject of specific type to attempt to recruit.

### 8)Judgemental sampling

### 

### We select samples based on his or her judgement. This technique lies on the researchers who chose the sample based on his own experience this approach is often used by media .

### ****17. What is the difference between parameter estimation and hypothesis testing?****

**In parameter estimation we calculate an estimate of unknown parameters based on sample data using mean , variance, standard deviation while in hypothesis testing it is used to make inferences about population parameter or test specific hypothesis about population.**

**In parameter estimation statistical method such as maximum likelihood estimation are used while in hypothesis testing null hypothesis and alternate hypothesis are used.**

**The outcome of parameter estimation is point estimate or interval estimate while the outcome of hypothesis testing is to reject or not null hypothesis based on the evidence drawn from sample data**

**Example :**

**To calculate average height of all students in class we can use estimation method.**

**A juice company claims that it contains 500ml of raw fruits extract , we can use hypothesis testing to find if this is true or not based on sample.**

### ****18. What is the p-value in hypothesis testing?****

### **P-value is a probability between 0 and 1 calculated with the assumption that the null hypothesis is true . It is an evidence against null hypothesis .**

### **P-value is also known as Probability value . It is the smallest level of significance at which null hypothesis is rejected .They are generally expressed as decimal.**

### **Smaller the p value stronger the evidence that you should reject null hypothesis. Usually p-values of 0.05 is used. In percentage it is 5% it means that there is 5 percent chance that your result could be random and would have happened by chance.**

### 

To make a decision in hypothesis testing, you compare the p-value to the chosen significance level (α). The decision rule is as follows:

* If p ≤ α, you reject the null hypothesis (i.e., evidence suggests that the null hypothesis is unlikely to be true).
* If p > α, you fail to reject the null hypothesis (i.e., the evidence is not sufficiently strong to reject the null hypothesis).

### ****19. What are Type I and Type II errors in hypothesis testing?****

In hypothesis testing there are two types of errors:  type I and type II errors. These errors result in incorrect conclusions.

## Type I Error :

It it called as false positive or error of the first kind.

In the type I error researchers reject a null hypothesis when its actually true.

The probability of making this error is the significance level (alpha or α). The significance level is at 0.05 or 5%. So if the null hypothesis is actually true ,results will have 5% chance of occurring or less than that.

The result are statistically significant and consistent if p value is less than the significance level with the alternative hypothesis.

The result are statistically non significant if p value is higher than the significance level.

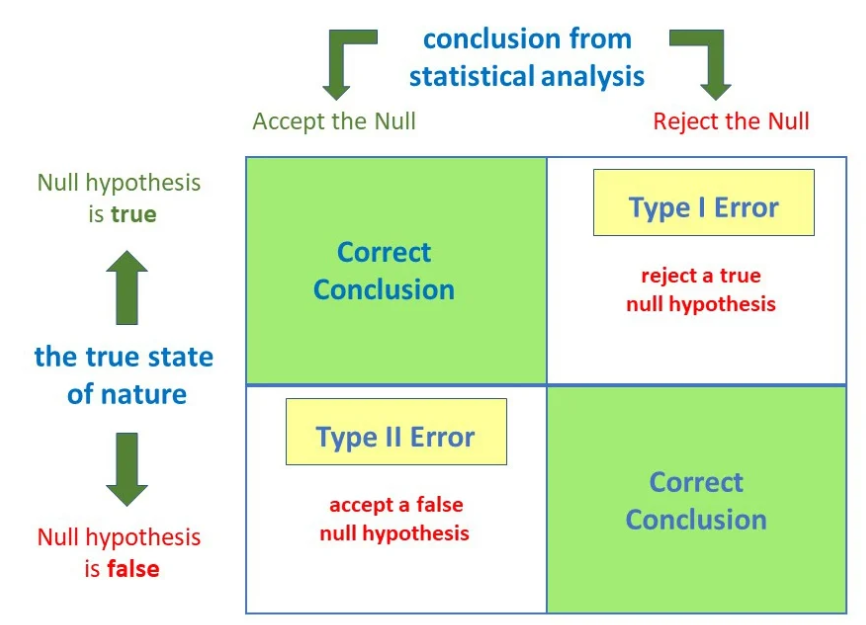
Type II Error :

It is called as false negative or error of the second kind.

When the null hypothesis should be rejected but it is accepted then this error happens.

The probability of making this error is Beta (β) that is related to power of the statistical test i.e. (power = 1- β).

If you ensure that test has enough power means that if sample size or the significance level is increased then risk of commiting this error can be reduced. Here power level of 80% or higher is acceptable.



