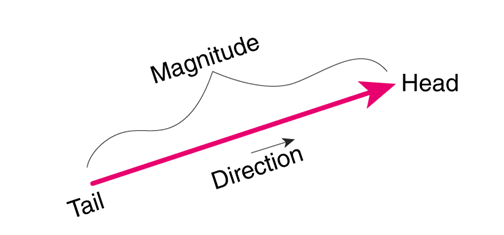
**Problem Statement: Write the Solutions to the Top 50 Interview Questions and explain any 5 Questions in a Video**

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

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



The direction of the vector is from the tail to the 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**.

The magnitude of the vector a is represented by **∥a∥**

### Vectors have a length, which is the magnitude, and have direction. In layman's terms, if a car is moving at a speed of 50km/hr and it's going towards home, then it is called a vector quantity.

### ****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 multidimensional. |
| With a change in quantity, has a change in magnitude. | With the change in quantity comes a change in either magnitude or direction or both. |
| Mathematical scalar operations result in a scalar only. | Vector operations can produce both scalar and vector results. The product of two vectors is a scalar. Subtraction, cross product, and summation of two vectors are also vector operations. |
| 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.

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

### The magnitude of a vector is the 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 the Pythagorean theorem in Euclidean space.

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

The main difference between a square matrix and a rectangular matrix is:

A square matrix has an equal number of rows and columns, and a rectangular matrix has an unequal number of rows and columns.

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

### It can be concluded in the end that shape and size are the basic differences between a square matrix and a rectangular matrix.

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

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

Eigenvector equation Av = λ**v**

**A is a square matrix, v is a non-zero eigenvector,** λ **is a scalar, and an eigenvalue of matrix A.**

**Eigenvectors and eigenvalues are used in many applications in linear algebra,** solving linear differential equations, and PCA.

### ****What is probability theory?****

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

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

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

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

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

### 

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

### **The conditional probability states the probability of an 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) is the conditional probability of event A given that event B has already occurred.**

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

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

### 

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

### Tossing a fair coin 3 times.

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

### The sample space is 8

### First Event (E) occurs at least 2 tails occur, the probability P(E) is 4/8, that is 1/2

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

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

### P(E/F) = P(E∩F)/P(F)

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

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

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

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

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

### 

### Uses of Bayes' theorem

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

### Weather forecasting: It is used to get the probability of future weather conditions based on current data.

### Spam filter: It is used to determine the probability that an incoming email is spam or not based on its content.

### ****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.

### Let's understand with an 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 1 to 6, it is random because the outcome of rolling a dice is uncertain.

### In a regular variable, let’s suppose x = x + 5, then with this information, we find x is 5, which is a fixed value.

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

### Discrete random variable: A 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: A Continuous random variable can take on any value in a continuous range. Example: the 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.

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

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

### **Let's understand it with the help of an example:**

### **If we flip a coin, there is a 50% chance of getting a head or a tail. So, 50% is the expected value (population mean)**

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

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

### **Now again tossing 1000 times, we get 543 heads, so the probability is 54.3%**

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

### ****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 the mean will always be normal, regardless of whether the population has a normal, binomial, or any other distribution.

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

### The sample size should be large enough, greater than or equal to 30.

### The sample size should be drawn from a finite population or an infinite population with a finite variance.

### 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 predict the population mean based on the sample data.**

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

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

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

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

### **The continuous probability distribution gives the probability of a continuous random variable. In a continuous probability distribution continuous random variable can take values that can be infinite and uncountable.**

### **Example: Time is infinite; it can be from 0 to trillion seconds, or the weight of a 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 the heights or weights of individuals.

### ****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 them by the number 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 a dataset in which, dataset is arranged in ascending order. If the dataset has an even number of values, then the median is the mean of the middle two values.

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

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

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

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

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

### ****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 percentiles 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 skewness.
* 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%.

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

### **Covariance of a 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 the strength of a relationship.

### It quantifies the degree to which variables tend to move concerning 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 the expected value**

### **μx is the mean of variable X**

### **μy is the mean of variable Y**

### **If covariance is positive, it means that if one variable increases, then the other variable will increase.**

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

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

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

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

### **Simple random sampling**

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

### 

### **Example: Population of different ages, sexes, and occupations**

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

### **Systematic sampling**

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

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

### 

### **Cluster sampling**

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

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

### 

### **Example: Broad category of population. Divide the entire population into groups, which are externally heterogeneous but internally homogenous. After identifying the cluster, we will pick all elements of the selected cluster.**

### **Stratified sampling**

### **In this method, first, researchers divide the 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 combine it to form their complete representative sample.**

### 

### **Example: Divide the population into groups: male, female, and elderly people. Select a sample from these samples. Select at least one person from each group.**

### Convenience sampling

### It involves selecting a sample based on convenience.

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

### 

### Snowball sampling

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

### 

### Quota sampling

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

### 

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

### Judgemental sampling

### 

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

### ****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, and standard deviation, while in hypothesis testing, it is used to make inferences about population parameters or test specific hypotheses about the population.**

**In parameter estimation, statistical methods such as maximum likelihood estimation are used, while in hypothesis testing null hypothesis and the alternative hypothesis are used.**

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

**Example:**

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

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

### ****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 evidence against the null hypothesis.**

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

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

### 

To decide on 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).

### ****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 is called a false positive or an error of the first kind.

In the type I error, researchers reject a null hypothesis when it's 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 true, results will have a 5% chance of occurring or less than that.

The results are statistically significant and consistent if the p-value is less than the significance level of the alternative hypothesis.

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

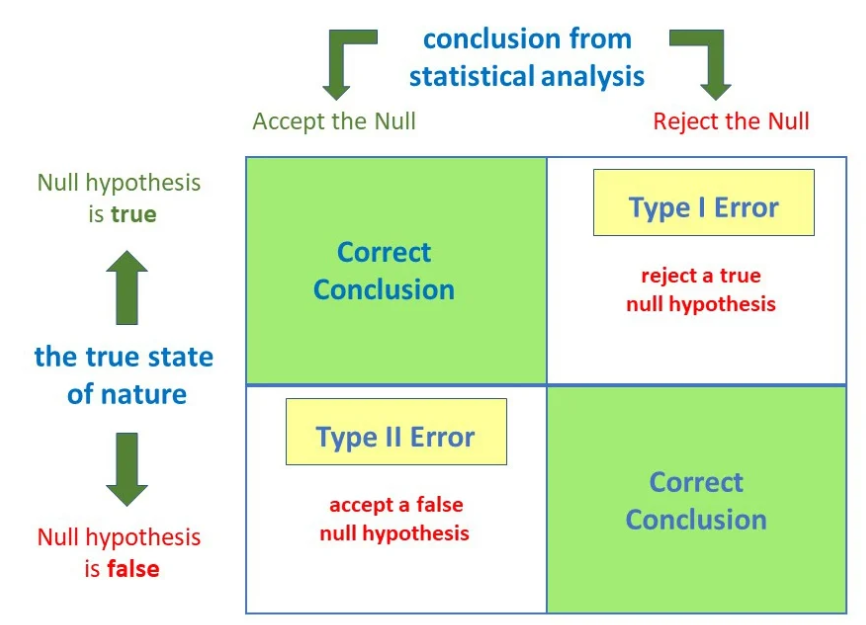
Type II Error:

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

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

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

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



### Example:

### Tested for COVID based on mild symptoms

### Two errors could occur:

### The test result says you have coronavirus, but you don’t. It is a Type I error (false positive).

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

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

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

### Causation is where one variable directly influences change in another variable. There is an effect relationship between variables.

### Correlation does not imply causation. If a correlation exists between variables, it does not mean that one variable has caused the change in the other variable.

### 

### Example:

### Data is collected on sales, and it is found that there is a positive correlation between the sales of ice cream and the sales of sunscreen. If we try to plot a scatter plot to visualize the relationship, we can see that as the sale of ice cream increases, the sale of sunscreen increases. It does not imply that an increase in ice cream has caused people to buy sunscreen; there is no direct causal link between ice cream sales and sunscreen sales. There is a third variable that is summer weather, because of that, sales increase in both products.