### Example:

### Tested for covid based on mild symptoms

### There are two errors that could occur:

### The test result says you have corona but you actually don’t . It its Type I error(false positive)

### The test result says you don’t have corona but you actually do. It is Type II error(false negative)

### ****20. What is the difference between correlation and causation?****

### Correlation is the relationship between variables. There is positive , negative and zero correlation. If one value goes up then other also goes up or vice-versa then it is positive correlation. If one value goes higher and other goes lower then it is negative correlation. If no relation exist then it is zero correlation.

### Causation is that where one variable directly influence change in another variable. There is cause and effect relation between variables.

### Correlation doesnot imply causation. If correlation exist between variables it doesnot mean that one variable has caused change to other variable.

### 

### Example :

### Data is collected of sales and it is found that there is positive correlation between sale of icecream and sale of sunscreen. If we try to plot a scatter plot to visualize the relationship we can see that as the sale of icecream increases sale of sunscreen increases. It doesnot imply that increase in icecream has caused people to buy sunscreen, there is no direct casual link between icecream sales and sunscreen sale. There is a third variable that is summer weather because of that sales increase in both product.

### ****21. How is a confidence interval defined in statistics?****

### **Confidence interval is range of values within which we expect a particular population parameter like mean to fall , it’s a way to express the uncertainity around estimate obtained from sample of data**

### **Confidence level is usually expressed as percentage that indicates how sure we are that the true value lies within the interval**

### 

### **Lets understand with example.**

### **If a youtube channel has number of subscribers and average age of them lies between 25 to 32 so this is confidence interval and confidence level is 95 % it means that we are 95 % confident that average age is between 25 to 32.**

### **2**2. What is hypothesis testing in statistics?****

### **Hypothesis testing is used to make conclusion based on sample data about the population parameter.**

### **It is an educated guess to detect significant differences by comparing sample statistics with population parameters. It should be testable either by experiment or observations. It is important to note that we are not making an absolute conclusion but we are either concluding the result we are getting are due to chances.**

### **It is used to infer the result of hypothesis performed on sample data from larger population by detecting significant difference and that difference didn’t occur by random choice.**

### **Null hypothesis and alternate hypothesis are two different types of hypothesis testing.**

### ****23. What is the purpose of a null hypothesis in hypothesis testing?****

### **In hypothesis testing, we state the assumed or hypothesised value of the population parameter before we begin sampling. The assumption we wish to test is called null hypothesis and is denoted by Ho**

**It refers to as Status Quo.**

H0:  p = p0 **This equation means that there is no difference between two population parameters**

### **If difference is large enough assuming that** H0 **is true then we have to reject** H0 **and conclude that there is difference between population.**

### **Example : we want to test the hypothesis that the population mean is equal to 500 . The null hypothesis is that population mean is 500 is written as:**

### H0: **u =500**

### ****24. What is the difference between a one-tailed and a two-tailed test?****

### **Two tailed test :**

### **When the test of hypothesis is made on the basis of rejection region represented by both sides of standard normal curve,it is called two tail test or two sided test.**

### **In simple words we can say that a test of statistical hypothesis where alternate hypothesis H1 is two sided. It is a non directional hypothesis . The specified parameter can be greater than or less than mean value.We use not equal to sign .**

### 

### **At both ends of curve the critical region has two parts. A two tailed test is applied in such cases when difference between sample and population statistic is tending to reject H0, the difference may be + or –.**

### **Example:**

### **There is a manufacturing company , it manufactures juicer parts , we have to check if if weighs 15 gms. The null hypothesis is that mean part weight is 15 gms . The alternative hypothesis is that the mean weight of part is not equal to 15 gm**

### **One tailed test :**

### **A test of statistical hypothesis where the alternate hypothesis is one sided is called one tailed or one sided test.**

### **It is unidirectional hypothesis where area of rejection is on sampling distribution.**

### 

### **The critical region is on one side either left tailed or right tailed. It has defined direction. The specified parameter can be either greater than or less than mean.**

### **We use either > or < sign for alternative hypothesis.**

### **One tailed test are of two types:**

### **Right tailed test : The rejection region or critical region lies entirely on right tail of normal curve.**

### **In left tailed test , critical region or rejection lies entirely on left tail of normal curve**

### **Example : There is a manufacturing company , it manufactures juicer parts , we have to check if if weighs less than 15 gms. The null hypothesis is that mean part weight is greater than or equal to 15 gms . The alternative hypothesis is that the mean weight of part is less than 15 gm.**

### ****25. What is the geometric interpretation of the dot product?****

The dot product of two vectors is a mathematical operation that takes two equal-length sequences of numbers (vectors) and returns a single number. Geometrically, the dot product has a significant interpretation related to the angle between two vectors.

  The geometric interpretation can be understood in terms of the angle θ:

* Parallel vectors (θ = 0∘ When the vectors are parallel, the dot product is maximized, and the cosine of θ is 1. This means that A⋅B=∣A∣⋅∣B∣, and the vectors are pointing in the same direction.
* Perpendicular vectors (θ = 90∘)  When the vectors are perpendicular, the dot product is zero, as the cosine of θ is 0. This indicates that the vectors are orthogonal or at a right angle to each other.
* Antiparallel vectors (θ = 180∘) When the vectors are pointing in opposite directions, the dot product is minimized, and the cosine of θ is -1. This results in A⋅B=−∣A∣⋅∣B∣.

## Geometrical Interpretation of Dot Product of Two Vectors

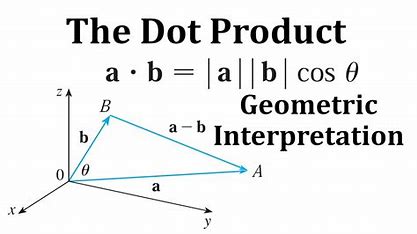
There are two vectors P and Q.

P = a1i + a2j + a3k , and Q = b1i + b2j + b3k

Then, P.Q will define the scalar product.

P.Q = (a1i + a2j + a3k).(b1i + b2j + b3k)

P.Q = a1b1 + a2b2 + a3b3



The angle between two vectors which is formed by intersection at one point

Cosθ = (a.b)/|a|\*|b|

### ****26. What is the geometric interpretation of the cross-product?****

The cross product is a binary operation on two vectors in three-dimensional space. It produces a vector that is perpendicular to the plane containing the input vectors. The magnitude of the cross product is equal to the area of the parallelogram formed by the input vectors, and the direction is determined by the right-hand rule.

Geometrical Interpretation of Cross Product of Two Vectors

Suppose there are two vectors, P and Q.

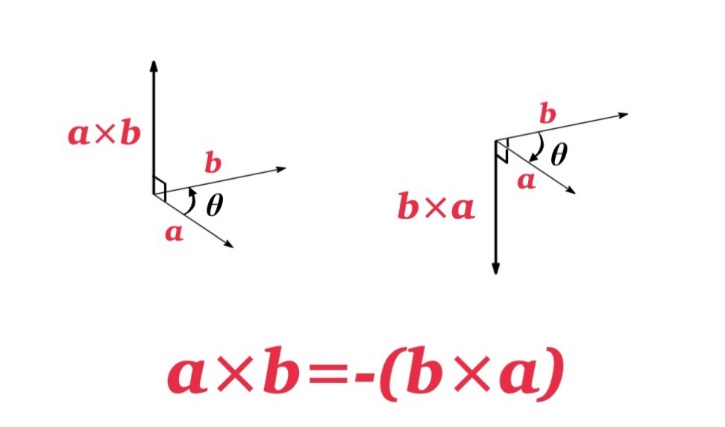
P = a1i + a2j + a3k, and Q = b1i + b2j + b3k

Then, P x Q will define the vector product.

P x Q = (a1i + a2j + a3k) x (b1i + b2j + b3k)

P x Q = (a2b3 – b2a3)i – (a1b3 – b1a3)j + (a1b2 – b1a2)k

If we place vector an on vector b from the point of their intersection, then the length of vector b occupied by vector a is the projection of vector a on vector b.

****

The resulting vector, A×B, is perpendicular to the plane formed by vectors A and B. The direction is determined by the right-hand rule, where you align your index finger with A, your middle finger with B, and your thumb points in the direction of A×B.

### ****27. What are observational and experimental data in statistics?****

### **Observational data is collected by observing what is happening and record the observations. There is no manipulation of researcher. It is used for descriptive and exploratory purposes. In observational study it may show that there is relation between variable but it is not necessary that it is cause and effect relation.**

### **Example : Study relation between smoking and lung cancer.**

### **The researchers collect observational data by surveying a large sample of adults about their smoking habits and monitoring their health outcomes over time.They are asked questions about their smoking history, number of cigarettes smoked per day, the duration etc. The researchers do not control their smoking behaviors.They observe and record smoking habits and health outcomes of the participants.After collecting the data, the researchers analyze the relationship between smoking habits and the incidence of lung cancer using statistical methods**

### **Experimental data is collected when researchers manipulate one or more variables to see what is effect on other variable , so here experiments are controlled and manipulated. This leads to cause and effect relation between variables. It is also called as scientific study.**

### **It involves two groups i.e. experiment group and control group**

### **Example :**

### **There are two groups, group1 and group2 of office employees. Group 1 was asked to follow strict routine for fixed period of time and group 2 was asked to follow normal routine as they used to have earlier. The researchers looked at which group will higher productivity in performance.**