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

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

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

### 

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

### **If a YouTube channel has several subscribers and the average age of them lies between 25 to 32, then this is a confidence interval, and the confidence level is 95 %, which means that we are 95 % confident that the average age is between 25 to 32.**

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

### **Hypothesis testing is used to make conclusions 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 observation. It is important to note that we are not making an absolute conclusion, but we are concluding that the results we are getting are due to chance.**

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

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

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

### **In hypothesis testing, we state the assumed or hypothesized value of the population parameter before we begin sampling. The assumption we wish to test is called the null hypothesis and is denoted by H0.**

**It is referred to as the Status Quo.**

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

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

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

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

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

### **Two-tailed test:**

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

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

### 

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

### **Example:**

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

### **One-tailed test:**

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

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

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

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

### **One-tailed tests are of two types:**

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

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

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

### ****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 the Dot Product of Two Vectors:

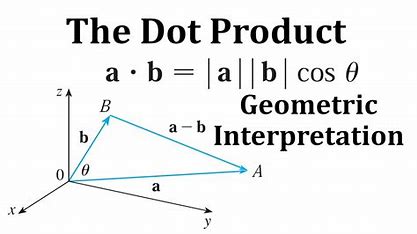
There are two vectors, P and Q.

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 that is formed by the intersection at one point

Cos θ = (a.b)/|a|\*|b|

### ****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 the Cross Product of Two Vectors:

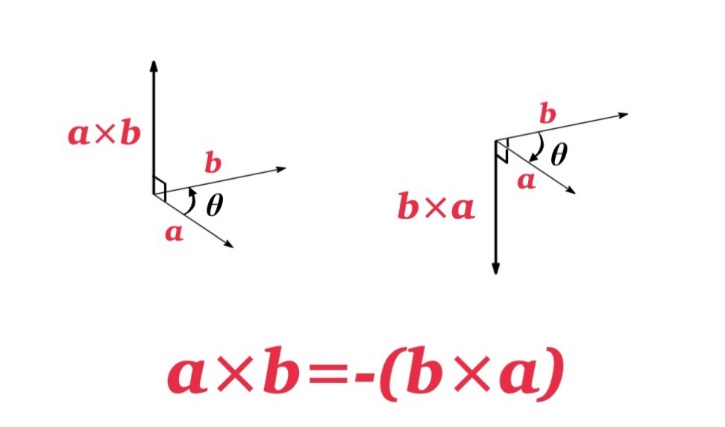
Suppose there are two vectors, P and Q.

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 a 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.

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

### **Observational data is collected by observing what is happening and recording the observations. There is no manipulation by the researcher. It is used for descriptive and exploratory purposes. In an observational study, it may show that there is a relation between variables, but it doesn't need to be a cause-and-effect relation.**

### **Example: Study the 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 behavior. 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 effect on other variables, so here experiments are controlled and manipulated. This leads to a cause-and-effect relationship between variables. It is also called a scientific study.**

### **It involves two groups, i.e. experimental group and the control group**

### **Example:**

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

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

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

### **The confidence test and the hypothesis test both use sample data collected from a population to conclude population parameters. They both are statistical methods and involve the calculation of probabilities to assess the likelihood of certain outcomes.**

### **The difference between confidence tests and hypothesis tests is:**

* Confidence test estimates a range of values that is a 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 parameters or test specific hypotheses about the population. The outcome of hypothesis testing is to reject or not the null hypothesis based on the evidence drawn from sample data.**
* In a confidence test, the result is expressed as a confidence interval, which is the 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 the confidence interval. In a hypothesis test, the decision-making process involves comparing the calculated p-value to a predetermined significance level.