### ****28. How are confidence tests and hypothesis tests similar? How are they different?****

### **Similarity in confidence tests and hypothesis tests**

### **The confidence test and hypothesis test both uses sample data collected from population to draw conclusions about population parameters. They both are statistical methods and involves calculation of probabilities to access the likelihood of certain outcomes.**

### **The difference between confidence tests and hypothesis test are:**

* Confidence test estimates a range of values that is confidence interval which contains the true value of a population parameter with a specified level of confidence. **In hypothesis testing it is used to make inferences about population parameter or test specific hypothesis about population. the outcome of hypothesis testing is to reject or not null hypothesis based on the evidence drawn from sample data**
* In a confidence test, the result is expressed as a confidence interval, which is range of values for the population parameter. In a hypothesis test, the result is expressed as a p-value,which is the strength of evidence against the null hypothesis.
* In a confidence test, the decision-making process focuses on confidence interval. In a hypothesis test, the decision-making process involves comparing the calculated p-value to a predetermined significance level.

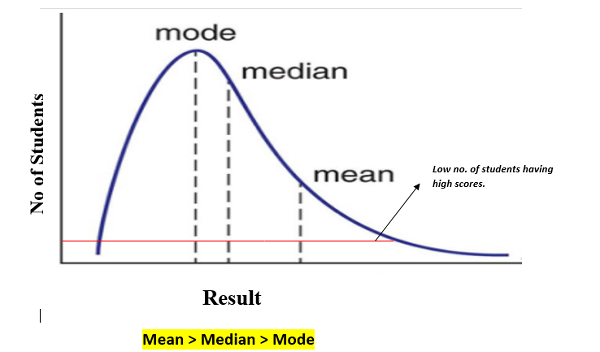
### ****29. What is the left-skewed distribution and the right-skewed distribution?****

Left-skewed and right-skewed distributions refer to the shapes of probability distributions.

Positive Skewed or Right-Skewed  (Positive Skewness) :

It has long right tail with the measures are dispersing . The mean , median , mode of distribution are positive instead of negative or zero.

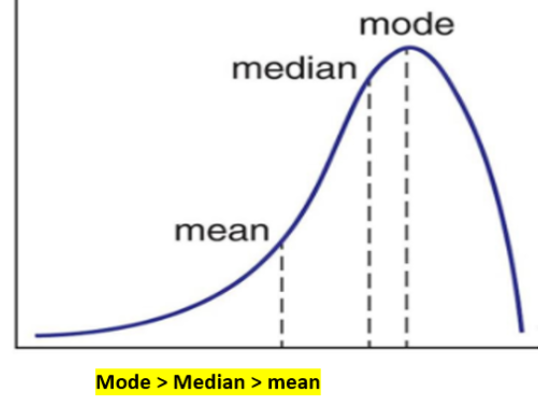
The mean is greater than median . The result is bent towards the lower side. So the mean will be more than median .

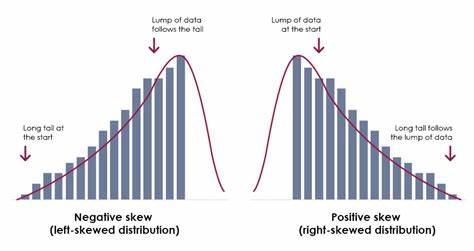


Negative Skewed or Left-Skewed :

It refers to the distribution model where tail of distribution is spreading on left side and more value are plots on right side of graph . It has long left tail.

The median is higher than mean due to unbalanced distribution. Mean , median and mode are negative instead of positive or zero.





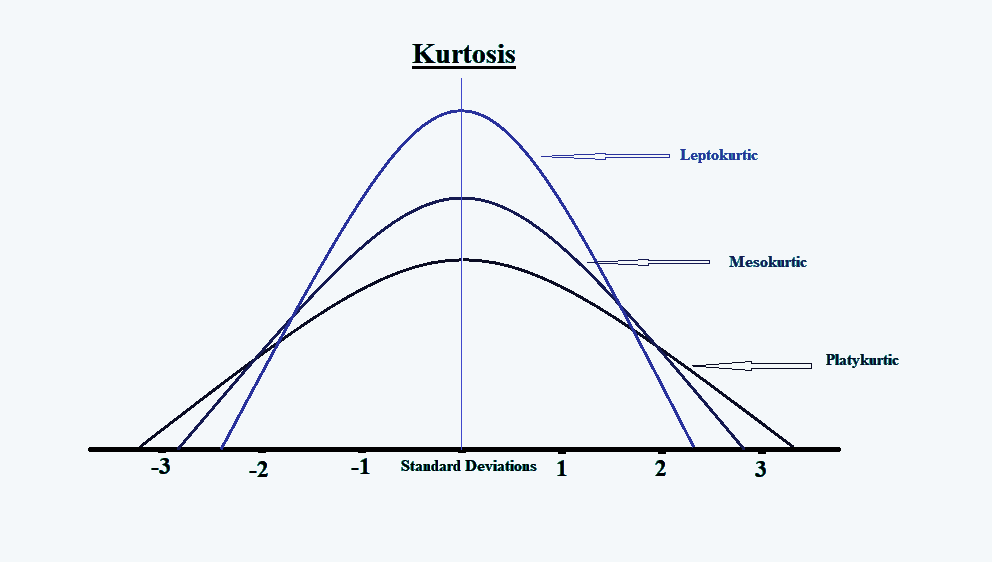
### ****30. What is kurtosis?****

Kurtosis is a statistical measure that quantifies the shape of probability distribution. It gives information about tails and peakedness of the distribution as compared to normal distribution . It is used in analysis of charactersitics and outliers of dataset.

The measure of kurtosis is tailedness of distribution and tailedness refers to how often outliers occurs.

Peakedness is degree to which data value are centered around mean. Dataset with low kurtosis have flat top near the mean and dataset with high kurtosis have distinct peak near mean which decline rapidly with heavy tails.

Positive kurtosis means heavier tails and more peaked distribution. Negative kurtosis means lighter tails and flatter distribution.



Mesokurtic, Leptokurtic, and Platykurtic are three different types of kurtosis

* Leptokurtic (Kurtosis>3): It is heavy tailed distribution i.e. more than normal distribution. It has long and thick tails , so it has more chances of outliers.
* Mesokurtic (Kurtosis =3): It has same as normal distribution. So kurtosis is near 0.
* Platykurtic (kurtosis<3): It is short tailed distribution which is less than normal distribution. It has thin tail . It is strechted around center , so most data points are in high proximity to mean

### ****31. What is the probability of throwing two fair dice when the sum is 5 and 8?****

### **When two dice is rolled , possible number of outcomes or sample space is 36.**

### **Formula for calculating probability :**

### **No of favourable outcomes / Total number of outcomes**

### **When sum is 5 , outcomes are (1,4) , (2,3) , (4,1), (3,2)**

### **So number of outcomes are 4.**

### **When sum is 8 , outcomes are (2,6),(6,2),(4,4),(5,3),(3,5)**

### **So number of outcomes are 5**

### **Probability of getting sum is 5 -> 4/36**

### **Probability of getting sum is 8 🡪5/36**

### **the probability of throwing two fair dice when the sum is 5 and 8:**

### **4/36 +5/36 = 9/36 = 1/4**

**32. What is the difference between Descriptive and Inferential Statistics?**

### Descriptive and Inferential Statistics are the types of Statistical analysis.

### Descriptive statistics deals with collection , organisation , analysis, interpretation and presentation of data , it focusses on summarizing and describing the main features of set of data

### Inferential Statistics deals with making conclusions and predictions about population based on sample.

### Descriptive statistics uses methods like measure of central tendency , measure of tendency while inferential statistics uses technique like probability , hypothesis testing, annova test to make predictions.

### Example : Students marks in class

### Descriptive statistics will give the information like average marks, range of marks, most common marks of student while the inferential statistics will predict the performance of students based on marks in future test.

**33. Imagine that Jeremy took part in an examination. The test has a mean score of 160, and it has a standard deviation of 15. If Jeremy’s z-score is 1.20, what would be his score on the test?**

Values given are :

μ = 160

σ = 15

Z = 1.20

Formula : X = μ + Zσ

Here:

μ: Mean

σ: Standard deviation

X: Value to be calculated

**X=160 + (15 \*1.2) =173.8**

**34. In an observation, there is a high correlation between the time a person sleeps and the amount of productive work he does. What can be inferred from this?**

### There is a high correlation between the time a person sleeps and the amount of productive work he does , it means there is positive relation i.e when the variables move together in the same direction. When the time a person sleeps increases, amount of productivity also increases. When time reduces, productivity also reduces.