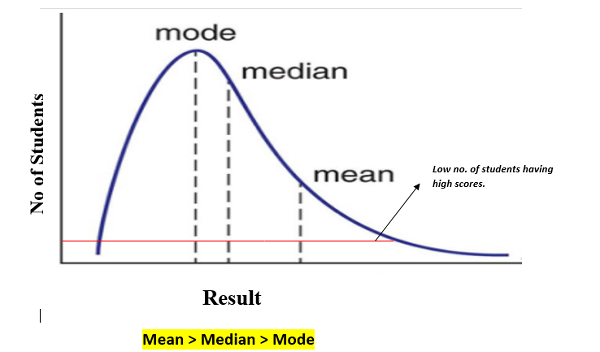
### ****What is the left-skewed distribution and the right-skewed distribution?****

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

Positive Skewed or Right-Skewed (Positive Skewness):

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

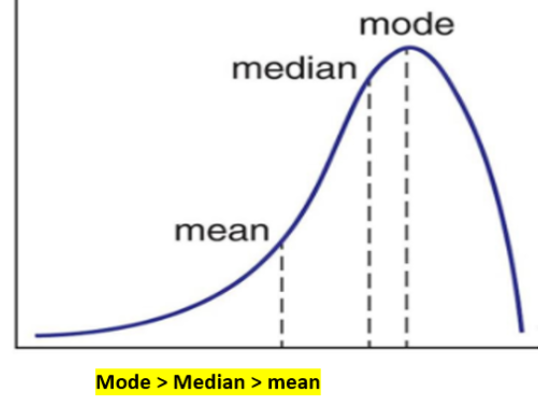
The mean is greater than the median. The result is bent towards the lower side. So, the mean will be more than the median.

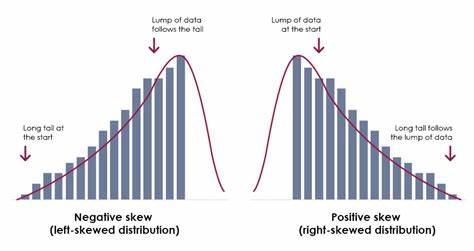


Negative Skewed or Left-Skewed:

It refers to the distribution model where the tail of the distribution is spreading on the left side, and more values are plotted on the right side of the graph. It has a long left tail.

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





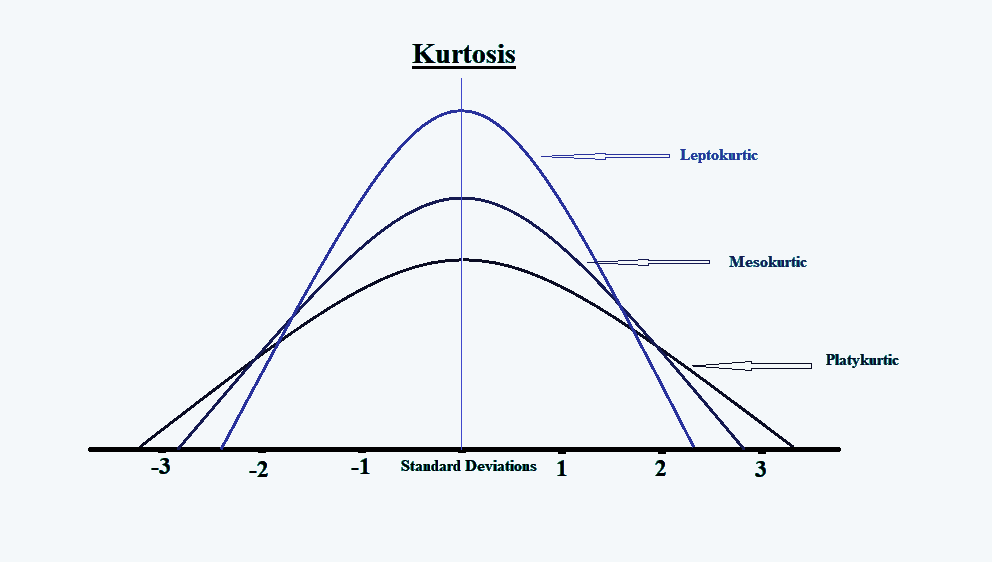
### ****What is kurtosis?****

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

The measure of kurtosis is tailed-ness of distribution, and tailed-ness refers to how often outliers occur.

Peaked-ness is the degree to which data values are centered around the mean. Datasets with low kurtosis have a flat top near the mean, and datasets with high kurtosis have a distinct peak near the mean, which declines rapidly with heavy tails.