**35. If there is a 30 percent probability that you will see a supercar in any 20-minute time interval, what is the proba­bility that you see at least one supercar in the period of an hour (60 minutes)?**

### The probability of not seeing a supercar in 20 min : 1- P(Seeing a car) =1-0.3 =0.7

### Probability of not seeing supercar in period of 60 minutes =0.7 \* 0.7 \* 0.7 =0.343

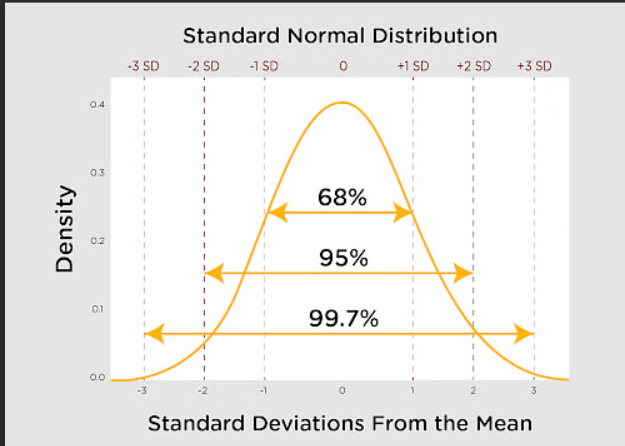
### So probability if seeing atleast one supercar in 60min :

### 1-P(not seeing any supercar) =1-0.343 = 0.657

**36. What is the empirical rule in Statistics?**

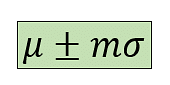
It is a statistical rule that says observed data for a normal distribution will fall within three standard deviations which is denoted by σ of the mean or average which is denoted by µ.

It is also known as three sigma rule or 68-95-99 rule that states that 68% of the data falls within one standard deviation, 95% of the data within two standard deviations and 99.7% data within three standard deviations from mean.



As we can see from the graph that normal distribution is related to it.

Empiral rule formula is:



Here µ is Mean ,σ is Standard deviation and m  is Multiplier

**37. How does increasing the confidence level affect the width of a confidence interval?**

### **Increasing the confidence level will make width of confidence interval more wider. Increasing the confidence level means more certainity and precision that interval will get true population parameter.**

### **Example : To get average height of people in city , we take ample and calculate average height we get 90% confidence interval of 5 feet to 6 feet. If we increase confidence interval to 95 % we need to widen the interval to maintain the higher confidence level , so new interval can be 5 feet 3inches to 5 feet 8inches.**

**38. Can a confidence interval be used to make a definitive statement about a specific individual in the population?**

### **No confidence interval cannot be used to make definitive statement about specific individual in population.**

### **Confidence interval is used for making inferences based on sample data but they don’t provide information about individual member**

### **Example : if we calculate 95% confidence interval for average height of people , we get the range from 160cm to 180cm we cannot say that any particular person has height within this range.**

**39. How does sample size influence the width of a confidence interval?**

### **Sample size influences the width of confidence interval . Confidence interval is narrow when sample size is large and it is wider when sample size is small.**

### **Large sample will be like more representative of population and it will give true parameter value and margin of error decreases . A samller sample will be prone to sampling eroor and variability.**

**40. Can two confidence intervals with different widths have the same confidence level?**

### Yes two confidence intervals with different width can have same confidence level. Width of confidence interval is effected by sample size and variability. A wider interval means more uncertainity while narrow interval means greater precison.

### Let’s understand with example.

### We are trying to catch fish using two different nets. both net have 90% chance of catching fish , one net is wider and other is narrower. So the nets are of different sizes but both give same level of confidence to catch fish . If net is wider it will catch large number of fish and with that it can catch large junk . If net is narrow it catch less fish but also less junk.

**41. What is a t-test?**

It is statistical hypothesis test . It is used to determine if there is any significant difference between means of two groups. To check if observed difference between groups has occurred by chance or are they statistically significant, t -test is used.

It is bases on t-distribution . It calculates t-statistic which is used to measure difference between means of two groups , if t-statistic is large it means that it is difference between groups’mean is not because of random chance.

Assumption for t-test:

* Data should be continuous or ordinal scale.
* Observation in data is randomly selected.
* When we plot the data it is normally distributed.
* Sample size is large.

There are three types of t-tests that can be performed:

* One sample t-test : In this we compare the average i.e mean parameter of one group with set average. Set average is any theoretical value or can be the population mean.
* Independent two-sample t-test :It is used for comparision of means of two different samples.

Example : compare the average scores of the male students to the average scores of the females. So, number of males and females should be equal for this comparison. This is where a two-sample t-test is used.

* Paired sample t-test : Here one group is measured at two different times. Separate means for a group is compared at two different times or under two different conditions.