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



Mesokurtic, Leptokurtic, and Platykurtic are three different types of kurtoses.

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

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

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

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

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

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

### **So, the number of outcomes is 4.**

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

### **So, the number of outcomes is 5.**

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

### **Probability of getting a 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**

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

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

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

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

### Descriptive statistics uses methods like measures of central tendency, measures of dispersion, while inferential statistics uses techniques like probability, hypothesis testing, and ANOVA test to make predictions.

### Example: Student’s marks in class

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

1. **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?**

The 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**

1. **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, which means there is a positive relation, i.e., when the variables move together in the same direction. When the time a person sleeps increases, the amount of productivity also increases. When time is reduced, productivity also reduces.

1. **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 will see at least one supercar in 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 a supercar in the period of 60 minutes =0.7 \* 0.7 \* 0.7 =0.343

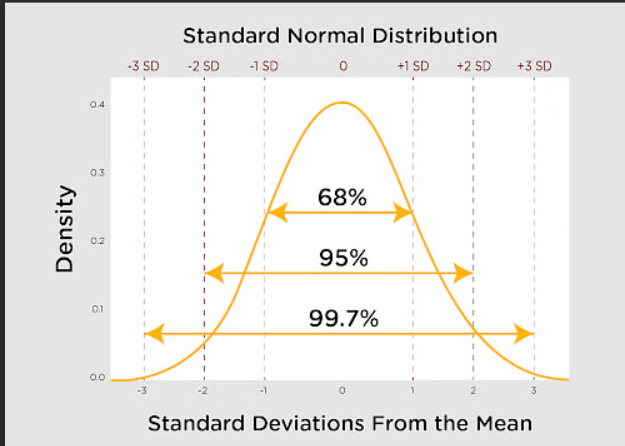
### So, the probability of seeing at least one supercar in 60 minutes is:

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

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

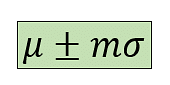
It is a statistical rule that says observed data for a normal distribution will fall within three standard deviations, which is denoted by σ of the mean or average, which is denoted by µ.

It is also known as the three-sigma rule or 68-95-99 rule, which 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 the mean.



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

The empirical rule formula is:



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

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

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

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

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

### **No confidence interval can be used to make a definitive statement about a specific individual in a population.**

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

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

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

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

### **The large sample will be more representative of the population, and it will give a true parameter value and a margin of error that decreases. A smaller sample will be prone to sampling error and variability.**

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

### Yes, two confidence intervals with different widths can have the same confidence level. The width of the confidence interval is affected by sample size and variability. A wider interval means more uncertainty, while a narrow interval means greater precision.

### Let’s understand with an example:

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

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

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

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

Assumptions for the t-test:

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

Three types of t-tests can be performed:

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

Example: compare the average scores of the male students to the average scores of the females. So, the 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 are compared at two different times or under two different conditions.

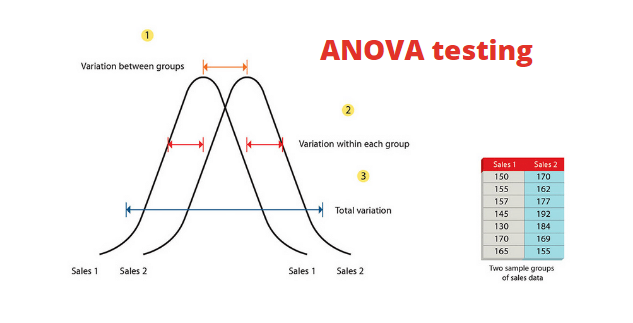
Example: To increase the productivity levels of an employee, a training program was conducted. Now, to measure if the productivity levels have increased, we will compare the productivity level before and after the training program. So, the same sample, i.e., the employees, is compared at two different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

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

ANOVA, or Analysis of variance, is a technique of testing hypotheses 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 populations.

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

Variance between samples is attributed to differences among sample means. Variance within the 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 the comparison of two means from two independent groups using the F distribution. The null hypothesis for the test will be that the two means are equal. So significant result means that the two means are unequal.