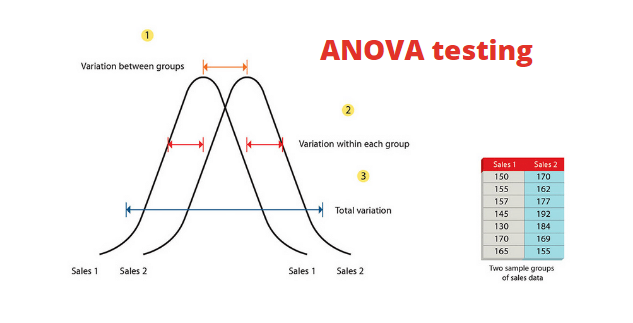
Example : To increase productivity levels of employee a training program was conducted . Now to measure if the productivity levels have increased we will compare the productivity level before and after the training program. So, same sample i.e the employees is compared at two different times which is before and after the training.

**42. What is the ANOVA test?**

ANOVA or Analysis of variance is a technique of testing hypothesis about the significant difference in several population means. The main purpose of analysis of variance is to detect the difference among various population means based on the information gathered from sample means of respective population.

It is also based on some assumption. Each population should have normal distribution with equal variances. The total variation in sample data can be on variance between samples and variance within samples.

Variance between sample is attributed to difference among sample means. Variance within sample is the difference due to chance or experimental errors.



ANOVA provides an F-statistic and a p-value. The p-value indicates whether the observed differences between group means are statistically significant. If the p-value is below a certain significance level (commonly 0.05), you would reject the null hypothesis and conclude that there are significant differences between at least two groups. ANOVA does not tell you which groups are different from each other; if it indicates a significant difference, additional post-hoc tests or pairwise comparisons may be conducted to identify specific group differences.

The technique of Anova is classified into two types:

One way Anova and Two way anova

One Way Anova is used for comparison of two means from two independent groups using F distribution. The null hypothesis for test will be that the two means are equal. So significant result means that two means are unequal.

Two Way Anova is extension of One Way Anova. There are two independents . There is one measurement variable and other is nominal variable. We can use two way Anova where the experiment has quantitative outcome and have two categorical variables.

**43. What is an inlier?**

### **An inliner also known as inlying value or an inlying point, is a data point or observation that fits well within a given dataset or model. Inliers are typically consistent with the majority of the data and are not considered outliers.**

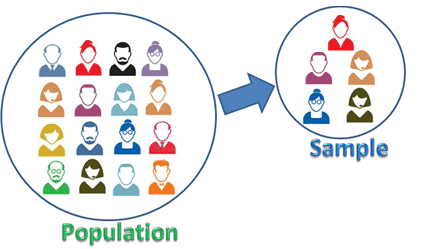
### **It is a point that is considered to be part of the normal behaviour of data. It aligns with the general pattern or trend exhibited by majority of dataset. Inliers contribute to a better understanding of the central tendency and relationship within data**

**44. What is sampling in statistics, and why is it important?**

**Population** : It is the group in which we are interested in studying.

Sample : It is the subset of population that represents whole population

Sampling in statistics is process to select a subset of population that is sample so that we can draw conclusion about larger population from this sample. In this method researchers infer about a population based on results from sample without having to investigate every individual .



Example : College students survey about food.

Suppose there are thousands of students in college. All students will be the population , but its not possible to ask each and every student about quality of food , as it may take weeks or months to gather data from each and every person , it will be time taking and practically impossible , so we select sample from population that represents whole population of students. Instead of asking each individual student , survey is done only with sample.

Importance of sampling:



Time efficient : Studying entire population will take a lot of time and studying sample from that population will be beneficial when timely decision are to be made.



**Cost efficient : It is impractical to collect information from entire population as it will be wastage of cost and resources , so to study sample will be cost effective.**

**Decision : Sampling provides valuable information for decision making in various domains.**

**Example : Ice cream wants to introduce a new icecream flavour . Before spending money in production , marketing first company wants to know will there be enough demand for it or not . For that if company surveys and reach out to each and every consumer then it may be in thousands or lakhs and it will be costly and impractical to study . So they select sample representative sample of consumers from target demographic .**

**Then conduct survey like taste tests to get the insights about demand and get feedback , whether they are interested in buying or not.**

**Feasbility: Sampling makes research feasible when the population is too large. For example, it's not feasible to survey every single person in a country to understand their opinions on a particular issue. Sampling allows researchers to study a subset of the population and draw conclusions.**