Two-way ANOVA is an extension of one-way ANOVA. There are two independents. There is one measurement variable and the other is a nominal variable. We can use two-way ANOVA where the experiment has a quantitative outcome and has two categorical variables.

1. **What is an inlier?**

### **An inliner, also known as an 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.**

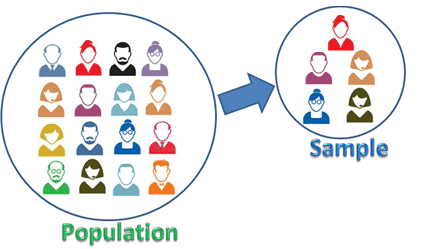
### **It is a point that is considered to be part of the normal behaviour of data. It aligns with the general pattern or trend exhibited by the majority of the dataset. Inliers contribute to a better understanding of the central tendency and relationship within the data.**

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

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

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

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



Example: A College student's survey about food.

Suppose there are thousands of students in college. All students will be the population, but it’s 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 student, the survey is done only with the sample.

Importance of sampling:



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



**Cost-efficient: It is impractical to collect information from the entire population as it will be a waste 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 ice cream flavour. Before spending money on production, the marketing company first wants to know whether there will be enough demand for it or not. For that, if the company surveys and reaches 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 the target demographic.**

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

**Feasibility: 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.**

1. **What is a chi-squared test?**

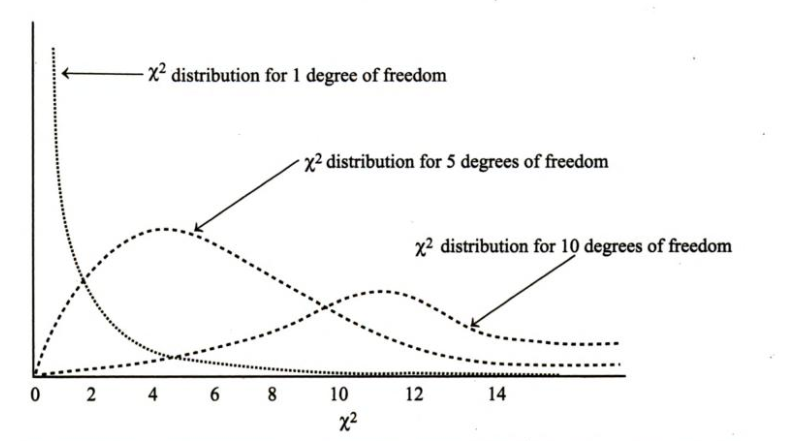
The chi-square test is a statistical test used to determine whether there is a significant association between two categorical variables. It is a non-parametric test, i.e., it doesn’t require prior knowledge about the population. It is used for the testing of hypotheses. It affords a measure of the correspondence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

**The χ2** distribution is a function of its degree of freedom. The distribution is skewed to the right. Being a sum of square quantities, the **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.

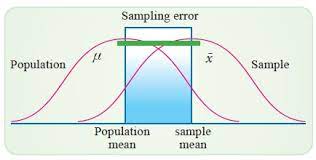


Application of the chi-square test:

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

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

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



Sampling error can be reduced by following these ways:

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

Random sampling: Reducing bias ensures that each member of the population has an 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 a sample frame helps to ensure that selection bias and sample error are reduced and all members of the population have an equal chance of getting selected in the sample.

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

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

### **Sample size: If the sample size is large, then the confidence interval will be narrow. The large sample provides more information and is more precise.**

### **Level of confidence: An Increasing level of confidence will make the confidence interval wider. A higher confidence level will capture a larger range of values and will be more confident.**

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

### **Standard Error: The standard error is small, which will result in a narrow confidence interval.**

1. **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 the product of their marginal probability distribution**

### **To determine if two random variables are independent based 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 a product of marginal distributions for all possible values of x and y, then variables X and Y are independent.**

1. **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 the joint probability distribution P(X, Y) assigns to each possible combination of values that X and Y can take.**

**Let’s understand with an example:**

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

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

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