**45. What is a Chi-Square test?**

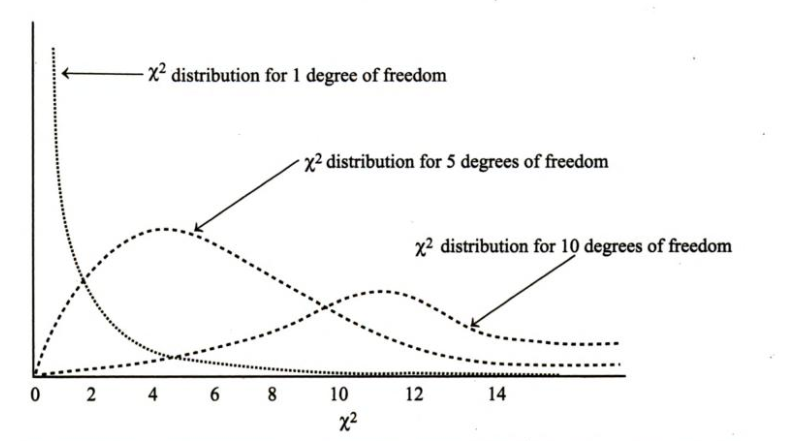
The chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables .It is non parametric test i.e it doesn’t require prior knowledge about the population .It is used for the testing of hypothesis. It affords a measure of the correspondence between theory and experiment.

**χ2**  distributiom is function of its degree of freedom . The distribution is skewed to right. Being a sum of square quantities **χ2** distribution can never be a negative value.

The formula for chi-square is:

**χ2 = ∑(Oi – Ei)2/Ei**

Oi 🡪 observed value Ei 🡪 expected value.

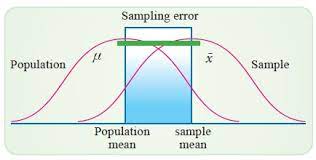


Application of chi square test :

* As a test of goodness of fit : It enables to ascertain whether the known probability distribution such as binomial , poisson , normal distribution fit or match with actual sample distribution . It provides a platform that can be used to ascertain whether probability distribution concide with empirical sample distribution. **χ2**  test compares the expected frequencies with actual to determine the difference between them.

**46. What is a Sampling Error and how can it be reduced?**

When sample does not truly represent entire population then statistical error arises which is called as sampling error. They are the difference between real value of population and values derived by using samples from population.



Sampling error can be reduced by following ways:

Increase sample size: As the sample size increases , it reduces variability and increase the precision of estimates.

Random sampling: Reducing bias ensures that each member of population has equal chance of being included , for that random sampling methods like simple random sampling, stratified sampling can be used.

Sampling Frames: A sampling frame is a list or database that includes all members of the population from which the sample will be drawn. Using sample frame helps to ensure selection bias and sample error is reduced and all members of the population have an equal chance of getting selected in the sample.

**47. What factors affect the width of a confidence interval?**

**Factors that effect width of confidence interval are sample size ,level of confidence , variability , standard error.**

### **Sample size : if sample size is large then confidence interval will be narrow. As the large sample provide more information and will be more precise.**

### **Level of confidence: Increasing level of confidence will make confidence interval wider. As higher confidence level will capture large range of value and will be more confident.**

### **Variability: If data points have large range the estimation will be less precise and will lead to wider interval , so greater variability results in wider confidence interval.**

### **Standard Error : The standard error is small then it will result in narrow confidence interval.**

**48. How do you determine if two random variables are independent based on their joint probability distribution?**

### **Let’s suppose there are two random variables X and Y , they are independent if their joint probability distribution is expressed as product of their marginal probability distribution**

### **To determine if two random variables are independent bases on their joint probability distribution there are two ways :**

### **Calculate marginal probability distribution :**

### **First calculate the marginal probability distribution for each random variable say X and Y. Then sum or integrate all possible values of other variables .**

### **For example To find P(X=x) , sum P(X=x,Y=y) over all possible value of y**

### **Comparison of marginal and joint probability distribution :**

### **If the joint distribution can be expressed as product of marginal distribution for all possible values of x and y then variables X and Y are independent.**

**49. What is a joint probability distribution?**

**It provides the probability of two or more random variables occurring simultaneously. It gives the likelihood of various combinations of outcomes for multiple random variables.**

**It is represented by tables, graphs.**

**If there are two random variables X , Y then joint probability distribution P(X,Y) assigns to each possible combination of values that X and Y can take.**

**Let’s understand with example.**

**Suppose there are two events , one is flipping coin and another is rolling a dice . So X is outcome of flipping coin and Y is outcome of rolling dice.**

**Then the joint probability distribution P(X,Y) are (heads,1) , (heads,2) , (head,3), (tail,1) etc.**

**50.What is Bessel’s correction ?**

### **It is the use of n-1 instead of n in the formula for sample variance and sample standard deviation . It is used for the correction of bias in the estimation of the population variance.**

### **The 1/n variance formula is systematically biased . It gives lower estimate than we would make if we had the population mean available. It is used because the mean of samples has error or bias from the true mean .Diving by n-1 gives the better estimate.**

### ****Video Presentation Link :****